university of groningen

campus fryslân

Fine-tuning Cantonese based on Wav2vec 2.0 XLRS model that pretrained on Mandarin Chinese to improve ASR performance

Qing Li

university of groningen / campus fryslân

University of Groningen - Campus Fryslân

Fine-tuning Guangdong Cantonese based on Wav2vec 2.0 XLRS model that pretrained on Mandarin Chinese to improve ASR performance

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. Shekhar Nayak (Voice Technology, University of Groningen)
with the second reader being
Associate Professor Matt Coler (Voice Technology, University of Groningen)

Qing Li (S5600502)

July 1, 2024

Acknowledgements

I would like to express my deepest gratitude to Dr. Shekhar Nayak for his invaluable guidance and mentorship throughout the course of my thesis. His insights and support have been instrumental in the successful completion of this work.
I am also profoundly thankful to my study advisor, Hieke Hoekstra, for her constant care and concern for my personal well-being, as well as her assistance during the process of applying for extensions.
Additionally, I would like to extend my heartfelt thanks to my fellow students for the stimulating discussions and mutual inspiration we shared. Their support and collaboration have greatly contributed to my academic journey.
A special thanks to Yue Fang, whose encouragement kept me going even when I felt like giving up. His support and assistance throughout this journey have been invaluable.

Abstract

This study investigates the effectiveness of cross-lingual transfer learning for Cantonese Automatic Speech Recognition (ASR) by comparing a baseline wav2vec2 XLRS model pre-trained on multiple languages with a transfer learning model pre-trained on Mandarin. The baseline model achieved a Character Error Rate (CER) of approximately 0.3, while the transfer learning model demonstrated a significantly lower CER of around 0.2 after 40 epochs of training. The transfer learning approach showed enhanced training efficiency, faster convergence, and robust generalization ability, despite the baseline model's slight advantage in validation loss during later stages. These findings validate the hypothesis that leveraging a pre-trained Mandarin model, fine-tuned with limited labeled Cantonese data, significantly outperforms the baseline model. This study underscores the potential benefits of cross-lingual transfer learning, particularly between linguistically similar languages, and highlights its importance for developing inclusive and diverse ASR systems for under-resourced languages.

Contents

1 Introduction

Language is not merely a collection of grammar and vocabulary; it is a vessel for cultural heritage. Acquiring a language entails understanding and preserving the culture and history it represents. Languages across the world, shaped by their unique cultural and regional contexts, hold significant cultural importance and linguistic research value. However, many languages are on the brink of extinction. Krauss (1992) initially estimated that only 10% of the world's languages are safe in the long term, with up to 50% already moribund. Recent data from the Ethnologue Lewis, Simons, and Fennig (2013), using the Expanded Graded Intergenerational Disruption Scale (EGIDS), provide more precise estimates confirming these dire predictions.

Regional languages often face encroachment from the dominant languages of neighboring powerful nations, a phenomenon exacerbated by globalization. For example, in Australia, Canada, and the United States, over 75% of languages are now extinct or moribund. Urbanization presents a new threat, where the necessity of acquiring dominant urban languages exerts pressure on minority languages. Besides, colonization has significantly impacted language loss. In settlement colonies, such as Australia and North America, the large-scale settlement of colonizers led to deep and prolonged language contact, resulting in significant language shift and loss. For instance, in North America, colonizers decimated indigenous populations through disease and warfare or forced relocations, severing ties to their native languages. This political phonology inflicted severe harm on the languages of vulnerable regions, highlighting the profound impact of colonial practices on language extinction (Simons & Lewis, 2013).

China is home to a multitude of dialects, each with its own distinct phonetic and phonological characteristics. From a linguistic perspective, many of these dialects can be considered independent languages due to their unique pronunciation systems and strong regional identities. The primary dialects in China include Northern Mandarin, Wu, Gan, Xiang, Min, Hakka, and Yue(Cantonese), among others. Each of these dialects exhibits significant variation in tone, vowel, and consonant structures. People from different regions of China often speak both their regional dialect and Mandarin, making them multilingual individuals, but ASR systems predominantly focus on Mandarin, leaving dialect ASR systems and data resources relatively scarceQ. Li, Mai, Wang, et al., 2024.

The Yue dialect(Cantonese) in China is a language of wider communication around the world, originated in China, Hong Kong, and Macao. It belongs to the Sino-Tibetan language family and is part of the Chinese macro-language. The language is used as a first language by all in the ethnic community and used as a language of instruction in education.(Ethnologue 2024) Due to differing political and social contexts and relative isolation, Cantonese has developed distinct spoken and written forms across various regions.

Cantonese holds a unique cultural and political significance among the people of Guangdong in mainland China, surpassing other Chinese dialects in its revered status and fostering a strong regional cultural identity. The "Protecting Cantonese Movement" (PCM), which began in 2010 and lasted more than 10 years, was initiated as a response to a proposal by the Guangzhou Committee of the Chinese People's Political Consultative Conference (CCPPCC) to switch local television broadcasts from Cantonese to Mandarin in an effort to attract global visitors during the 2010 Asian Games.Y. Li, Kang, Ding, and Zhang 2022 Such a movement is unlikely to occur among speakers of other Chinese dialects. Promoting Cantonese speech recognition in Guangdong can aid in the preservation and development of linguistic information related to the regional Cantonese dialect through technological means, amidst the widespread promotion of Mandarin. This effort supports the transmission of

Cantonese and the protection of its associated cultural heritage.

Cantonese is the predominant and most widely used language in Hong Kong, spoken by 90% of its approximately 6.5 million ethnic Chinese residents as their daily language. It holds a unique status due to its distinctive vocabulary, indigenous Chinese characters, colloquial phonetic features, conventionalized written form, extensive English loanwords, and a tradition of lexicography with romanization. Despite its prevalence, there is a concerning trend of increasing numbers of schools switching their medium of instruction from Cantonese to Putonghua.(Bauer 2016)And in recent years, Language education policies in Hongkong have evolved to promote Chinese-medium instruction in schools, with recent shifts allowing more flexibility for schools to choose their medium of instruction. Putonghua(Mandarin Chinese) was Increasingly promoted in schools, reflecting broader national policies aiming for linguistic unification.(Bolton 2024) This shift marks a significant change since Hong Kong's return to Chinese sovereignty in 1997.

With such significant international influence, surpassing that of a typical Chinese "dialect," Cantonese can be recognized as an independent linguistic entity. However, its development in the field of speech technology remains underdeveloped. This gap is particularly critical given Cantonese's extensive usage in regions such as Hong Kong and Guangdong, where it serves not only as a medium of daily communication but also as a cultural identifier. Enhancing speech recognition technology for Cantonese is essential to preserving its linguistic heritage and ensuring technological inclusivity for its speakers.

Over the past few decades, automatic speech recognition (ASR) has made remarkable strides, achieving impressive milestones. The evolution from Gaussian Mixture Model-Hidden Markov Models (GMM-HMM) to Deep Neural Networks (DNN) has enabled machines to learn hierarchical feature representations, which can capture more abstract features at higher layers. This transition has significantly enhanced the ability of ASR systems to recognize and interpret speech with greater accuracy and efficiency(Hinton et al., 2012). One application of this is Deep Speech 2, which replaces the traditional pipeline of hand-engineered components (feature extraction, acoustic models, language models, etc.) with a single end-to-end deep learning model. With such methods, DS2 matches or exceeds the transcription accuracy of human workers in several benchmarks(Amodei et al. 2015). Another notable development is Jasper (Just Another SPeech Recognizer), created by NVIDIA, which is a convolutional neural network-based model designed for sequence-to-sequence ASR tasks. Jasper is known for its high accuracy and efficiency in recognizing speech(J. Li et al. 2019). But DS2 and Jasper rely heavily on supervised learning and require large amounts of labeled data to train their models. While they perform well with enough labeled data, their effectiveness drops significantly when data is scarce. In contrary, Wav2vec 2.0, developed by Facebook AI, leverages self-supervised learning to train on vast amounts of unlabeled audio data, followed by fine-tuning on smaller labeled datasets(Baevski, Zhou, Mohamed, and Auli 2020a). This two-stage training paradigm has proven effective in improving ASR performance, especially in low-resource settings where labeled data is scarce.

Despite the advancements in automatic speech recognition (ASR) technology, many ASR systems primarily focus on high-resource languages such as English and Mandarin, which benefit from abundant and high-quality datasets. In contrast, under-resource languages and dialects, such as Cantonese, often lack sufficient training data, hampering the development of robust ASR systems.

Under-resourced languages often lack standardized writing systems, have limited online presence, and face shortages of linguistic expertise and electronic resources. Challenges include the absence of monolingual corpora, bilingual dictionaries, transcribed speech data, pronunciation dic-

tionaries, and sufficient vocabulary lists. Innovative methods such as crowd-sourcing for data collection, cross-lingual transfer learning, and leveraging multilingual resources have been crucial in addressing data scarcity. Successful cross-lingual acoustic modeling leverages data from well-resourced languages to bootstrap models for under-resourced languages(Besacier, Barnard, Karpov, & Schultz, 2014). Cantonese, spoken by over 100 million people, is considered a under-resource language due to the scarcity of standardized textual and audio corpora.

Cantonese, like many other Chinese dialects, belongs to the Sino-Tibetan language family. Transfer learning shows more significant improvements when there are stronger linguistic similarities between the high-resource language and the under-resource language being trained. Therefore, leveraging existing ASR models for Mandarin can provide a solid foundation for developing Cantonese speech recognition systems. This approach exploits the linguistic commonalities within the Sino-Tibetan language family, enabling more efficient and accurate model adaptation for Cantonese. To achieve this, I plan to use the wav2vec 2.0 model.This model's ability to learn powerful audio representations from large amounts of unlabeled data and fine-tune on smaller labeled datasets makes it particularly suitable for under-resource languages like Cantonese. By fine-tuning a pre-trained Mandarin ASR model using wav2vec 2.0, we can effectively improve the performance of Cantonese speech recognition, addressing the challenges posed by the limited resources available for this dialect. This method leverages the shared phonetic and phonological features within the Sino-Tibetan language family, ensuring a robust and efficient adaptation process.

Now that a brief motivation for this research has been presented, the structure of the thesis is the following: subsection 1.1 introduces the research question posed along with a hypothesis on the outcome of the research. Section 2 provides an extensive literature review that frames the research question and hypothesis in the state-of-the-art. In section 3, the methodology is covered and the underlying models used are explained. Then, section 4 describes the experimental setup developed to answer the research questions and validate the hypothesis. Section 5 describes the results obtained in detail and compares them to the baseline. Lastly, section 6 summarizes the thesis and presents the conclusions drawn, along with recommended future work.

1.1 Research Question and Hypothesis

In light of the preceding discussion, the research question at the core of this study can be formulated as follows:

> **Can using pretrained Mandarin wav2vec2 model improve the performance of Cantonese ASR than using wav2vec2 XLRS as a pretrained model?**

From which the following subquestions are derived:

- What is the baseline CER achieved when using wav2vec2 XLRS as a pretrained model?
- Can the model pretrained on Mandarin improve over the baseline model?

My hypothesis is that using pretrained Mandarin wav2vec2 model and fine-tune it with limited labeled Cantonese speech dataset will improve over using wav2vec2 XLRS as pretrained model and fine-tune it in the same way significantly. The falsification of the hypothesis would suggest that wav2vec2 XLRS model is better than pretrained Mandarin wav2vec2 model.

2 Literature Review

This section is dedicated to providing a comprehensive review of the existing research pertaining to ASR for under-resourced languages, with a specific focus on Cantonese, particularly Guangzhou Cantonese. The emphasis of this thesis is on the transfer learning approach from Mandarin to Cantonese. By conducting a thorough and critical analysis of the literature in this field, this review aims to offer valuable insights into the methods and effectiveness of applying transfer learning to Cantonese based on Mandarin models. This approach leverages the linguistic similarities within the Sino-Tibetan language family to improve ASR performance in low-resource settings. The review will explore various techniques and models, including the use of state-of-the-art technologies such as wav2vec 2.0, to demonstrate how they have been applied and the outcomes achieved in enhancing Cantonese ASR.

To those ends, the section is structured as follows. To begin, I will delineate the keywords used during the comprehensive literature search described above and describe the inclusion/exclusion criteria used in selecting the literature. After that, I offer a succinct overview of the key findings and contributions of the selected papers (in subsections 2.1-2.X).

I have grouped the keywords according to the topic they are related to. The topics are highlighted in bold, after which the keywords for that topic are mentioned. Thus, the topics and their corresponding keywords are:

- **Transfer learning:** transfer learning ASR, transfer learning speech recognition;
- **low-resource language ASR:** Cantonese speech recognition, Guangzhou Cantonese, dialect ASR, Cantonese acoustic modeling;
- **Mandarin to Cantonese transfer:** Mandarin to Cantonese ASR, Mandarin-based ASR for Cantonese, language adaptation;
- **wav2vec 2.0 and ASR technologies:** wav2vec 2.0, self-supervised learning, end-to-end ASR, neural network models in ASR, speech representation learning;

To streamline the paper selection process, I organized the papers based on their relevance to specific topics and keywords. However, not all retrieved literature was directly related to the research question topic. Therefore:

1. To maintain coherence, I excluded papers that pertained to different tasks;
2. To ensure the inclusion of recent research, the publication dates were limited to papers from 2010 onwards. This decision was made to reflect the latest advancements and methodologies in ASR for under-resourced languages;

These criteria ensured that the literature review was both current and highly relevant to the focus of this research, providing a solid foundation for understanding the state-of-the-art in ASR technology and its application to Cantonese.

These keywords guided the literature search and helped identify relevant studies that contribute to the understanding and advancement of ASR for Cantonese, particularly through the use of transfer learning from Mandarin.

Next, I describe the inclusion and exclusion criteria used to select the literature. The inclusion criteria were: (1) studies focusing on ASR for under-resourced languages, specifically Cantonese, (2) research involving transfer learning techniques, (3) papers discussing the use of wav2vec 2.0 or similar advanced models in ASR. The exclusion criteria were: (1) studies not directly related to ASR or transfer learning, (2) papers lacking sufficient experimental results or methodological details, and (3) non-English publications.

Following this, I provide an overview of the key findings from the selected papers, organized by the aforementioned topics. Each subsection (2.1-2.X) will delve into specific aspects such as the effectiveness of transfer learning techniques, the unique challenges and solutions in developing Cantonese ASR, and the impact of utilizing advanced models like wav2vec 2.0 on performance improvements. This structured approach aims to offer a comprehensive understanding of the current state and future directions of ASR research for Cantonese and other under-resourced languages.

The literature review is organized into different subsections based on the general topics they cover. Subsection 2.1 discusses the literature regarding low-resource language automatic speech recognition (ASR), exploring the challenges and methodologies in developing ASR systems for languages with limited resources. Subsection 2.3 presents an overview of various techniques introduced to enhance ASR performance, including advanced neural network models especiallly wav2vec2. Moving towards the broader subfield of multilingual and cross-lingual approaches.

2.1 low resource language Automatic Speech Recognition(ASR)

Low-resource language ASR refers to the development and implementation of automatic speech recognition systems for languages that lack extensive and high-quality linguistic resources, such as large annotated datasets, comprehensive lexicons, and robust language models (Besacier et al., 2014). These languages often have insufficient funding for large-scale data collection and annotation efforts. Although some of these languages may have many speakers and significant cultural impact, they still suffer from a lack of comprehensive linguistic resources. As a result, they are often overlooked in the development of ASR systems, primarily due to the lack of commercial incentives. However, developing ASR for low-resource languages remains critically important as it promotes equal access to technology and helps preserve diverse cultural heritages (Kwon & Chung, 2023).

To address these challenges, recent studies have explored innovative approaches to improve low-resource spoken language understanding (SLU) through multitask learning and transfer learning. For instance, Meeus, Moens, and Van hamme (2022) proposed a multitask learning model that jointly performs automatic speech recognition (ASR) and either intent classification or sentiment classification. Their approach showed significant improvements over single-task models, especially in low-resource scenarios. With as few as two examples per class, their multitask model outperformed baselines trained on text features or using a pipeline approach. Notably, their model achieved comparable performance to an end-to-end model with ten times fewer parameters on sentiment classification tasks.

Building on the concept of leveraging larger datasets, Wang, Long, Li, and Wei (2023) introduced the Aformer architecture for low-resource accented speech recognition. This approach combines a general encoder trained on large non-accented datasets with an accent encoder adapted to limited accented data. Their multi-pass training strategy and cross-information fusion methods effectively utilize both large-scale non-accented and limited accented speech data. This method achieved significant improvements in low-resource settings, with up to 24.5% relative WER reduction on unseen

accents compared to conventional fine-tuning.

These studies demonstrate the potential of multitask learning and transfer learning approaches in improving SLU and ASR performance for low-resource scenarios, offering promising directions for future research in this field.

2.1.1 Chinese Dialect Speech Recognition

China is home to a rich tapestry of dialects, many of which are distinct languages from a linguistic perspective, rather than mere dialects of Mandarin. In southern Chine, the dialects include Cantonese, Hakka, and Minnan, possess unique phonetic, lexical, and grammatical features. Among these, Cantonese is particularly noteworthy due to its widespread use in multiple countries and regions, its significant cultural and political influence, and its distinct dialectal variations developed in different cultural contexts(Q. Li et al., 2024).

In Guangdong, for instance, Cantonese has developed several regional accents, each with unique characteristics. Despite these variations, most current Cantonese automatic speech recognition (ASR) systems focus primarily on recognizing the Hong Kong variant of Cantonese. This narrow focus can result in suboptimal recognition performance and a lack of robustness in ASR applications across different Cantonese-speaking regions(Yu et al., 2022).

The Common Voice dataset, developed by Mozilla, addresses this issue by incorporating contributions from speakers with various accents. This crowd-sourced approach ensures that the dataset includes a wide range of Cantonese accents, thereby maximizing the diversity of Cantonese speech patterns. By leveraging such a rich and varied dataset, ASR systems can be trained to recognize and accurately transcribe the different regional accents of Cantonese, leading to more robust and versatile ASR applications across Guangdong and other Cantonese-speaking areas. This inclusivity is crucial for improving the overall performance and user experience of Cantonese ASR systems(Ardila et al., 2020).

2.2 Transfer Learning

Multilingual ASR models have shown great potential in improving recognition performance for low-resource languages. In a large-scale study covering 51 languages, Pratap et al., 2020 demonstrated that low-resource languages can significantly benefit from joint multilingual training. Their multi-headed model achieved an average relative Word Error Rate (WER) reduction of 28.76% on low-resource languages. Furthermore, they proved that such multilingual models can effectively transfer to unseen low-resource languages, further improving recognition performance.

Building on these findings, recent approaches to low-resource ASR have explored multilingual training strategies to leverage shared acoustic and linguistic properties across languages. For instance, Diwan et al., 2021 demonstrated this approach using six Indian languages from different language families. Their work employed hybrid DNN-HMM models, specifically time-delay neural networks (TDNNs) with the lattice-free MMI objective function for acoustic modeling. Additionally, they showed that transfer learning, where pre-trained multilingual models are fine-tuned on new low-resource languages, can significantly improve ASR performance with limited data.

Transfer learning has further shown significant promise in improving ASR performance for low-resource languages. In particular, the use of self-supervised pre-trained models, such as wav2vec 2.0 (Baevski, Zhou, Mohamed, & Auli, 2020b), has emerged as a powerful approach. Bartelds

and Wieling (Bartelds & Wieling, 2022) demonstrated the effectiveness of fine-tuning wav2vec 2.0 models for low-resource ASR tasks. Building on this, recent work by Bartelds, San, McDonnell, Jurafsky, & Wieling, 2023 explored the application of the multilingual XLS-R model, which is based on the wav2vec 2.0 architecture, to extremely low-resource scenarios. They found that fine-tuning this pre-trained model on as little as 24 minutes of transcribed speech from the target language could yield substantial improvements over traditional approaches. Furthermore, they investigated the potential of continued pre-training on the target language, although the gains from this method were limited compared to the computational cost. These findings underscore the power of transfer learning from large-scale multilingual models to resource-scarce languages, particularly when leveraging the wav2vec 2.0 architecture.

Additionally, Gupta & Boulianne, 2022 explored multilingual training approaches for low-resource ASR, focusing on three morphologically complex languages: Kurmanji Kurdish, Cree, and Inuktut. They investigated the transfer of knowledge from 12 languages by comparing separate language-specific phone sets versus merged common phones, finding that the optimal strategy varies depending on the target language. The authors demonstrated significant improvements through transfer learning, achieving word error rate (WER) reductions of up to 10.5% absolute for Kurmanji Kurdish and 8.6% absolute for Inuktut compared to monolingual training. Furthermore, they showed that fine-tuning the multilingual model with target language data for just one epoch can lead to substantial WER reductions, highlighting the effectiveness of transfer learning in low-resource scenarios.

While these transfer learning approaches have shown significant promise, the field of ASR continues to evolve rapidly with the development of more advanced models. These state-of-the-art models leverage self-supervised learning and large-scale training to further improve performance, especially in low-resource scenarios. Among these models, wav2vec 2.0 stands out for its effectiveness in low-resource settings and will be the focus of our study.

2.3 State-of-the-Art ASR Models

In the field of automatic speech recognition (ASR), significant advancements have been made with the development of models like wav2vec2 and Whisper. This section will provide an overview of many cutting-edge ASR models, highlighting their development, architecture, and contributions to the field.

2.3.1 Wav2vec Development History

wav2vec: The original wav2vec model introduced a self-supervised framework for learning speech representations directly from raw audio data. It focused on predicting future audio samples from past ones, using a contrastive loss to distinguish between true and false samples. This model demonstrated that useful speech features could be learned without extensive labeled data, setting the stage for future developments in unsupervised and self-supervised learning for ASR.(Schneider, Baevski, Collobert, and Auli (2019))

wav2vec2: Building on the success of wav2vec, the wav2vec2 model introduced several key improvements. baevski(Baevski et al. (2020b)) utilized a two-stage training process: self-supervised pretraining on unlabeled audio followed by supervised fine-tuning on a smaller labeled dataset. The model architecture incorporated a convolutional neural network (CNN) feature encoder and a transformer network, enabling it to capture more complex and contextually rich speech representations.

The wav2vec2 model significantly improved ASR performance, especially in low-resource settings where labeled data is scarce.
wav2vec2 XLSR: The wav2vec2 XLSR (Cross-Lingual Speech Representation) model represents a further advancement, specifically designed for multilingual ASR. This model is pretrained on a large, diverse set of languages, leveraging cross-lingual transfer to improve recognition performance across different languages. By learning from a wide array of linguistic data, wav2vec2 XLSR is particularly effective for low-resource languages, benefiting from the shared features across languages.(von Platen (2021))

2.3.2 wav2vec 2.0 Architecture

wav2vec 2.0 consists of a multi-layer convolutional feature encoder that transforms raw audio into latent speech representations. These representations are then passed through a Transformer-based context network, which generates contextualized embeddings by capturing long-range dependencies in the audio sequence. The model also includes a quantization module that discretizes the latent speech representations, which are used as targets during pre-training.(Schneider et al. (2019))

In conclusion, the advancements in the wav2vec framework, particularly with the introduction of wav2vec 2.0 and XLSR, offer promising solutions for improving ASR systems in low-resource languages. By leveraging these state-of-the-art models, we aim to enhance the recognition performance and robustness of Cantonese ASR, contributing to the preservation and accessibility of this linguistically rich language.

2.3.3 Other Notable ASR Models

While our study focuses on wav2vec 2.0, it's important to contextualize it within the broader landscape of ASR models:

Hinton et al. (2012) introduce several significant architectural innovations in acoustic modeling for automatic speech recognition (ASR) using deep neural networks (DNNs). The key advancements include the use of deep belief networks (DBNs) for unsupervised pre-training of DNNs, which initializes network weights through layer-wise stacking of restricted Boltzmann machines (RBMs), addressing the challenges of training deep architectures.

The authors propose replacing traditional Gaussian mixture models (GMMs) with multi-layer DNNs, typically consisting of 5-7 layers, enabling more complex feature representations. A crucial innovation is the integration of DNNs with hidden Markov models (HMMs) to form hybrid DNN-HMM systems, where DNNs compute posterior probabilities of HMM states. Hinton et al. employ context-dependent states as output targets, often numbering in the thousands, enhancing discriminative capability. They also utilize extended context windows for input features, usually spanning 11 frames, to better capture long-term speech dependencies. In some tasks, they show that log mel-filter bank features outperform traditional MFCCs as DNN inputs.

The paper also introduces an autoencoder bottleneck (AE-BN) feature extraction method for GMM-HMM systems and explores the application of convolutional neural networks in speech recognition, particularly using convolution and pooling operations in the frequency domain. These innovations collectively lead to significant performance improvements in ASR systems, consistently outperforming traditional GMM-HMM approaches across various benchmark tests including Switchboard,

Bing Voice Search, and Google Voice Input. The proposed approaches demonstrate effectiveness and robustness in large vocabulary continuous speech recognition tasks.

Similar to wav2vec2, Whisper, developed by OpenAI, is another notable ASR model that has been applied to low-resource scenarios. Whisper was trained on a large and diverse dataset, comprising 680,000 hours of supervised multilingual and multitask data. This extensive training allows Whisper to perform well in various acoustic conditions and linguistic contexts (Radford et al. (2022a)).

Whisper's zero-shot learning capability is particularly noteworthy. It can significantly reduce error rates across different datasets and languages without needing fine-tuning for specific datasets. Whisper achieves a 50% reduction in errors compared to other models in zero-shot settings. Additionally, Whisper is capable of handling multiple tasks such as speech transcription, voice activity detection, and speaker diarization, which simplifies the overall speech processing pipeline (Radford et al. (2022a)).

In another study, Whisper demonstrated robustness in noisy environments and with adversarial inputs, maintaining high accuracy and low WER under challenging conditions (Radford et al. (2022b)). These attributes make Whisper a powerful tool for ASR, particularly in low-resource and multilingual scenarios.

In addition to transfer learning approaches, significant progress has been made in the development of end-to-end Automatic Speech Recognition (ASR) models. These models aim to simplify the ASR pipeline by directly mapping input audio to text, eliminating the need for separate acoustic, pronunciation, and language models. A notable contribution in this area is the Jasper model, introduced by J. Li et al. (2019). Jasper is an end-to-end convolutional neural acoustic model that achieves state-of-the-art results on LibriSpeech among models without external language models. The authors demonstrate that deep convolutional neural networks can be highly effective for ASR tasks, challenging the notion that recurrent or transformer-based models are necessary for achieving top performance. Jasper's architecture, consisting of blocks of 1D convolutions, batch normalization, ReLU, and dropout, provides a streamlined approach to ASR that is both efficient and accurate. This work exemplifies the ongoing trend towards more integrated and efficient ASR systems, potentially offering new avenues for improving recognition in low-resource scenarios.

The choice of wav2vec2 as the model for our experiments is driven by its proven effectiveness in low-resource settings, robust architecture, and successful application in various languages. The advancements in the wav2vec framework, particularly with the introduction of wav2vec 2.0 and XLSR, provide a solid foundation for enhancing Cantonese ASR. By leveraging these state-of-the-art models, we aim to enhance the recognition performance and robustness of Cantonese ASR, contributing to the preservation and accessibility of this linguistically rich language. Wav2vec 2.0's self-supervised learning approach allows it to leverage large amounts of unlabeled data, making it particularly suitable for languages with limited transcribed resources. Its proven effectiveness in low-resource settings, robust architecture, and successful application in various languages make it an ideal choice for our experiments on Cantonese ASR.

3 methodology

In this section, I will outline the methodology used to address the research question and validate the hypothesis on a high level. First, in subsection 3.1, I will discuss the datasets utilized for training and testing the models. Next, subsection 3.2 will focus on the feature extractor employed in the models, namely wav2vec 2.0. Following that, in subsection 3.2.1, I will delve into the XLSR models and provide a comparative analysis. Subsection 3.3 will then elaborate on the evaluation method and metric employed, specifically the character error rate (CER), and justify why CER is preferred over word error rate (WER) for Cantonese. Finally, in subsection 3.4.2, I will illustrate the concrete steps of the experiments.

3.1 Dataset

3.1.1 Common Voice dataset

For this study, I utilized the Common Voice dataset developed by Mozilla. Common Voice is an open-source, multilingual voice dataset designed to address the under-representation of various languages and demographics in voice technology. The data collection process is community-driven, involving volunteers who contribute voice recordings by reading predefined sentences in their native languages. This crowd-sourced approach creates a more diverse and extensive dataset compared to most commercial datasets, and it has been particularly beneficial in gathering and organizing speech data for under-resourced languages and dialects.

As of the latest release, the Common Voice dataset comprises over 31,000 recorded hours of voice data, with more than 20,000 validated hours covering 124 languages. Each data entry consists of an MP3 audio file and its corresponding text file, accompanied by demographic metadata such as age, sex, and accent. This rich dataset allows for comprehensive analysis and training of speech recognition models, ensuring inclusivity and diversity.

Despite making it possible to collect a significant amount of speech data for low-resource languages, the quantity and quality of data still vary greatly among different languages. For example, the Common Voice dataset includes over 2,000 hours of English recordings and more than 1,000 hours of Mandarin recordings. In contrast, Cantonese only has about 177 hours, and Minnan (Hokkien) even less, with only around 22 hours. This discrepancy highlights the ongoing challenges in achieving balanced and comprehensive language representation in speech datasets.

Utilizing the Common Voice Cantonese dataset allows me to leverage both the pre-existing Mandarin ASR model and the extensive Cantonese-specific data. This approach aims to enhance the performance of the Cantonese ASR system by incorporating the linguistic similarities between Mandarin and Cantonese, thus making efficient use of available resources and addressing the challenges associated with low-resource languages.

3.2 Model Framework - wav2vec 2.0

Wav2Vec 2.0, developed by Facebook AI, introduces a novel approach to Automatic Speech Recognition (ASR) through self-supervised learning. As illustrated in Figure 1, the model architecture comprises three main components: a shared CNN encoder, a shared quantizer, and a shared Transformer encoder.

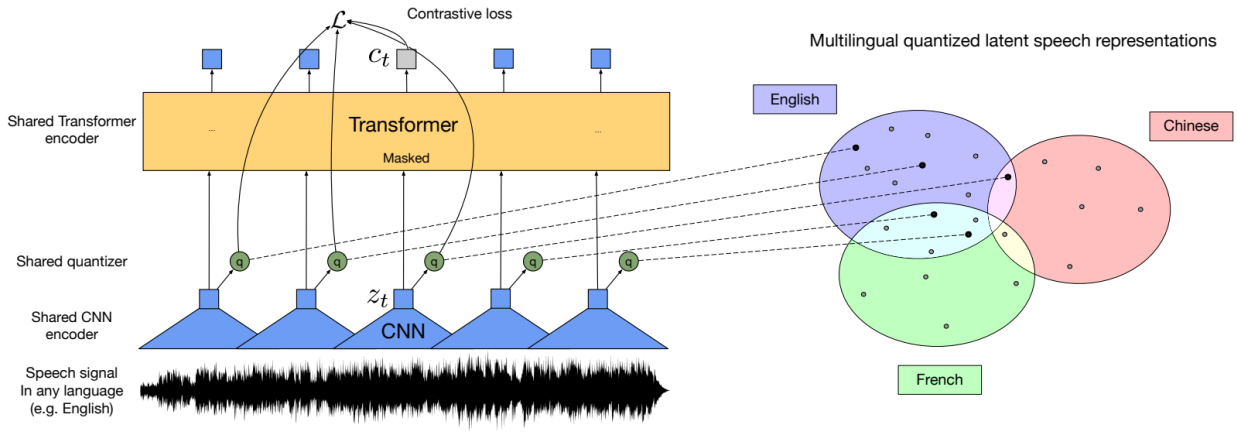

Figure 1: The model structure of wav2vec2

Shared CNN Encoder: The CNN encoder processes the raw audio waveform, extracting latent speech representations. These representations are crucial for capturing the essential features of the audio signal.

Shared Quantizer: The quantizer discretizes the latent speech representations into a finite set of learned speech units. This step helps in transforming continuous audio signals into discrete representations that are easier to handle in subsequent steps.

Shared Transformer Encoder: The Transformer encoder further processes these quantized representations to capture long-range dependencies and contextual information. are easier to handle in subsequent steps.

During pre-training, the model solves a contrastive task, which involves distinguishing true latent speech representations from distractors. This self-supervised pre-training enables Wav2Vec 2.0 to learn powerful audio representations from unlabeled audio data. These representations can then be fine-tuned with labeled data for specific ASR tasks, significantly improving the model's performance.

The Gaussian Error Linear Unit (GELU) is an activation function used in neural networks, particularly in Transformer models like Wav2Vec 2.0. The GELU activation function is defined by the formula:

$$GELU(x) = xP(X \leq x) = x\phi(x)$$

where $\Phi(x)$ is the cumulative distribution function (CDF) of the standard normal distribution.

Explanation of the Formula

- x is the input to the activation function.
- $P(X \leq x)$ represents the probability that a standard normal random variable X is less than or equal to x.
- $\Phi(x)$ is the cumulative distribution function (CDF) of the standard normal distribution.

The GELU activation function combines linear and non-linear components, preserving the properties of the input signal while introducing smooth non-linearity. This helps reduce the likelihood

of dead neurons and enhances the model's learning capabilities. By integrating the GELU activation function, models like Wav2Vec 2.0 benefit from improved convergence and performance, contributing to state-of-the-art results in speech recognition tasks.

3.2.1 Large-Scale Cross-Lingual Models - XLSR-53 & XLS-R

The XLSR (cross-lingual speech representations) model extends this approach by pre-training on multiple languages, thereby enhancing its ability to generalize across different linguistic contexts. This multilingual pre-training framework has shown promising results in cross-lingual and low-resource ASR tasks, making it a suitable candidate for fine-tuning on Cantonese speech recognition.

3.3 Evaluation - Character Error Rate

In this study, the evaluation of the Cantonese Automatic Speech Recognition (ASR) system is conducted using Character Error Rate (CER) rather than Word Error Rate (WER). The primary reason for this choice lies in the unique linguistic characteristics of Cantonese, which make CER a more suitable and accurate metric for assessing ASR performance. Character Error Rate (CER) is a common metric used to evaluate the performance of Automatic Speech Recognition (ASR) systems. CER is calculated by comparing the recognized text (hypothesis) with the reference text (ground truth) at the character level. The formula for CER takes into account the number of substitutions (S), deletions (D), and insertions (I) needed to transform the hypothesis into the reference text, divided by the total number of characters (N) in the reference text.

$$\text{CER} = \frac{S + D + I}{N} \tag{1}$$

where:

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

Cantonese, like Mandarin Chinese, does not use spaces to separate words in written text. Instead, it relies on characters that form words and phrases, which can vary significantly in length and meaning. This lack of clear word boundaries complicates the use of WER, as it is based on the accurate recognition of individual words, typically separated by spaces in languages such as English. In contrast, CER measures the accuracy at the character level, evaluating the number of character insertions, deletions, and substitutions needed to match the reference text.

Furthermore, the use of CER is particularly relevant for Cantonese ASR due to the high variability and complexity of Cantonese characters. These characters often carry significant meaning on their own, and errors at the character level can drastically alter the intended message. By focusing on CER, we ensure a more granular and precise assessment of the ASR system's performance, capturing the nuances of character recognition that are critical for accurate transcription in Cantonese.

3.4 Training Setup

In this subsection, I will provide details about the setup and configuration used for training the speech recognition models. The code for my model is publicly accessible on GitHub[1].

3.4.1 Model Configuration

The model configuration for both the baseline and transfer learning models is designed to leverage the powerful capabilities of the Wav2Vec 2.0 architecture and the XLSR model for cross-lingual speech recognition tasks.

- **Architecture:** Both models are based on the Wav2Vec 2.0 architecture, utilizing the XLSR model pre-trained on multiple languages. The architecture consists of 24 hidden layers, each with a hidden size of 1024 units and 16 attention heads. Activation functions used include GELU, with layer normalization and dropout techniques applied to improve model robustness and generalization.

- **Pre-trained Weights:** The models are initialized with pre-trained weights from the Wav2Vec 2.0 and XLSR models. For the baseline model, the pre-trained weights are from "facebookwav2vec2-large-xlsr-53". The transfer learning model uses weights fine-tuned on Mandarin data, with additional tokens added for commonly used Cantonese characters. This Mandarin-wav2vec2.0 model is pre-trained using 1000 hours of data from the AISHELL-2 dataset Du, Na, Liu, and Bu (2018). This pre-training process involves using the wav2vec 2.0 framework, which is a self-supervised learning approach allowing the model to leverage large amounts of unlabeled speech data. The detailed steps for pre-training include segmenting raw audio into discrete units and training the model to predict these units from masked segments of the audio Lu and Chen (2022). This model is then fine-tuned on 178 hours of labeled data from the AISHELL-1 dataset Bu, Du, Na, Wu, and Zheng (2017) to improve its performance for Mandarin ASR tasks[2].

- **Hyperparameters:** Key hyperparameters include a learning rate of 3×10^{-4}, batch size adjusted through gradient accumulation steps (effectively increasing the batch size without requiring additional memory), and the number of training epochs set to 40. Optimization is performed using the Adam optimizer, with a warm-up period of 500 steps to stabilize training. Dropout rates for attention and hidden layers are set to 0.1, with additional configuration for gradient checkpointing to manage memory usage during training.

3.4.2 Training Process

- **Data Loading:** The training and validation datasets are loaded from the local disk using the `load_from_disk` function. Unnecessary columns such as "accent," "age," and "up_votes" are removed to streamline the dataset. Special characters are filtered out from the text data to ensure clean input for training. The audio files are resampled to a uniform sampling rate of

[1] https://github.com/Erin-lab-design/Cantonese-ASR-Wav2vec2-XLRS-transfer-learning-project

[2] The pre-trained Mandarin-wav2vec2.0 model is publicly available at https://github.com/kehanlu/mandarin-wav2vec2

16 kHz to maintain consistency. A custom vocabulary is created from the dataset and used to tokenize the data. This preprocessing ensures that the model receives high-quality and standardized input data.

- **Optimization:** The optimization of the model is performed using the Adam optimizer with a learning rate of 3×10^{-4}. The learning rate is warmed up over the first 500 steps to allow the model to adjust gradually to the training data. Gradient accumulation is used to effectively increase the batch size without requiring additional memory, which is particularly useful given the large size of the model and the extensive dataset. These techniques help in efficiently updating the model weights and improving the training stability and convergence.

- **Validation:** The validation process includes evaluating the model's performance at regular intervals during training. Specifically, the Character Error Rate (CER) metric is computed on the validation set every 400 steps. This frequent evaluation allows for monitoring the model's progress and making necessary adjustments during training. The CER is calculated by comparing the predicted transcriptions to the ground truth labels, providing a measure of the model's accuracy at the character level. This metric is particularly suitable for the Cantonese language due to its unique linguistic characteristics.

3.5 Objective

This study aims to develop a robust ASR system for Cantonese by fine-tuning the Wav2Vec 2.0 XLSR-53 model on the Common Voice Cantonese dataset as a baseline model. Following this, transfer learning techniques are employed using an existing Mandarin ASR model, specifically the Wav2Vec 2.0 model fine-tuned on Mandarin, and applying the same Cantonese dataset for further fine-tuning. The primary objectives include:

- Evaluating the performance of the fine-tuned XLSR-53 model on Cantonese ASR tasks as the baseline.

- Investigating the impact of applying transfer learning from the Mandarin ASR model to Cantonese on model performance.

- Identifying potential challenges and proposing solutions for improving Cantonese ASR systems.

By addressing these objectives, this research seeks to contribute to the broader effort of developing effective ASR systems for low-resource languages and dialects, ultimately enhancing the accessibility of speech recognition technology to a more diverse range of speakers.

This concludes the methodology section which explains at a high-level the methods employed during this research. In the next section, the experimental setup will be presented which will include more low-level details about the dataset used and the parameters of the models.

4 Experimental Setup

4.1 Overview

In this section, I provide a detailed breakdown of the experimental setup used for our study on Cantonese Automatic Speech Recognition (ASR) using the Wav2Vec2.0 XLSR-53 model. The subsections cover the following aspects:

- **Training Setup:** This part details the baseline and transfer learning models, data splitting strategy, and dataset used for training and evaluation.
- **Experimental Setup:** Here, I discuss the hardware and software environment, including the computing resources, frameworks, and dependencies used for the experiments.
- **Training and Evaluation Process:** This subsection outlines the training procedures, evaluation methods, and performance monitoring strategies employed to assess the models' effectiveness.

Each subsection is designed to provide a thorough understanding of the methodologies and tools applied in this research, ensuring transparency and reproducibility of the results.

4.2 Training Setup

4.2.1 Baseline Model - XLSR-53 Fine-Tuning

For the baseline model, I utilize the Wav2Vec2.0 XLSR-53, pre-trained on 53 languages, and fine-tune it with the Common Voice Cantonese dataset. This step serves as the foundation for evaluating the initial performance of cross-lingual transfer learning for Cantonese ASR.

4.2.2 Transfer Learning Model - XLSR-53 Fine-Tuned on Mandarin

Next, I fine-tune the XLSR-53 model, which has been pre-trained on Mandarin, using the Common Voice Yue dataset. This model aims to leverage the linguistic similarities between Mandarin and Cantonese to improve performance.

4.2.3 Data Splitting and Subsets

The dataset used for this study is the Common Voice Cantonese dataset as of March 2024, with a total duration of approximately 177 hours.

To ensure a balanced evaluation, the dataset was split into three subsets:

- **Train Dataset:** Consisting of 80% of the total data, the training dataset contains about 142 hours of audio.
- **Dev Dataset:** The development dataset, used for validation during training, makes up 10% of the total data and contains round 17 hours of audio.
- **Test Dataset:** The test dataset, also comprising 10% of the total data, is used to evaluate the final model performance and contains round 17 hours of audio.

This splitting strategy ensures that the model is trained, validated, and tested on distinct and appropriately proportioned datasets, facilitating robust and reliable performance assessment.

4.2.4 Hardware

- **Computing Resources**: The experiments were conducted using computing resources with 64 GB of GPU memory. Jobs were submitted to a cluster with these specifications to ensure adequate processing power for the models.

4.2.5 Software Environment

- **Frameworks and Libraries:** The models were implemented using PyTorch and the Hugging Face Transformers library.
- **Dependencies:** Key dependencies include Python 3.6.8, `PyTorch` 1.10.1, and `Transformers` 4.18.0. Additional libraries used for data handling and processing include `NumPy` 1.19.5, `pandas` 1.1.5, and `SciPy` 1.5.4. The dataset management and processing were facilitated using the `datasets` library version 2.4.0. Other notable libraries include `torchaudio` 0.10.1 for audio processing, `librosa` 0.9.2 for music and audio analysis, and `pycantonese` 3.3.1 for handling Cantonese language data.
- **Operating System:** The experiments were conducted on a Linux system with kernel version 4.18.0-513.24.1.el8_9.x86_64, indicating a variant of Red Hat Enterprise Linux (RHEL) 8 or CentOS 8.

4.3 Training and Evaluation Process

4.3.1 Training Procedure

The training procedure for each model would be illustrated as follows:

Recreate the Wav2Vec2 Cantonese model: The baseline model for this study is fine-tuned from the Wav2Vec2.0 XLSR -53 model. The pre-trained model used for this process can be accessed at the huggingface[3]. The training script I used to reproduce the model is available on GitHub[4]. For the baseline model, I utilized the Wav2Vec2.0 XLSR - 53, pre-trained on 53 languages. This model was fine-tuned on the Common Voice Cantonese dataset. The fine-tuning process involved training the model over 40 epochs. The training script was executed as follows: After setting up the necessary environment and dependencies, the pre-trained XLSR-53 model was loaded and fine-tuned using the specified Cantonese dataset. The model parameters were updated through back-propagation over 40 epochs, allowing the model to learn the characteristics of Cantonese speech. This process was crucial in adapting the multilingual pre-trained model to perform well on Cantonese ASR tasks.

Transfer Learning Model - XLSR-53 Fine-Tuned on Mandarin: For the transfer learning model, I utilized a Wav2Vec2.0 XLSR-53 model that had been specifically fine-tuned on Mandarin. The pre-trained Mandarin model used as the starting point for this fine-tuning process can be found

[3]Wav2vec2 XLRS Cantonese model is on: `https://huggingface.co/ctl/wav2vec2-large-xlsr-cantonese`
[4]Training script available at: `https://github.com/chutaklee/CantoASR/blob/main`

at the huggingface main page[5]. This approach leverages the model's ability to learn general acoustic features across multiple languages, combined with the linguistic similarities between Mandarin and Cantonese, to potentially enhance ASR performance for Cantonese.

To address the challenge of recognizing Cantonese-specific characters not included in the Mandarin fine-tuned model's vocabulary, I extracted characters from Cantonese Wikipedia articles that were missing from the model's existing `vocab.json`. These additional characters include many commonly used Cantonese expressions and colloquial terms, essential for accurate Cantonese ASR.

The same Common Voice Cantonese dataset used for the baseline model was employed to fine-tune this transfer learning model, ensuring a consistent basis for comparison. The training process involved running 40 epochs, during which the model parameters were updated to learn the specific features of Cantonese speech. The best model was saved based on the lowest Character Error Rate (CER) achieved during the training process.

4.3.2 Evaluation Method

The evaluation of the models is conducted using the Character Error Rate (CER) metric, which is calculated by comparing the predicted transcriptions to the ground truth labels at the character level. This metric is particularly suitable for Cantonese due to its unique linguistic characteristics.

- **Validation Frequency:** The CER is computed on the validation set at the end of each epoch. This frequent evaluation allows for monitoring the model's progress and making necessary adjustments during training.
- **Performance Monitoring:** The performance of the models is logged using Weights & Biases (W&B), providing real-time monitoring and comparison of different runs. Detailed logs capture essential information such as training and validation metrics, system resource utilization, and any anomalies encountered during the runs.
- **Hyperparameter Tuning:** Automated hyperparameter tuning is conducted using tools such as Optuna, allowing systematic exploration of the hyperparameter space to identify optimal settings for the models.

This section provides a comprehensive overview of the training and evaluation processes for both the baseline and transfer learning models, detailing the key steps and methods used to ensure robust and reliable ASR performance.

[5]The pre-trained Mandarin model: `https://huggingface.co/jonatasgrosman/wav2vec2-large-xlsr-53-chinese-zh-cn`

5 Results & Discussion

5.1 Results

This section presents the results of our experimental setup, comparing the performance of the baseline model and the transfer learning model on Cantonese Automatic Speech Recognition (ASR). The comparison is based on three key metrics: training loss, evaluation loss, and Character Error Rate (CER).

Training Loss Comparison: The training loss for both models across 40 epochs is illustrated in Figure 2. The baseline model, fine-tuned from the Wav2Vec2.0 XLSR-53 pretrained on 53 languages, shows a higher initial training loss which decreases significantly over the epochs. The transfer learning model, fine-tuned from the Wav2Vec2.0 XLSR-53 pretrained on Mandarin, demonstrates a much lower initial training loss and converges more quickly. This indicates that the transfer learning model benefits from the linguistic similarities between Mandarin and Cantonese, leading to faster adaptation during training.

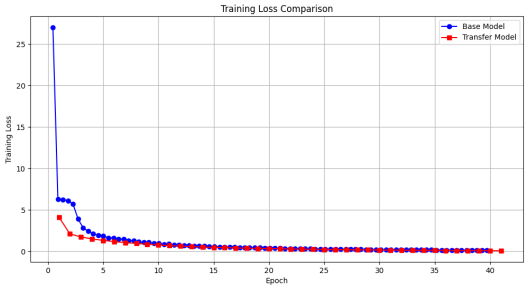

Figure 2: Training Loss Comparison between Baseline and Transfer Learning Models

Evaluation Loss Comparison: Figure 3 presents the evaluation loss for both models. Similar to the training loss, the transfer learning model achieves much lower evaluation loss compared to the baseline model at the starting point and the early stage of the training. But the baseline model converges a little bit better than the transfer learning one afterwards. This suggests that the transfer learning model trains faster but may not generalizes better to unseen data during validation compared to the baseline one.

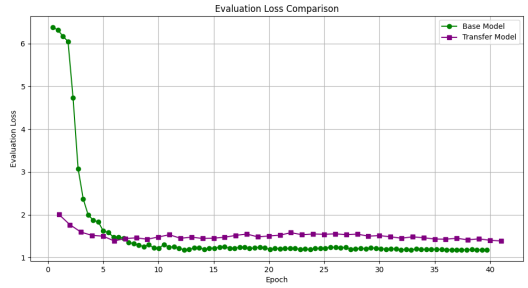

Figure 3: Evaluation Loss Comparison between Baseline and Transfer Learning Models

Evaluation Character Error Rate (CER) Comparison: The CER is a critical metric for assessing the performance of ASR models. Figure 4 shows the CER for both models over the epochs. The transfer learning model consistently outperforms the baseline model, achieving lower CER values across most epochs. And at the end of the training process the transfer learning model still get apparent lower CER results than the other, with the transfer learning model gets around 0.2 CER while the baseline model gets a CER about 0.3. This demonstrates the effectiveness of leveraging the pre-trained Mandarin model to enhance the ASR performance for Cantonese.

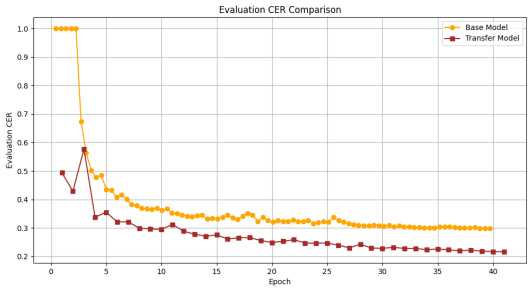

Figure 4: Evaluation CER Comparison between Baseline and Transfer Learning Models

5.2 Discussion

This finding validates our hypothesis that leveraging a pre-trained Mandarin wav2vec2 model would improve the performance of Cantonese ASR. The lower initial training loss and faster convergence observed in the transfer learning model suggest that it benefits from the linguistic similarities between Mandarin and Cantonese. This supports the notion that cross-lingual transfer learning can be particularly effective when the source and target languages share phonetic and syntactic characteristics.

The results indicate several key advantages of the transfer learning approach. The transfer learning model demonstrates much lower initial training loss and faster convergence compared to the baseline model, indicating enhanced training efficiency. Despite a slight advantage of the baseline model in validation loss during later stages of training, the transfer learning model consistently achieves better Character Error Rate (CER), suggesting a robust generalization ability to unseen data. The significant reduction in CER from around 0.3 to around 0.2 at the last epoch highlights the superior overall performance of the transfer learning model.

These findings underscore the potential benefits of cross-lingual transfer learning, particularly between linguistically similar languages. The use of a pre-trained Mandarin model provides a strong foundation for recognizing Cantonese speech, demonstrating how prior knowledge from a related language can be effectively transferred to improve ASR performance. This approach not only enhances the training process but also contributes to the development of more inclusive and diverse speech recognition systems, capable of supporting under-resourced languages and dialects. In conclusion, the study confirms that a pre-trained Mandarin wav2vec2 model, when fine-tuned with Cantonese data, significantly outperforms the baseline model pre-trained on multiple languages. This validates the effectiveness of cross-lingual transfer learning and highlights its potential in multilingual ASR tasks. The results emphasize the importance of leveraging linguistic similarities to

improve model performance and advance speech recognition technology for under-represented languages.

This integrated discussion encapsulates the results, validates the hypothesis, and highlights the implications and benefits of the research, providing a comprehensive understanding of the study's outcomes and significance.

6 Conclusion

This section provides a comprehensive illustration of the key aspects and outcomes of the study. It is divided into three main parts: Challenges, Limitations and Recommendations, and Future Work.

6.1 Challenges

Throughout the course of this study, several significant challenges were encountered, primarily related to technical capabilities and data handling. These challenges are outlined as follows:

Technical Challenges in Code and Environment Configuration: The most substantial challenge I faced was related to my coding skills, particularly in configuring the necessary environment for the experiments. The existing Python version on the server did not support some of the required packages, which necessitated the search for and implementation of alternative solutions. Managing the dependencies between various packages and Python versions proved to be exceptionally time-consuming and demanding. Resolving these compatibility issues consumed a considerable amount of effort and focus, often requiring creative problem-solving to find suitable alternatives that would not compromise the functionality of the experimental setup.

Data Collection and Processing Issues: Another significant challenge was related to the collection and processing of the dataset. Initially, the available disk space on the server was insufficient to download the entire Common Voice dataset directly. As a workaround, I resorted to using a series of .arrow files, which are essentially cache files, to load the dataset in parts. This method, while effective temporarily, was not ideal. Eventually, I managed to secure additional disk space, which allowed for the full download and proper handling of the dataset.

Challenges in Tokenization and Model Adaptation for Cantonese: When incorporating Cantonese tokens into the pre-trained Mandarin model, I faced significant challenges in the task of gathering appropriate Cantonese text proved to be more complex than initially anticipated. Unlike Mandarin, which has abundant standardized text resources, Cantonese, being primarily a spoken dialect, lacks a standardized written form and has limited digital text corpora.

The process of organizing the collected text into usable tokens was equally challenging. Cantonese, with its unique phonetic and tonal system, required careful consideration in tokenization. The technical aspects of this task stretched the limits of my coding abilities. It required scripting for web scraping to gather Cantonese text from various online sources, such as wikipedia. Data cleaning was another crucial step, involving the normalization of different written forms of Cantonese, and removing punctuations. For the tokenization process itself, I had to modify the existing tokenizer to recognize and properly handle Cantonese-specific tokens while maintaining compatibility with the pre-trained Mandarin model.

Despite these challenges, the study successfully demonstrated the potential of using a pre-trained Mandarin wav2vec2 model to improve Cantonese ASR. The experience underscored the importance of robust coding skills and adequate computational resources in conducting advanced machine learning research. Future work could focus on overcoming these technical barriers and exploring the integration of auxiliary language models to further enhance ASR performance.

6.2 Limitations and Recommendations:

A notable limitation of this study is the absence of fixed random seeds during experiments. This omission affects the reproducibility of the results, as variations in randomness can lead to different outcomes in data shuffling, weight initialization, and training processes. Consequently, it may be challenging for others to replicate the exact findings of this study. Future research should incorporate fixed random seeds to ensure consistent and reproducible results.

Due to time constraints, both models were only trained for about 40 epochs. While further training could potentially yield better results, the current performance is sufficiently robust for effective comparative analysis.

Another limitation of the study is the exclusive reliance on Character Error Rate (CER) as the evaluation metric. While CER is useful, it does not provide a complete picture of the model's performance. Other metrics, such as Word Error Rate (WER), Phoneme Error Rate (PER), and Sentence Error Rate (SER), can offer additional insights. For instance, in Cantonese speech recognition, WER can effectively gauge the overall sentence recognition accuracy, as it assesses word-level errors and thus captures practical usability issues of the ASR system. PER, on the other hand, provides a finer granularity by focusing on phonetic accuracy, which might show lower error rates if many recognized phonemes are correct, even if the resulting characters are incorrect.

Additionally, employing different methods of calculating CER, such as the Levenshtein distance, offers advantages by accounting for insertions, deletions, and substitutions needed to transform one string into another. This approach gives a more detailed measure of the similarity between the transcribed text and the reference, thereby providing a nuanced understanding of the model's accuracy.

Including a range of these metrics would yield a more comprehensive assessment of the models' performance, helping to identify specific strengths and weaknesses in different aspects of speech recognition. This multi-metric approach would facilitate a more nuanced understanding and guide more targeted improvements in future research.

Last but not least, one significant limitation I encountered in this research was the inability to successfully integrate a language model with the acoustic model, despite extensive efforts. Initially, I trained a Cantonese BERT language model with the intention of incorporating it into the post-processing stage of the ASR pipeline. This approach was motivated by my observation that while the ASR system often made accurate phonetic predictions, it frequently selected incorrect characters due to the high degree of homophony in Cantonese.

My goal was to develop an end-to-end acoustic-language joint model, combining the strengths of the wav2vec2.0-based acoustic model with the contextual understanding provided by the Cantonese BERT model. In theory, this integrated approach would allow for more holistic processing of speech input, considering both acoustic features and linguistic context simultaneously.

However, I faced significant technical challenges in implementing this joint model:

- Lack of computational resources for training joint ASR model
- Integrating the fundamentally different architectures of wav2vec2.0 and BERT models
- Developing an effective algorithm for joint inference

Despite dedicating approximately four weeks to this endeavor, I was unable to overcome these hurdles within the project's timeframe and available resources. Consequently, I had to abandon the planned end-to-end acoustic-language joint model.

If given more time, additional computational resources, and access to a wider range of Python versions on the server, I believe I could potentially complete this integration. Extended time would allow for more thorough exploration of different integration techniques and optimization strategies. Greater computational resources would enable more extensive experimentation with model architectures and hyperparameters. Access to various Python versions would provide flexibility in using different libraries and frameworks that might be crucial for successful integration. Furthermore, with these additional resources, I could explore more sophisticated approaches such as:

- Developing custom layers or modules to bridge the gap between acoustic and language models
- Implementing more advanced training techniques to better align the two models
- Experimenting with alternative language model architectures that might be more compatible with the acoustic model
- Investigating dynamic fusion techniques that adaptively balance the contributions of acoustic and linguistic information

While the implementation of this joint model remains an aspirational goal, the insights gained from this attempt have deepened my understanding of the challenges in integrating acoustic and linguistic processing in ASR systems. Future work in this area could build upon these lessons, potentially leading to more effective and efficient ASR systems for Cantonese and other languages with complex phonological characteristics.

6.3 Future Work

Future work will explore several promising directions to enhance ASR systems for under-resourced languages. This includes investigating the efficacy of transfer learning for Hokkien ASR using a Mandarin pre-trained model, integrating task-specific language models to improve performance in specialized environments, and developing ASR systems capable of handling code-switching and dialect recognition. These efforts aim to create more versatile, accurate, and user-friendly ASR systems, promoting linguistic diversity and inclusion in speech technology.

Transfer Learning for Hokkien ASR: One potential direction is to investigate the efficacy of transfer learning using the wav2vec2 model pre-trained on Mandarin for Hokkien, another low-resource language. Both Cantonese and Hokkien exhibit significant linguistic similarities with Mandarin, particularly in their phonetic systems and vocabulary. These similarities can be exploited to facilitate more effective transfer learning. For instance, both Cantonese and Hokkien have tones and phonemic structures that, while distinct, are sufficiently similar to Mandarin to benefit from a shared pre-training phase. By understanding these linguistic relationships, we can better tailor transfer learning techniques to maximize performance improvements. Exploring how the pre-trained Mandarin model can be adapted for Hokkien with a smaller dataset will provide insights into the scalability and versatility of transfer learning approaches for different Chinese dialects.

Task-Specific Language Models: Another area for future research is the integration of task-specific language models, such as BERT, to enhance ASR performance in specialized environments. For example, in automotive settings where there is a demand for robust Cantonese and other low-resource language ASR systems, incorporating a language model fine-tuned for such specific tasks can significantly improve recognition accuracy. This approach leverages contextual understanding and domain-specific vocabulary, making the ASR system more reliable in practical applications.

Handling Code-Switching and Dialect Recognition: In many regions where low-resource languages like Cantonese are spoken, there is often a prevalence of code-switching with dominant languages such as Mandarin. Addressing this phenomenon requires the development of ASR systems capable of recognizing and seamlessly switching between languages and dialects. Future work should focus on creating models that can detect and adapt to code-switching, ensuring accurate recognition and improving user convenience. Additionally, incorporating dialect recognition capabilities will further enhance the system's usability in multilingual environments.

By addressing these areas, future research can contribute to the development of more versatile, accurate, and user-friendly ASR systems for under-resourced languages, ultimately promoting linguistic diversity and inclusion in speech technology.

References

Amodei, D., et al. (2015). *Deep speech 2: End-to-end speech recognition in english and mandarin* (Tech. Rep.). Baidu Research – Silicon Valley AI Lab. Retrieved from https://arxiv.org/abs/1512.02595 (Retrieved from the provided PDF document)

Ardila, R., Branson, M., Davis, K., Kohler, M., Meyer, J., Henretty, M., ... Weber, G. (2020, May). Common voice: A massively-multilingual speech corpus. In *Proceedings of the twelfth language resources and evaluation conference* (pp. 4218–4222). Marseille, France: European Language Resources Association. Retrieved from https://aclanthology.org/2020.lrec-1.520

Baevski, A., Zhou, H., Mohamed, A., & Auli, M. (2020a). *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

Baevski, A., Zhou, H., Mohamed, A., & Auli, M. (2020b). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, *33*, 12449–12460.

Bartelds, M., San, N., McDonnell, B., Jurafsky, D., & Wieling, M. (2023). Making more of little data: Improving low-resource automatic speech recognition using data augmentation. *arXiv preprint arXiv:2305.10951*.

Bartelds, M., & Wieling, M. (2022). Quantifying language variation acoustically with few resources. In *Proceedings of the 2022 conference of the north american chapter of the association for computational linguistics: Human language technologies* (pp. 3735–3741).

Bauer, R. S. (2016). The hong kong cantonese language: Current features and future prospects. *Global Chinese*, *2*(2), 115-161. Retrieved from https://doi.org/10.1515/glochi-2016-0007 doi: 10.1515/glochi-2016-0007

Besacier, L., Barnard, E., Karpov, A., & Schultz, T. (2014). Automatic speech recognition for under-resourced languages: A survey. *Speech Communication*, *56*, 85-100. doi: 10.1016/j.specom.2013.07.008

Bolton, K. (2024). Language policy and planning in hong kong: Colonial and post-colonial perspectives. *Journal of Asian Pacific Communication*, *34*(2), 115-161. doi: 10.1075/japc.34.2.03bol

Bu, H., Du, J., Na, X., Wu, B., & Zheng, H. (2017). Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline. In *2017 20th conference of the oriental chapter of the international coordinating committee on speech databases and speech i/o systems and assessment (o-cocosda)* (pp. 1–5).

Diwan, A., Vaideeswaran, R., Shah, S., Singh, A., Raghavan, S., Khare, S., ... others (2021). Multilingual and code-switching asr challenges for low resource indian languages. *arXiv preprint arXiv:2104.00235*.

Du, J., Na, X., Liu, X., & Bu, H. (2018). Aishell-2: Transforming mandarin asr research into industrial scale. *arXiv preprint arXiv:1808.10583*.

Ethnologue. (2024). *Yue chinese.* Retrieved from https://www.ethnologue.com/language/yue/ (Accessed: 2024-05-21)

Gupta, V., & Boulianne, G. (2022). Progress in multilingual speech recognition for low resource languages kurmanji kurdish, cree and inuktut. In *Proceedings of the 13th conference on language resources and evaluation (lrec 2022)* (pp. 6420–6428).

Hinton, G., Deng, L., Yu, D., Dahl, G., rahman Mohamed, A., Jaitly, N., ... Kingsbury, B. (2012).

Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal Processing Magazine*, *29*(6), 82-97. doi: 10.1109/MSP.2012.2205597

Krauss, M. (1992). The world's languages in crisis. *Language*, *68*(1), 4–10.

Kwon, Y., & Chung, S.-W. (2023). Mole: Mixture of language experts for multi-lingual automatic speech recognition. In *Icassp 2023 - 2023 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 1–5). Rhodes Island, Greece. doi: 10.1109/ICASSP49357.2023.10096227

Lewis, M. P., Simons, G. F., & Fennig, C. D. (2013). *Ethnologue: Languages of the world*. SIL international Dallas, TX.

Li, J., Lavrukhin, V., Ginsburg, B., Leary, R., Kuchaiev, O., Cohen, J. M., ... Gadde, R. T. (2019). Jasper: An end-to-end convolutional neural acoustic model. *ArXiv*, *abs/1904.03288*. Retrieved from https://api.semanticscholar.org/CorpusID:102352277

Li, Q., Mai, Q., Wang, M., et al. (2024). Chinese dialect speech recognition: a comprehensive survey. *Artificial Intelligence Review*, *57*(25). Retrieved from https://doi.org/10.1007/s10462-023-10668-0 doi: 10.1007/s10462-023-10668-0

Li, Y., Kang, Y., Ding, D., & Zhang, N. (2022). An overview of the "protecting cantonese movement" in guangzhou (2010–2021). *Asian-Pacific Journal of Second and Foreign Language Education*, *7*(1), 36. Retrieved from https://doi.org/10.1186/s40862-022-00165-2 (Epub 2022 Sep 15) doi: 10.1186/s40862-022-00165-2

Lu, K.-H., & Chen, K.-Y. (2022). A context-aware knowledge transferring strategy for ctc-based asr. *arXiv preprint arXiv:2210.06244*.

Meeus, Q., Moens, M.-F., & Van hamme, H. (2022). Multitask learning for low resource spoken language understanding. *arXiv preprint arXiv:2211.13703*.

Pratap, V., Sriram, A., Tomasello, P., Hannun, A. Y., Liptchinsky, V., Synnaeve, G., & Collobert, R. (2020). Massively multilingual asr: 50 languages, 1 model, 1 billion parameters. *ArXiv*, *abs/2007.03001*. Retrieved from https://api.semanticscholar.org/CorpusID:220380826

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022a). Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022b). Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*.

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

Simons, G. F., & Lewis, M. P. (2013). The world's languages in crisis: A 20-year update. In E. Mihas, B. Perley, G. Rei-Doval, & K. Wheatley (Eds.), *Responses to language endangerment. in honor of mickey noonan* (Vol. Studies in Language Companion Series 142, pp. 3–19). Amsterdam: John Benjamins. (Paper presented at the 26th Linguistics Symposium: Language Death, Endangerment, Documentation, and Revitalization, University of Wisconsin, Milwaukee, 20–22 October 2011. Final revision for proceedings volume: 27 Aug 2013)

von Platen, P. (2021, Nov). *Fine-tune XLSR-WAV2VEC2 for low-resource ASR with Huggingface transformers.* Retrieved from https://huggingface.co/blog/fine-tune-xlsr-wav2vec2

Wang, X., Long, Y., Li, Y., & Wei, H. (2023). Multi-pass training and cross-information fusion for low-resource end-to-end accented speech recognition. *arXiv preprint arXiv:2306.11309*.

Yu, T., Frieske, R., Xu, P., Cahyawijaya, S., Yiu, C. T., Lovenia, H., ... Fung, P. (2022,

June). Automatic speech recognition datasets in Cantonese: A survey and new dataset. In *Proceedings of the thirteenth language resources and evaluation conference* (pp. 6487–6494). Marseille, France: European Language Resources Association. Retrieved from https://aclanthology.org/2022.lrec-1.696