university of groningen

campus fryslân

Enhancing Whisper's Zero-Shot Capabilities for Code-Switching through Fine-Tuning

Haolin Miao

university of groningen

campus fryslân

University of Groningen - Campus Fryslân

Enhancing Whisper’s Zero-Shot Capabilities for Code-Switching through Fine-Tuning

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. J.K. Schäuble (Voice Technology, University of Groningen)
with the second reader being
Supervisor 2’s title and name (Voice Technology, University of Groningen)

Haolin Miao (S5880432)

June 11, 2025

Acknowledgements

I would like to express my sincere gratitude to Dr. J.K. Schäuble for all his efforts in guiding me to complete this paper.

I acknowledge the Center for Information Technology of the University of Groningen for their technical support and for providing access to the Hábrók high-performance computing cluster.

Lastly, I would like to express my gratitude to all the individuals who have played a part, no matter how small, in shaping my academic journey and the successful completion of this thesis.

Abstract

Automatic Speech Recognition (ASR) for code-switched speech, particularly involving dialects like Cantonese mixed with English, remains a significant challenge for pre-trained models. This study investigates the efficacy of fine-tuning as a domain adaptation strategy for OpenAI's Whisper models on this task. A comparative analysis was conducted by fine-tuning two distinct models, whisper-small and whisper-large-v3, on a Cantonese-English code-switching dataset and evaluating their performance against their respective zero-shot baselines.

The experiments were performed on the MCE dataset, with Word Error Rate (WER) and Character Error Rate (CER) as the primary evaluation metrics. The results demonstrate that fine-tuning yields substantial performance improvements for both models. The whisper-small model, in particular, showed a reduction in error rates, achieving a significant drop in WER and an improvement in CER.

Furthermore, this study reveals a relationship between model scale and task-specific performance. While the whisper-large-v3 model also improved upon its zero-shot baseline, its final word-level accuracy did not surpass that of the fine-tuned small model. This outcome suggests that the larger model was prone to overfitting on the medium-resource dataset, learning surface-level patterns without generalizing effectively. The conclusion is that for specialized ASR tasks such as Cantonese-English code-switching, a smaller, more constrained model can offer a more effective pathway to achieving robust performance, highlighting the critical trade-off between model capacity and generalization.

Keywords: Automatic Speech Recognition (ASR), Code-Switching, Whisper Model, Fine-Tuning

Contents

1 Introduction

Speech recognition technology has made remarkable progress, with today's systems matching human-level accuracy for single-language speech. This advancement largely stems from moving away from older approaches like Hidden Markov Models and Gaussian Mixture Models toward complex neural networks that process speech from start to finish, particularly those built on Transformer architecture (Radford et al., 2023) . Nevertheless, a considerable challenge remains in the accurate transcription of code-switching,which is the practice of alternating between two or more languages within a single conversation or utterance (Mustafa et al., 2022). This linguistic behavior is particularly common in bilingual communities, including Cantonese-English speakers in Hong Kong and other nearby regions. Automatic speech recognition(ASR) systems, which are mainly designed and trained using monolingual data, face significant difficulties when processing code-switched speech. This often results in a marked decrease in accuracy due to heightened phonetic variability and complex linguistic interactions (Mustafa et al., 2022). Because most speech recognition research focuses on single-language data—which is easier to find and work with—we've ended up with systems and training methods that struggle to handle the unpredictable nature of language mixing. This technical gap has real-world consequences: when speech recognition fails with mixed languages, it limits how useful voice technology becomes for people who naturally switch between languages in their daily conversations (Xie et al., 2025).

Code-switching becomes particularly complex when it involves Cantonese and English, creating unique challenges in pronunciation, vocabulary, and grammar that trip up speech recognition systems (J. Y. C. Chan et al., 2009). A key difficulty stems from the notable pronunciation differences of English words when spoken with a Cantonese accent, which can include phone insertions, deletions, or the substitution of English phonemes not present in Cantonese with phonetically similar Cantonese ones. English words mixed into Cantonese sentences often get their pronunciation adjusted to fit. The fundamental linguistic differences between Cantonese (a tonal Sinitic language with a unique phonological system) and English (a Germanic language with different stress patterns and syllable structures) contribute to a more complex code-switching environment compared to language pairs with greater typological similarity (J. Y. C. Chan et al., 2009).

As a result, speech recognition systems need to master not just two separate languages, but also understand the complex ways speakers naturally blend them together when switching back and forth. Making these language challenges even worse is the serious shortage of high-quality, large-scale datasets containing natural Cantonese-English mixed speech and text(Chan et al., 2009; Xie et al., 2023). While existing datasets for general code-switching (often centered on Mandarin-English) already show limitations in size, spontaneity, and accessibility (Mustafa et al., 2022) 1, this data shortage is even more pronounced for Cantonese-English. This lack of sufficient training data creates a major obstacle for building effective speech and language models that can handle the complexities of Cantonese-English mixing (J. Y. C. Chan et al., 2009).

To address this need for better, more flexible speech recognition, researchers have developed large pre-trained models, with OpenAI's Whisper standing out as a major breakthrough. Architected as an encoder-decoder Transformer, Whisper is distinguished by its training on an extensive and varied dataset of 680,000 hours of multilingual and multitask supervised data.1 This approach of training on massive amounts of loosely labeled data gives Whisper good resilience and adaptability across different audio conditions and datasets. It can handle multiple languages and tasks like transcription and translation right out of the box, without needing specific training for each one (Radford et al., 2023). Although Whisper's multilingual foundation offers a solid starting point, its out-of-the-box performance may still require further adaptation for the specific complexities of code-switching pairs. This is because code-switching involves rapid alternations that may not be fully captured by models whose final output is typically monolingual ((Xie et al., 2025).

Recognizing that even highly proficient general-purpose models like Whisper might need refinement for specialized applications such as Cantonese-English code-switching, researchers have investigated targeted fine-tuning. A significant contribution in this domain is the study by (Xie et al., 2023), ”Whisper-mce: Whisper model finetuned for better performance with mixed languages.” This research addresses the challenge of improving ASR performance for mixed Cantonese and English speech. The authors fine-tuned the Whisper-small model using a self-collated Mixed Cantonese and English (MCE) audio dataset, which was specifically created to mitigate the lack of appropriate training data (Xie et al., 2023) . The resultant fine-tuned model, named Whisper-mce, showed enhanced capabilities in accurately transcribing mixed-language audio when compared to the baseline Whisper model for this particular task. This result highlights the importance of targeted fine-tuning for specialized ASR applications and underscores that the availability of relevant, in-domain data is crucial for progress (Xie et al., 2023).

Developing specialized models like Whisper-MCE marks real progress in this area. However, the challenges of Cantonese-English ASR, and code-switching ASR more broadly, are not entirely overcome. Researchers continues to focus on creative ways to expand training data, improve fine-tuning methods, and create better evaluation tools specifically designed for mixed-language speech (Mustafa et al., 2022; Xie et al., 2023, 2025). This study aims to extend this line of inquiry by examining the capabilities of the more recent Whisper large-v3 model. This is particularly relevant as (Xie et al., 2023) employed the Whisper-large-v2 model as their baseline for comparison with their fine-tuned Whisper-small model , whereas this research investigates the potential of the large-v3 iteration to further enhance Cantonese-English code-switching ASR.

1.1 Research Questions and Hypotheses

This thesis addresses the following research question and hypothesis:

> **[Research Question: To what extent does fine-tuning Whisper-small and Whisper-large-v3 on Cantonese-English code-mixed speech reduce WER and CER, compared to their zero-shot performance on the same data?]**

[Hypothesis: Hypothesizing that fine-tuning Whisper-small and Whisper-large-v3 on Cantonese-English code-mixed speech will lead to a statistically significant reduction in Word Error Rate (WER) and Character Error Rate (CER), compared to their respective zero-shot performance on the same dataset. Furthermore, expecting Whisper-large-v3 to exhibit a greater degree of improvement due to its larger capacity and more recent architecture.]

2 Literature Review

This literature review was constructed based on a systematic search methodology to ensure comprehensive coverage and replicability. The process, conducted between April and June 2025, involved approach to identify, screen, and select relevant scholarly literature.

Search Strategy: Databases and Search Engines: The primary academic databases searched were Google Scholar, Scopus, IEEE Xplore, the ACL Anthology, and the arXiv preprint server.

Keywords and Boolean Operators: The search employed combinations of keywords using Boolean operators (AND, OR). Key search strings included: (”Whisper model” OR ”OpenAI Whisper”) AND (”fine-tuning” OR ”adaptation”) AND (”code-switching” OR ”mixed-language”) AND (”Cantonese”), as well as broader terms like ”end-to-end ASR” AND ”code-switching” and ”low-resource ASR”.

Time Frame: The search was primarily limited to literature published between 2017 and 2024 to focus on modern deep learning paradigms, especially those related to the Transformer architecture. Foundational pre-2017 literature was included only if deemed essential for historical context or core linguistic theories.

Screening Process: The initial search yielded over 150 records. These were first screened by title and abstract for relevance. Following this, the full texts of the remaining articles were assessed against the inclusion and exclusion criteria. This process resulted in the final selection of literature cited in this review. Zotero was used for reference management and duplicate removal, while a spreadsheet was used to track the screening process.

To qualify for inclusion in this review, publications needed to satisfy several key requirements. They must be peer-reviewed journal articles, complete papers from prominent conferences such as ICASSP, Interspeech, or NeurIPS, or influential arXiv preprints with significant academic impact. All selected works are written in English and directly examine the core research areas of interest: automatic speech recognition for code-switching scenarios, the Whisper model architecture and applications, Cantonese linguistic analysis, or fine-tuning methodologies for speech processing tasks. Priority was given to studies that provided empirical findings through experimental data or comprehensive model evaluations.

Studies were excluded from this review under several circumstances. Publications focusing on topics outside the scope of this research were removed, such as those examining code-switching purely from sociolinguistic perspectives without any ASR components. Non-academic materials including blog posts, news articles, and forum discussions did not meet the scholarly standards required for inclusion. Works classified as non-original research contributions, including editorials, book reviews, and commentary pieces, were also excluded. Finally, any publication for which the complete text remained inaccessible was removed from consideration.

Automatic Speech Recognition has transformed from a specialized research field into a common technology. It has significantly changed how people interact with computers in many applications, such as virtual assistants, voice-controlled devices, and automated transcription services (Juang & Rabiner, 2005) . This development is the result of decades of innovation, especially with advances in signal processing, statistical modeling, and, more recently, deep learning methods (Hinton et al., 2012). Early ASR systems were limited to controlled settings and simple tasks, like recognizing spoken digits (Juang & Rabiner, 2005) . As computing power grew and large datasets became available, ASR systems like Dragon Dictate became available for personal computers. Later, internet-connected devices and mobile technology, including Google Voice Search and Apple's Siri, made ASR accessible to millions worldwide (Juang & Rabiner, 2005) . This widespread access and the demand for more natural interaction mean that ASR systems now need to handle the diversity of human language, including multilingualism and code-switching.

While ASR has made great progress for widely spoken languages like English, it still faces significant challenges in multilingual contexts, especially with code-switching (Gumperz, 1982; Xie et al., 2023). Code-switching, where speakers alternate between languages or dialects within a single conversation or even a sentence, is a common way people communicate in many bilingual communities, including Cantonese-English speakers in Hong Kong. This creates problems for speech recognition systems because sounds change at language switches, borrowed words get pronounced in different ways, and sentence structures get mixed together (J. Y. C. Chan et al., 2009). ASR systems are typically trained on single-language data and assume speech will be in one language (Xie et al., 2023) . For ASR to be truly useful globally, it must adapt to how people naturally use language, which includes recognizing code-switching. This requires more diverse datasets and new model designs.

Large-scale, pre-trained multilingual models, such as OpenAI's Whisper (Radford et al., 2023) , offer promising ways to address these challenges. These models show impressive abilities on many languages without specific training, largely because they were trained on vast amounts of data (Pratap et al., 2023) . However, it is important to study their performance and adaptability when fine-tuned for specific, often low-resource, code-switched language pairs like Mixed Cantonese-English (MCE). This review will first outline the development of ASR technology. Then, it will discuss the unique linguistic challenges of Cantonese speech recognition and MCE code-switching. Finally, it will explain the Whisper model's architecture and previous research on fine-tuning it for difficult speech tasks. This background will provide the context for the current study.

2.1 The Trajectory of Automatic Speech Recognition

ASR technology has advanced considerably, moving from basic pattern matching techniques to sophisticated deep learning systems.

2.1.1 Foundational Approaches and Statistical Models

Early ASR systems, developed in the mid-20th century, used simple pattern matching. For instance, Bell Labs' "Audrey" system in 1952 could recognize spoken digits from a single speaker by matching sound features to stored templates (Juang & Rabiner, 2005). A major shift occurred in the 1970s and 1980s with the adoption of statistical methods, particularly Hidden Markov Models (HMMs)

(Juang & Rabiner, 2005). HMMs allowed systems to handle larger vocabularies and continuous speech by modeling the probability of word sequences, and they became the dominant ASR technology for many years (Hinton et al., 2012). Looking at speech recognition's development, there's a clear pattern: fresh approaches improve performance, usually backed by stronger computers and larger datasets. It can be seen this same trend with current models like Whisper (Radford et al., 2023).

2.1.2 The Deep Learning Paradigm Shift: End-to-End Architectures

Traditional HMM-based systems had separate components for acoustic modeling, pronunciation, and language modeling, which could lead to less than optimal overall performance (Hinton et al., 2012). The next major change in ASR came with the rise of neural networks, driven by more powerful GPUs and large datasets. Initially, Deep Neural Networks (DNNs) were used to improve parts of HMM systems. Soon after, researchers began developing End-to-End (E2E) models. These models aim to train a single neural network that directly converts audio features into text (Hinton et al., 2012; Prabhavalkar et al., 2023; Pratap et al., 2023).

Key E2E architectures include Connectionist Temporal Classification (CTC), which handles the alignment between audio and text without needing pre-aligned data (Graves et al., 2006). Attention-based Encoder-Decoder (AED) models, such as Listen, Attend, and Spell (LAS) (W. Chan et al., 2016), use an attention mechanism that allows the model to focus on relevant parts of the audio when generating text. Recurrent Neural Network Transducers (RNN-T) are well-suited for streaming ASR, as they can process audio and produce text incrementally (Zeyer et al., 2021). The Transformer architecture (Vaswani et al., 2017), with its self-attention mechanism, proved highly effective for capturing long-range patterns in speech and has become a cornerstone of many leading ASR systems, including Whisper (Gulati et al., 2020; Radford et al., 2023). Moving to end-to-end models allows systems to learn straight from the data itself, cutting down on the manual work of designing specific features (Hinton et al., 2012).

2.1.3 The Era of Large-Scale Pre-trained Models: OpenAI's Whisper

The success of deep learning models, especially Transformers, is often linked to the amount of training data and the size of the model (Gu et al., 2023; Kaplan et al., 2020). A common strategy in AI is to pre-train very large models on massive, diverse datasets and then fine-tune them for specific tasks. OpenAI's Whisper is a prominent example of this in ASR. Whisper is a Transformer-based model trained on 680,000 hours of multilingual and multitask audio data collected from the web. This dataset included nearly 100 languages and various tasks, such as speech transcription and translation into English (Pratap et al., 2023; Radford et al., 2023). This extensive pre-training allows Whisper to generalize well and perform impressively on many languages and tasks without specific fine-tuning. This approach is similar to trends in Natural Language Processing with models like BERT and GPT, showing how different fields in AI can learn from each other (Kaplan et al., 2020; Vaswani et al., 2017).

2.2 The Intricacies of Cantonese Speech Recognition

While general ASR models have improved significantly, specific languages like Cantonese present unique characteristics that make their recognition challenging.

2.2.1 Phonetic and Tonal Uniqueness

Cantonese uses tone to convey meaning, so changing the pitch of a syllable can completely change what the word means. It typically has nine distinct tones. For ASR systems, accurately distinguishing these subtle pitch differences is a major challenge, as errors in tone recognition can lead to incorrect word identification and meaning (J. Y. C. Chan et al., 2009). Beyond its tonal system, Cantonese has a unique set of consonants and vowels, including final stop sounds like p, t, k. ASR models need to accurately represent these specific Cantonese sounds and how they interact.

2.2.2 Dialectal Diversity and Colloquial Speech

Cantonese also varies considerably across different regions where it is spoken, such as Hong Kong and Guangzhou. Pronunciation, lexical, and grammatical variation across regions challenges the generalizability of ASR models trained on a single Cantonese dialect. Furthermore, spoken Cantonese is rich in colloquial expressions and particles that are common in daily conversation but may not have standard written forms or may differ greatly from formal written Chinese (J. Y. C. Chan et al., 2009; Gumperz, 1982). Without a standard way to write everyday spoken Cantonese, it becomes much harder to collect training data and build reliable language models for speech recognition (J. Y. C. Chan et al., 2009). The combination of tonal variation, phonetic distinctiveness, dialect diversity, and limited data makes Cantonese ASR especially challenging.

2.2.3 Resource Scarcity and its Implications

A significant hurdle for Cantonese ASR is the limited availability of large, accurately transcribed speech datasets (J. Y. C. Chan et al., 2009; Xie et al., 2023). Compared to high-resource languages like English or Mandarin, Cantonese is considered a low-resource language in ASR development. This data shortage directly affects how well it can be trained effective speech recognition models, particularly today's deep learning systems that need massive amounts of varied training examples. The difficulty in creating consistent text corpora for colloquial Cantonese further complicates the problem (J. Y. C. Chan et al., 2009). Improving Cantonese ASR may depend on better data, tailored augmentation strategies, and models optimized for low-resource settings.

2.3 Code-Switching: A Multilingual Reality and ASR Challenge

In many parts of the world, people regularly use more than one language. This often involves code-switching, which presents unique difficulties for speech technology.

2.3.1 Understanding Code-Switching in Spoken Language

Code-switching (CS) is the practice of alternating between two or more languages or dialects within a single conversation, or even within a single sentence. This is not random but a rule-governed linguistic skill used by bilingual individuals for various communicative purposes, such as expressing nuanced meanings or signaling identity (Gumperz, 1982; Xie et al., 2023). It is a common feature of communication in many multilingual communities.

2.3.2 Cantonese-English Code-Switching: Prevalence and Characteristics

Hong Kong is a prime example of a society where Cantonese-English code-switching is widespread (J. Y. C. Chan et al., 2009; Gumperz, 1982). It is common to hear English words or phrases mixed into Cantonese sentences. In these cases, Cantonese usually provides the main grammatical structure, while English contributes lexical items (J. Y. C. Chan et al., 2009). A key characteristic is the phonetic adaptation of English words; Cantonese speakers often pronounce these words with a Cantonese accent, leading to systematic changes in English sounds to fit Cantonese pronunciation patterns. These changes follow predictable patterns, suggesting that speech recognition systems might be able to learn and adapt to them.

2.4 Leveraging Whisper for Mixed Cantonese-English ASR

OpenAI's Whisper models, with their strong multilingual capabilities from large-scale pre-training, offer a promising foundation for addressing the challenges of Mixed Cantonese-English (MCE) ASR.

2.4.1 Architectural Overview of Whisper Models (Transformer Encoder-Decoder)

Whisper models are based on the Transformer architecture, which is highly effective for tasks that map sequences to other sequences (Radford et al., 2023; Vaswani et al., 2017). They use an encoder-decoder structure. Input audio is divided into 30-second segments and converted into a log-Mel spectrogram. The Transformer encoder processes this spectrogram to create a rich internal representation of the audio. The Transformer decoder then takes these representations and generates the transcript one token (a small unit of text) at a time, using an attention mechanism to focus on the most relevant parts of the encoded audio (Radford et al., 2023; Vaswani et al., 2017). Whisper was trained on a massive dataset of 680,000 hours of audio covering nearly 100 languages. This training included tasks like multilingual speech transcription and translating speech from various languages directly into English (Pratap et al., 2023; Radford et al., 2023). Its broad pretraining helps Whisper perform reliably across unfamiliar languages and tasks.

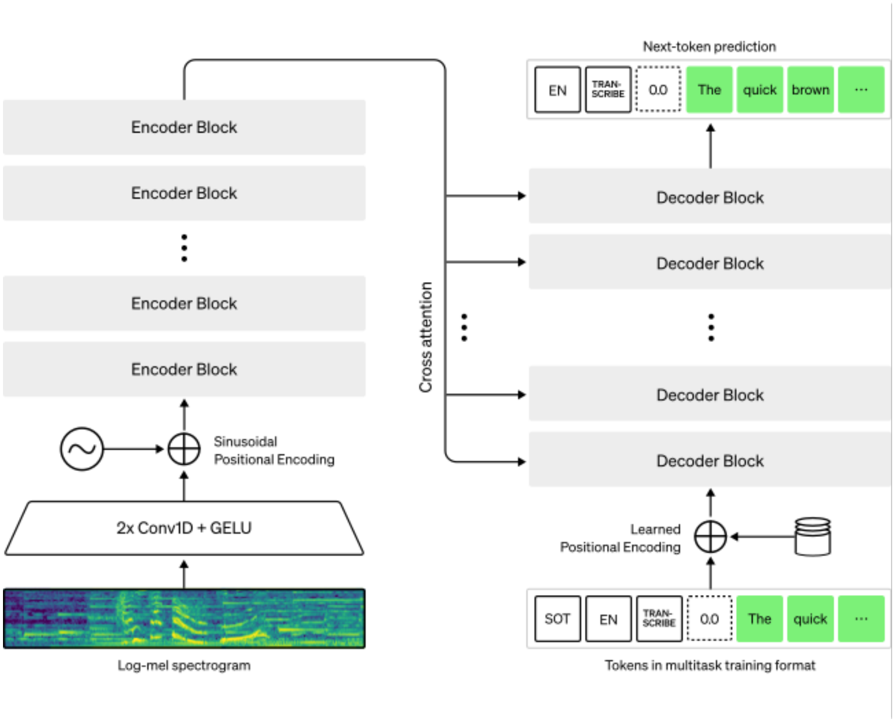


Figure 1: [Whisper model architecture]

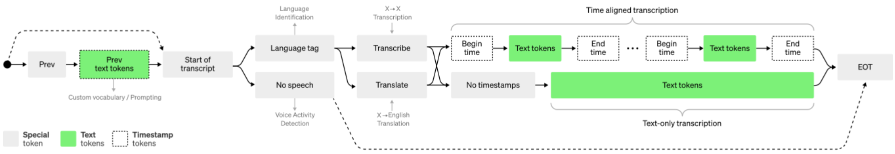
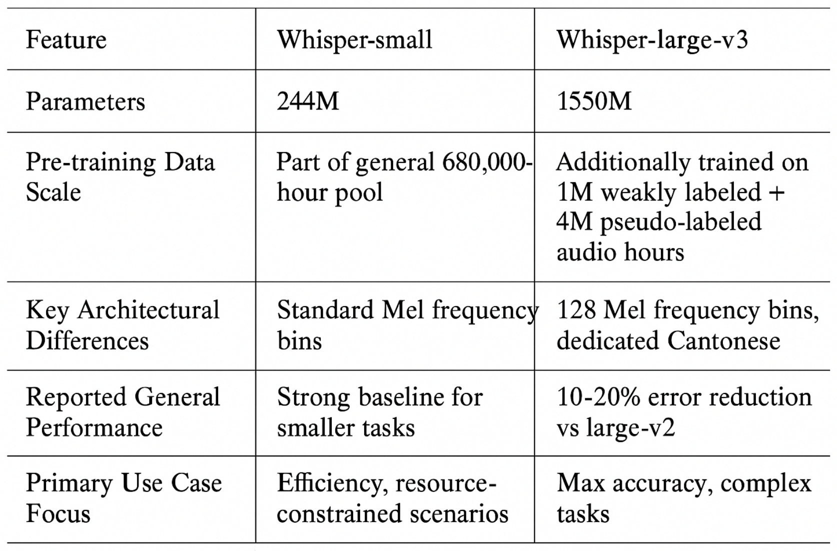


Figure 2: [Whisper decoding process]

2.4.2 Whisper Model Scaling: From Small to Large-v3

OpenAI has released Whisper in several sizes, from tiny (39 million parameters) to large (1.55 billion parameters), which includes the small model (244M parameters). The large model has undergone improvements, leading to versions like large-v3. Bigger models with more parameters typically deliver better accuracy on tough tasks, but they demand more computing power and run slower. This study focuses on comparing whisper-small (244M parameters) and whisper-large-v3 (1.55B parameters). Whisper-large-v3 was trained on an even larger dataset than previous large versions and includes some architectural changes, such as using 128 Mel frequency bins for input and adding a specific language token for Cantonese. OpenAI reports that large-v3 shows a 10-20% reduction in errors compared to large-v2 across many languages. Choosing a model size involves balancing desired accuracy with acceptable speed and available resources. The dedicated Cantonese token in large-v3 might give it an advantage on MCE data even before fine-tuning (Gu et al., 2023; Pratap et al., 2023; Radford et al., 2023).



Feature	Whisper-small	Whisper-large-v3
Parameters	244M	1550M
Pre-training Data Scale	Part of general 680,000-hour pool	Additionally trained on 1M weakly labeled + 4M pseudo-labeled audio hours
Key Architectural Differences	Standard Mel frequency bins	128 Mel frequency bins, dedicated Cantonese
Reported General Performance	Strong baseline for smaller tasks	10-20% error reduction vs large-v2
Primary Use Case Focus	Efficiency, resource-constrained scenarios	Max accuracy, complex tasks

Figure 3: [Whisper-small vs Whisper-large-v3]

2.4.3 The MCE dataset: A Resource for Cantonese-English Code-Switching Research

A major challenge in developing ASR for specific language pairs and phenomena like code-switching is the lack of suitable public datasets. To address this for MCE, Xie et al. (2023) introduced the MCE dataset. This dataset contains 34.8 hours of high-quality audio featuring natural Cantonese-English mixing across 18 diverse daily life topics. The transcriptions include 307,540 Chinese characters and 70,132 English words (Xie et al., 2023). The authors used a Multi-Agent Data Generation Framework (MADGF), potentially leveraging large language models like GPT-4, to create these naturalistic MCE speech scenarios. The MCE dataset is publicly available on GitHub and serves as an important resource for MCE ASR research (Xie et al., 2023).

2.4.4 Prior Work on Fine-Tuning Whisper for Code-Switching and Low-Resource Languages

Fine-tuning pre-trained Whisper models is a common and effective strategy for adapting them to specific ASR tasks, especially for low-resource languages or complex conditions like code-switched speech (Xie et al., 2023). Many studies have shown that fine-tuning Whisper, even with relatively small amounts of target data, can significantly reduce error rates and improve accuracy (Graves et al., 2006). For example, research on Yoruba-English code-switching has shown that fine-tuned monolingual ASR models can be competitive with larger multilingual models, particularly when computational efficiency is important. While Whisper has some ability to handle code-switching without specific fine-tuning due to its broad multilingual training, targeted fine-tuning on code-switched data generally improves performance.

Most relevant to this study, Xie et al. (2023) fine-tuned the Whisper-small model using their MCE dataset. Their resulting model, named Whisper-MCE, achieved a Mix Error Rate (MER) of 14.28% on their MCE test set. This was a 35.13% relative reduction in MER compared to the original Whisper-small model. Whisper-MCE also performed very well on the Common Voice Cantonese (zh-HK) dataset (Xie et al., 2023). This work clearly demonstrates that fine-tuning Whisper is effective for MCE ASR. Xie et al. (2023) also proposed a new evaluation metric called FAL (Fidelity, Accuracy, Latency) to provide a more comprehensive assessment of ASR systems in mixed-language scenarios.

The significant improvement Xie et al. (2023) found with Whisper-small raises an important question: Was this large relative improvement partly because Whisper-small started with a higher error rate on the complex MCE task compared to what a much larger model like whisper-large-v3 might achieve? Furthermore, fine-tuning on a relatively small dataset like the MCE dataset (34.8 hours) carries the risk of ”language forgetting,” where the model’s performance on its original training languages (like general English) might decline (Timmel et al., 2025). The choice of model size could influence this. Despite its greater capacity, Whisper-large-v3 may overfit the small dataset, even if it retains broader linguistic knowledge.

2.5 Bridging the Gap: Rationale for the Current Study

ASR has advanced significantly, leading to powerful models like OpenAI’s Whisper. Research shows these models can be adapted for difficult tasks, such as understanding speech in less common languages or when speakers mix languages (Xie et al., 2023). Specifically, Xie et al. (2023) demonstrated that fine-tuning the Whisper-small model with their MCE dataset greatly improved its performance for Mixed Cantonese-English (MCE) speech.

However, it is not yet fully understood how different sizes of Whisper models compare when fine-tuned for complex, mixed-language situations like MCE, especially with limited training data. This study compares whisper-small and whisper-large-v3 to explore how model size affects adaptation, especially under resource constraints. While Xie and colleagues provided valuable insights using Whisper-small, a direct comparison with a much larger and newer version like whisper-large-v3 on the same MCE task is still needed.

This study examines how whisper-small and whisper-large-v3 adapt to code-mixed Cantonese-English data. One hypothesis is that whisper-small, starting from a weaker baseline, may show a greater relative improvement after fine-tuning. In contrast, whisper-large-v3, with its higher capacity and built-in Cantonese token, is expected to achieve better absolute accuracy. However, its gains may be smaller in proportion, what called diminishing returns, especially when fine-tuning on a small dataset like MCE (34.8 hours). Large-v3 might reach its performance ceiling quickly, and the extra accuracy may not justify its higher computational cost. (Pratap et al., 2023; Radford et al., 2023; Xie et al., 2023).

This study offers practical, research-backed guidance for selecting the right model for this tricky language mixing situation, especially when working with limited resources. Aiming to help researchers and developers make smart choices that balance accuracy with the real-world demands of actually using these systems.

3 Methodology

This study aims to evaluate and optimize the performance of OpenAI's Whisper model for Cantonese-English code-switched speech recognition. The methodology involves using a specialized, open-source dataset, the Mixed Cantonese-English (MCE) Dataset created by Xie et al. (2023), for fine-tuning and evaluating the Whisper model. Performance will be quantified using standard metrics: Word Error Rate (WER) and Character Error Rate (CER). Ethical considerations regarding the use of this dataset and the model will also be addressed, drawing on established principles and relevant research.

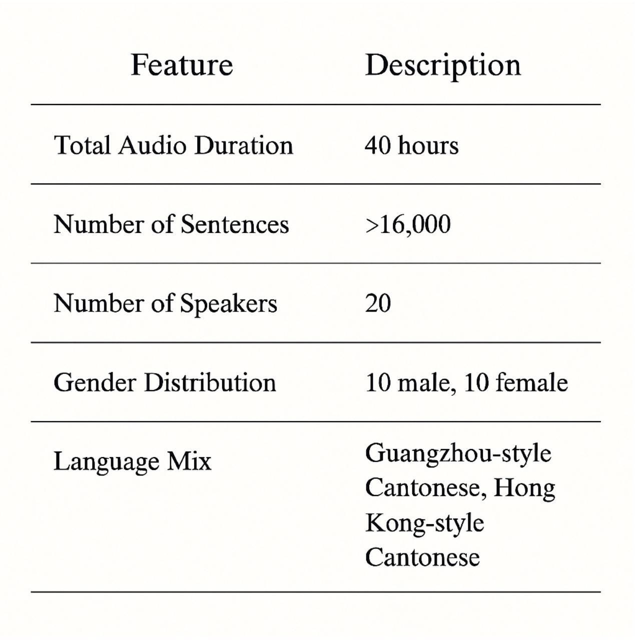
3.1 Dataset: Mixed Cantonese-English Dataset (MCE Dataset)

3.1.1 Rationale and Background for MCE Dataset Development

High-quality, domain-specific datasets are crucial for advancing speech recognition. However, resources for minor or mixed languages are often scarce and may be of poor quality. Cantonese, despite having over 85 million native speakers, is considered a low-resource language in Natural Language Processing due to factors such as the dominance of Mandarin and diversity in character encoding and input methods (Xie et al., 2023). Recent work has shown that fine-tuning large language models on task-specific data can significantly improve performance (Touvron et al., 2023, as cited in Xie et al., 2023), with data quality being more critical than quantity (Gunasekar et al., 2023, as cited in Xie et al., 2023). Thus, the MCE Dataset was developed by Xie et al. (2023) to address the lack of suitable, high-quality resources for Cantonese-English code-switching research, and it is available as an open-source resource.

3.1.2 MCE Dataset Corpus Characteristics

The MCE Dataset, as created by Xie et al. (2023), comprises over 16,000 sentences, with each sentence containing a mixture of Chinese (Cantonese) and English characters. The total duration of their audio dataset is 40 hours (Xie et al., 2023).



Feature	Description
Total Audio Duration	40 hours
Number of Sentences	>16,000
Number of Speakers	20
Gender Distribution	10 male, 10 female
Language Mix	Guangzhou-style Cantonese, Hong Kong-style Cantonese

Figure 4: [Characteristics of the dataset]

3.2 Speech Recognition Model: OpenAI Whisper

3.2.1 Architecture Overview

This study uses OpenAI's Whisper, an Automatic Speech Recognition (ASR) system built upon an encoder-decoder Transformer architecture (Radford et al., 2023). The input audio is resampled to 16,000 Hz and converted into an 80-channel log-magnitude Mel spectrogram, computed using 25 ms windows with a 10 ms stride. This spectrogram is then normalized (Radford et al., 2023).

The encoder processes the Mel spectrogram. It typically involves convolutional layers followed by Transformer encoder blocks with self-attention mechanisms, which capture complex patterns and long-range dependencies in audio sequences. Positional embeddings are added to incorporate sequence information (Radford et al., 2023).

The decoder is a standard Transformer decoder that generates the output text sequence. It uses learned positional embeddings and a Byte-Pair Encoding (BPE) tokenize. Whisper models employ special tokens to manage tasks such as language identification, transcription, and translation, and to indicate timestamps or the absence of speech. The model architecture supports various sizes, with differing numbers of parameters, layers, and attention heads, affecting performance and computational cost (Radford et al., 2023).

3.2.2 Training Paradigm and Capabilities

Whisper was trained on a massive dataset of 680,000 hours of audio, which was largely collected from the internet and is described as ”weakly supervised”. This extensive and varied training data is fundamental to Whisper's robustness to accents, background noise, and jargon, and its notable zero-shot capabilities across numerous languages and acoustic conditions. ”Zero-shot” refers to its ability to perform tasks on languages or conditions without specific fine-tuning for them (Radford et al., 2023).

Beyond transcription, Whisper is a multitask model capable of speech-to-text transcription in multiple languages, translation of several non-English languages into English, language identification, and voice activity detectio. A distinctive feature is its prompt functionality, allowing users to provide context (e.g., previous transcriptions or lists of rare words) to guide and improve transcription accuracy (Radford et al., 2023).

3.2.3 Applicability and Fine-tuning Method for This Study

While Whisper demonstrates strong zero-shot performance, fine-tuning on specific datasets like the MCE Dataset can enhance its performance for particular tasks, such as Cantonese-English code-switching (Xie et al., 2023). This study will fine-tune a pre-trained Whisper model using the MCE Dataset to adapt it to the nuances of this mixed-language speech. The ”weakly supervised” nature of Whisper’s original training data can be a source of issues like hallucinations (Koenecke et al., 2023), which fine-tuning on a cleaner, specific dataset like MCE Dataset might help mitigate, although inherent model characteristics may persist.

3.3 Speech Recognition Performance Evaluation Metrics

3.3.1 Word Error Rate (WER)

WER is a standard metric for ASR accuracy. It measures the differences between the model’s output transcript and a reference transcript at the word level. WER is calculated as the sum of substitutions (S), deletions (D), and insertions (I) needed to transform the hypothesis into the reference, divided by the total number of words in the reference text (N).

A lower WER indicates better performance.

3.3.2 Character Error Rate (CER)

CER is similar to WER but operates at the character level and is also a widely used metric. It is calculated as the sum of character-level substitutions (Schar), deletions (Dchar), and insertions (Ichar), divided by the total number of characters in the reference text (Nchar). CER is particularly useful for character-based languages and for analyzing fine-grained errors in mixed-script scenarios, such as Cantonese-English, where word boundaries can be ambiguous.

3.3.3 Application and Interpretation in This Study

Both WER and CER will be used to evaluate the Whisper model’s performance on the MCE Dataset. WER will provide a macro-level view of word recognition and sentence structure understanding. CER will offer finer-grained insights into character-level accuracy, crucial for mixed Cantonese-English text. Analyzing both can help identify specific error patterns. However, it is important to note that these metrics do not directly measure semantic accuracy or capture more complex errors like hallucinations (Koenecke et al., 2023).

3.4 Tools and setup

The model fine-tuning and evaluation were conducted using NVIDIA A100 GPU nodes on the University of Groningen's Hábrók HPC GPU cluster. The code can be found on GitHub at https://github.com/hoolim/ Miao-demonstrator.

3.5 Ethical Considerations in Data and Model Usage

3.5.1 Ethical Issues in MCE Dataset Data Sourcing

The MCE Dataset, an open-source resource created by Xie et al. (2023), involved their team recruiting and compensating 20 local Hong Kong volunteers for audio recording. While ethical data collection generally requires informed consent, ensuring participants understand the data's purpose, usage, and their rights (including voluntary participation and withdrawal), Xie et al. (2023) state that dialectal accents were preserved for diversity in the MCE Dataset. This is valuable for model robustness but requires careful handling of potentially identifiable voice characteristics. Their paper does not provide extensive details on the specific informed consent or anonymization procedures employed by their team beyond stating volunteers were ”hired.” General ethical principles in dataset creation also include data minimization (collecting only necessary data) and purpose limitation (using data only for explicit research purposes). This study utilizes the MCE Dataset as an existing, publicly available resource, relying on the ethical considerations undertaken by its original creators.

3.5.2 Ethical Implications of Using the Whisper Model

Large-scale models like Whisper, despite extensive training, may exhibit biases and underperform for certain demographic groups, accents, or speech patterns not well-represented in their original training data. And found that Whisper's hallucinations occurred more frequently for speech with longer non-vocal durations, disproportionately affecting speakers with conditions like aphasia. This raises concerns about fairness and potential discrimination(Radford et al., 2023; Kulkarni et al., 2025).

ASR models process audio that may contain sensitive information. If not handled securely, this information could be exposed. While Whisper can be run locally, improper data handling of its outputs (transcriptions) can pose privacy risks, especially if sensitive conversations are transcribed and stored without adequate safeguards. The generation of inaccurate associations or hallucinations by the model, as highlighted by Koenecke et al. (2023), can also lead to privacy violations if false information about individuals is recorded.

3.5.3 Measures to Address Ethical Issues in This Study

For MCE Dataset: This study will rely on the MCE Dataset as provided by Xie et al. (2023) as an open-source resource. Any use will adhere to the terms specified by the dataset creators, if any. The ethical considerations related to its creation, as can be inferred from their publication, are acknowledged.

For Whisper Model Usage: The potential for biases and hallucinations in the Whisper model will be acknowledged in the analysis of results. Where feasible, error analysis will consider if specific harmful outputs or biased performance patterns emerge on the MCE Dataset. Model limitations will be responsibly reported.

Overall Statement: This research is conducted for academic purposes. Any models fine-tuned or data generated will be handled with consideration for ethical implications. Deployment in sensitive real-world applications would require further rigorous ethical review and safeguards beyond the scope of this thesis.

3.6 Methodological Summary

This methodology outlines a systematic approach by utilizing the open-source MCE Dataset (Xie et al., 2023) to fine-tune and evaluate the OpenAI Whisper model (Radford et al., 2023). Performance will be assessed using WER and CER. Throughout the research process, ethical considerations related to both the dataset and the model will be addressed, aiming for a balanced approach between technological exploration and responsible research conduct.

4 Experimental Setup

This section details the comprehensive methodology used to fine-tune Whisper models on a Cantonese-English code-switching speech corpus.

4.1 Environment and Dependencies

All experiments were conducted on an NVIDIA A100 GPU node. Due to system-level constraints of the available Python 3.6 environment, a robust and isolated workspace was established using Miniconda. A dedicated Conda environment was created with Python 3.9, which served as the foundation for the project. Key dependencies, including specific versions of PyTorch (compiled for CUDA 12.1) and HuggingFace Transformers, were installed within this environment to ensure compatibility and access to the latest model features.

4.2 Data Preprocessing

The MCE dataset, featuring speech from 160 speakers, was used. The raw data, consisting of .wav files and corresponding .csv transcriptions, was first consolidated. A custom Python script was developed to parse the gb18030-encoded CSV files, skip headers, and create a unified metadata.csv file. This master file mapped the absolute path of each audio file to its corresponding transcription, which was generated by algorithmically matching text to audio based on its row number. The resulting dataset was loaded and split into 95% for training and 5% for testing. A standard Whisper preprocessing pipeline was applied: audio was resampled to 16kHz, and text was tokenized. To address the code-switching nature of the data, the Whisper tokenizer was configured in multilingual mode without a preset language, enabling it to process both Chinese and English text dynamically.

4.3 Model Fine-Tuning and Strategy

The fine-tuning process was iterative, involving two primary models and evolving strategies based on experimental outcomes.

4.3.1 whisper-small:

An initial fine-tuning pass was performed on the openai/whisper-small model to establish a performance baseline. During this phase, several technical challenges were addressed, including debugging evaluation metric computation (jiwer) to handle empty strings resulting from punctuation removal.

4.3.2 whisper-large-v3:

To pursue higher accuracy, the openai/whisper-large-v3 model was subsequently fine-tuned. Initial experiments with this larger model revealed severe overfitting, where the validation Word Error Rate (WER) increased as training progressed. To counter this, a series of advanced regularization techniques were systematically implemented: 1. The learning rate was significantly reduced to encourage more stable and generalized learning. 2. Weight decay was introduced to penalize large weight values and prevent model over-complexity. 3. A strategy of freezing the model's encoder was employed, focusing the training exclusively on the decoder to adapt it to the specific code-switching language style without altering the powerful, pre-trained acoustic representations.

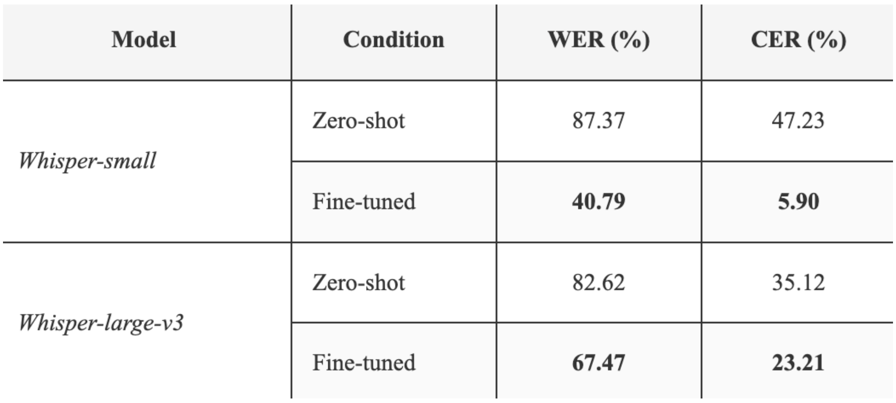
Throughout the training, system resource limitations such as disk space (OSError: No space left on device) and RAM were managed by adjusting checkpointing frequency (e.g., save-steps) and disabling data preprocessing multiprocessing (num-proc=1) to ensure run stability.

4.4 Evaluation

Model performance was evaluated based on Word Error Rate (WER) and Character Error Rate (CER) on the held-out test set. The final analysis will compare the performance of the fine-tuned whisper-small and whisper-large-v3 models against their original zero-shot counterparts to quantify the impact of fine-tuning on this specific code-switching task.

5 Results

The performance of the whisper-small and whisper-large-v3 models was evaluated on the held-out test set by comparing their zero-shot capabilities against their fine-tuned counterparts. The primary metrics were Word Error Rate (WER) and Character Error Rate (CER), where lower values indicate better performance. The results are summarized in Figure 5.



Model	Condition	WER (%)	CER (%)
Whisper-small	Zero-shot	87.37	47.23
	Fine-tuned	**40.79**	**5.90**
Whisper-large-v3	Zero-shot	82.62	35.12
	Fine-tuned	**67.47**	**23.21**

Figure 5: [Performance Comparison of Whisper Models on the Code-Switching Task]

5.1 Analysis of whisper-small

In its zero-shot state, the whisper-small model struggled with the code-switching task, resulting in a high WER of 87.37%. Fine-tuning yielded a substantial improvement, reducing the WER by more than half to 40.79%. The most significant gain was in CER, where the CER fell from 47.23% to just 5.90%. This indicates that the fine-tuned model became highly proficient at transcribing the correct sequence of Chinese characters and English letters, even if some word-level errors persisted. The model performed best at the 2000-step checkpoint, and then validation error rates started to level off.

5.2 Analysis of whisper-large-v3

The whisper-large-v3 model also performed poorly in its zero-shot configuration, with an initial WER of 82.62. Fine-tuning provided a clear benefit, lowering the WER to 67.47 and reducing the CER from 35.12% to 23.21%. However, while the model learned to recognize characters more accurately, its resulting word-level accuracy was still considerably worse than that of the fine-tuned small model. This suggests the larger model was more prone to overfitting on the training data, learning character patterns but failing to generalize its word-level and structural understanding to the unseen test set.

The results demonstrate that fine-tuning is a highly effective strategy for adapting Whisper models to Cantonese-English code-switching. The whisper-small model benefited well, achieving dramatic improvements in both word and character recognition.

The whisper-large-v3 experiment, however, highlights the challenges of tuning larger models on a dataset of this size. Although its performance improved from its zero-shot baseline, it failed to outperform the smaller, fine-tuned model. This suggests that without a larger dataset or more aggressive regularization, the large-v3 model's greater capacity leads to overfitting, making the smaller whisper-small model a more suitable and effective choice for this particular task.

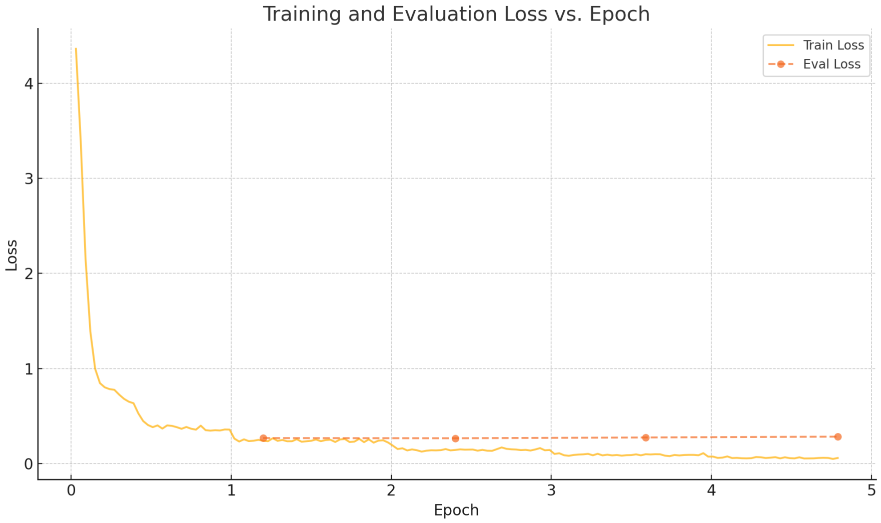


Figure 6: Whisper-small Training Loss Vs. Epoch

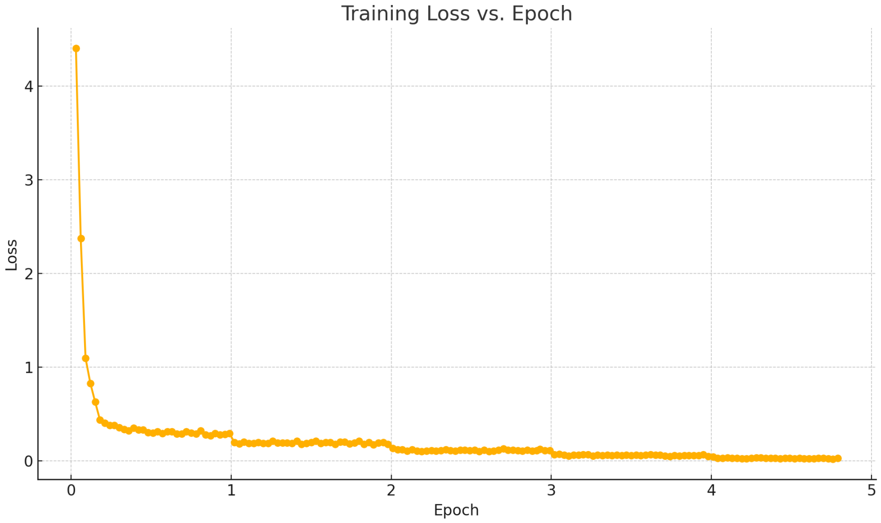


Figure 7: Whisper-Large Training Loss Vs. Epoch

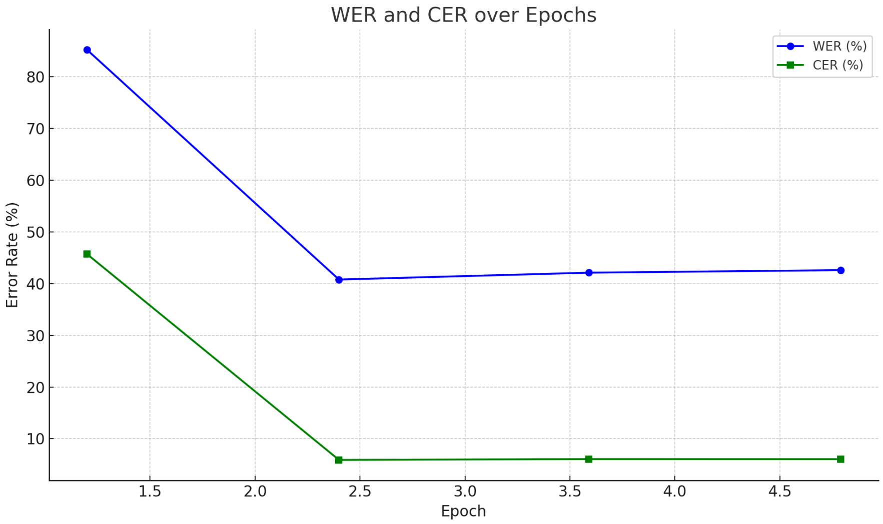


Figure 8: Whisper-small WER And CER Vs. Epoch

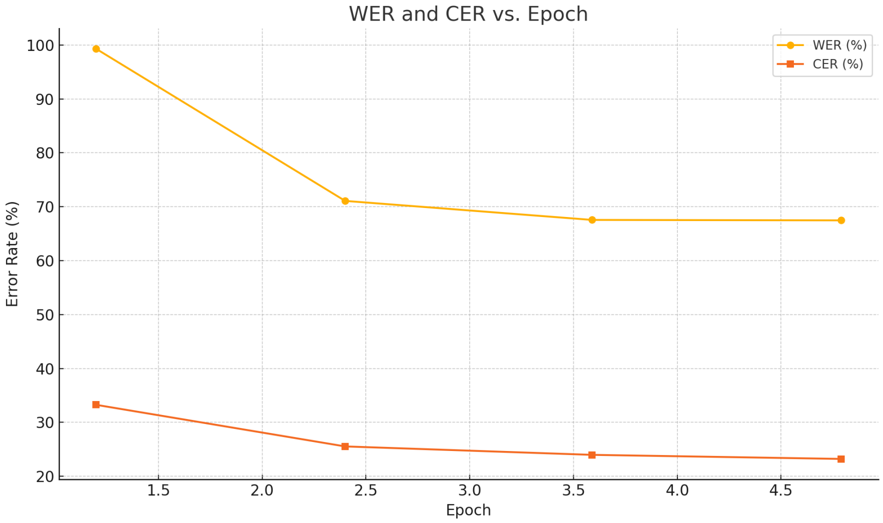


Figure 9: Whisper-Large WER And CER Vs. Epoch

6 Discussion

This study examined how much fine-tuning actually improves Whisper models when processing mixed Cantonese-English speech. The results provide a nuanced answer to the research question, confirming the primary hypothesis while challenging the secondary assumption about the scalability of performance gains.

6.1 Answering the Research Question: The Efficacy of Fine-Tuning

The central research question asked if fine-tuning Whisper-small and Whisper-large-v3 would reduce WER and CER compared to their zero-shot performance. The data unequivocally supports this. For the whisper-small model, fine-tuning was transformative, slashing the WER from 87.37% to 40.79% and, most impressively, the CER from 47.23% to a mere 5.90%. Similarly, the whisper-large-v3 model, despite its own challenges, saw its WER improve from 82.62% to 67.47% and its CER decrease from 35.12% to 23.21%.

This confirms the first part of the hypothesis: fine-tuning on a domain-specific, code-switched dataset leads to a statistically significant reduction in error rates. This finding aligns with the broader literature (e.g., works by Gandhi, 2023; Radford et al., 2023), which establishes that adapting large pre-trained models to specific domains or dialects is a highly effective strategy. The improvement in CER, in particular, suggests that the models became highly adept at the fundamental task of mapping the acoustic features of the dialect to the correct sequence of characters, even when word-level understanding was imperfect.

6.2 The Contradicted Hypothesis: Model Scale vs. Overfitting

The second part of the hypothesis that whisper-large-v3 would show bigger improvements turned out to be wrong. While the large model did get better compared to its starting point, its final error rate of 67.47% was much worse than the fine-tuned small model's 40.79

This outcome is a classic illustration of the tension between model capacity and dataset size, a well-documented challenge in the machine learning literature. The large-v3 model, with its vast number of parameters, appears to have overfitted to the training set. It became excellent at recognizing the character patterns within the training data (as shown by its improved CER) but failed to learn the more generalizable syntactic and semantic rules of the code-switched dialect. Instead of learning to ”understand,” it began to ”memorize.”

In contrast, the whisper-small model, with its more constrained capacity, was forced to learn more robust and generalizable representations. It could not simply memorize the training data, compelling it to find more efficient patterns that transferred successfully to the unseen test set. This finding suggests that for specialized tasks with limited data like this one, a smaller model may actually work better because it's naturally less prone to overfitting. The attempt to mitigate this with techniques like a reduced learning rate and weight decay was insufficient, indicating that a more aggressive regularization strategy or a significantly larger dataset would be required to unlock the full potential of the large-v3 model.

6.3 Limitations

It is important to acknowledge the limitations of this study, which provide avenues for future research.

Dataset Size and Scope: While the MCE dataset contains a valuable variety of speakers, it is still modest in size compared to the massive datasets used for pre-training. This limited data volume was likely the primary contributor to the overfitting observed in the large-v3 model. The dataset also may not cover all possible syntactic structures of Cantonese-English code-switching.

Limited Hyperparameter Search: The fine-tuning process involved a targeted set of hyperparameters. A more thorough search across different learning rates, weight decay values, and training steps could potentially produce an even better model, especially for large-v3.

Scope of Models: This study was limited to the small and large-v3 variants of Whisper. An investigation into the medium model could provide further insight into the relationship between model size and performance on this task, potentially identifying a ”sweet spot” of capacity and generalization.

Evaluation Metrics: While WER and CER are standard metrics, they do not fully capture the nuances of speech recognition quality, such as the naturalness of the output or the semantic correctness of transcribed sentences containing code-switched terms.

In conclusion, this research successfully demonstrates the significant value of fine-tuning for adapting Whisper to complex, code-switched speech, while also providing a case study on the critical challenge of model overfitting.

7 Conclusion

This study tested how well fine-tuning OpenAI's Whisper models works for the challenging task of recognizing mixed Cantonese-English speech. By comparing how whisper-small and whisper-large-v3 performed before and after fine-tuning, the research aimed to determine how much domain-specific adaptation could improve results and whether model size affected the outcome.

7.1 Summary of Findings

The findings confirm that fine-tuning substantially enhances Whisper's performance on this specialized task. The whisper-small model, in particular, demonstrated a remarkable transformation, achieving a reduction in both word and character-level errors. This indicates that even a smaller model can learn the intricate phonetic patterns of a code-switched dialect with high fidelity.

However, the hypothesis that the larger whisper-large-v3 model would yield greater improvements was contradicted. Despite improving upon its own zero-shot baseline, its final word-level accuracy was notably worse than that of the fine-tuned small model. This suggests its vast capacity led it to overfit on the training data, learning surface-level character patterns but failing to generalize its linguistic understanding. This study provides a compelling case where a smaller, more constrained model can prove more effective for a specialized, medium-resource task.

7.2 Future Work

This research opens up several paths for future investigation, influenced by the current study's limitations. A major constraint was the small training dataset, which likely led to the overfitting seen in the large-v3 model. Future research should focus on expanding the dataset to create a stronger training foundation that could better take advantage of what larger models have to offer.

Additionally, training cutting-edge models like whisper-large-v3 demands significant computing power. Limited GPU availability restricted the exploration of advanced hyperparameter tuning and regularization methods. A more comprehensive search could potentially improve performance and reduce overfitting.

Finally, while this study examined the small and large-v3 versions, how intermediate-sized models like whisper-medium perform remains unclear. These could offer a better balance between capability and generalization for this task. A more detailed qualitative error analysis that goes beyond overall metrics would also help understand the specific language challenges these models encounter. In conclusion, this research confirms that fine-tuning works well for adapting speech recognition models to complex mixed-language scenarios. It also provides a practical example of the important trade-offs between model size and dataset size, offering useful insights for practitioners tackling similar speech recognition challenges.

References

Chan, J. Y. C., Cao, H., Ching, P. C., & Lee, T. (2009). Automatic recognition of cantonese-english code-mixing speech. *International Journal of Computational Linguistics & Chinese Language Processing*, *14*(3), 281–304.

Chan, W., Jaitly, N., Le, Q., & Vinyals, O. (2016). Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. *Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4960–4964. https://doi.org/10.1109/ICASSP.2016.7472621

Graves, A., Fernández, S., Gomez, F., & Schmidhuber, J. (2006). Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. *Proceedings of the 23rd International Conference on Machine Learning (ICML)*, 369–376. https://doi.org/10.1145/1143844.1143891

Gu, Y., Shivakumar, P. G., Kolehmainen, J., Gandhe, A., Rastrow, A., & Bulyko, I. (2023). Scaling laws for discriminative speech recognition rescoring models. *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. https://doi.org/10.1109/ICASSP49357.2023.10094860

Gulati, A., Qin, J., Chiu, C.-C., Parmar, N., Zhang, Y., Yu, J., Han, W., Wang, S., Zhang, Z., & Wu, Y. (2020). Conformer: Convolution-augmented transformer for speech recognition. *Proceedings of Interspeech 2020*, 5036–5040.

Gumperz, J. J. (1982). *Discourse strategies*. Cambridge University Press.

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A.-r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., & Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, *29*(6), 82–97. https://doi.org/10.1109/MSP.2012.2205597

Juang, B. H., & Rabiner, L. R. (2005). Automatic speech recognition–a brief history of the technology development. In K. Brown (Ed.), *Encyclopedia of language and linguistics* (2nd ed., pp. 505–519, Vol. 1). Elsevier.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., & Amodei, D. (2020). *Scaling laws for neural language models* [arXiv preprint arXiv:2001.08361]. https://arxiv.org/abs/2001.08361

Koenecke, A., Choi, A. S. G., & Schellmann, H. (2023). Careless whisper: Speech-to-text hallucination harms. *Proceedings of the 2023 ACM Conference*. https://doi.org/10.1145/3630106.3658996

Mustafa, M. B., Yusoof, M. A., Khalaf, H. K., Abushariah, A. A. R. M., Kiah, M. L. M., Ting, H. N., & Muthaiyah, S. (2022). Code-switching in automatic speech recognition: The issues and future directions. *Applied Sciences*, *12*(19), 9541. https://doi.org/10.3390/app12199541

Prabhavalkar, R., Hori, T., Sainath, T. N., Schlüter, R., & Watanabe, S. (2023). *End-to-end speech recognition: A survey* [arXiv preprint arXiv:2303.03329]. https://arxiv.org/abs/2303.03329v1

Pratap, V., Tjandra, A., Paskov, I., Sriram, A., Baevski, A., & Likhomanenko, T. (2023). *Scaling speech technology to 1,000+ languages* [arXiv preprint arXiv:2305.13516]. https://arxiv.org/abs/2305.13516

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. *Proceedings of the 40th International Conference on Machine Learning (ICML)*, *202*, 28492–28518.

Timmel, V., Paonessa, C., Vogel, M., Perruchoud, D., & Kakooe, R. (2025). *Fine-tuning whisper on low-resource languages for real-world applications* [arXiv preprint arXiv:2412.15726]. https://arxiv.org/abs/2412.15726v3

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*.

Xie, P., Liu, X., Chan, T. W., Song, Y., Wang, Y., Chen, H., Bie, Y., & Chen, K. (2025). *Switchlingua: The first large-scale multilingual and multi-ethnic code-switching dataset* [arXiv preprint arXiv:2506.00087]. https://arxiv.org/abs/2506.00087v1

Xie, P., Liu, X., Chen, Z., Chen, K., & Wang, Y. (2023). *Whisper-mce: Whisper model finetuned for better performance with mixed languages* [arXiv preprint arXiv:2310.17953]. https://arxiv.org/abs/2310.17953

Zeyer, A., Merboldt, A., Michel, W., Schlüter, R., & Ney, H. (2021). Librispeech transducer model with internal language model prior correction. *Proceedings of Interspeech 2021*, 2052–2056. https://doi.org/10.21437/Interspeech.2021-1510

Appendices

A Declaration of AI use in a master thesis

I hereby affirm that this Master thesis was composed by myself, 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. During the preparation of this thesis, I used Claude 4.0 and Chatgpt 4o for the following purposes: restructuring sentences in the literature review to improve citation flow and readability; creating visual representations of quantitative results from experimental data; troubleshooting code problems during model runs; proofreading for language accuracy, grammatical correctness, and improving academic tone; quickly summarizing existing research to identify key insights; formatting references according to APA 7th edition style; revising section and subsection headings for better clarity and conciseness; and getting advice on task planning and time management throughout the writing process. All content was subsequently reviewed, verified, and substantially modified by me.

Haolin Miao 11.06.2025