university of groningen

campus fryslân

Streaming Speech Recognition for Smart Glasses: A Fine-tuning Approach Based on Pre-trained FastConformer

Ruoxin Kang

university of groningen / campus fryslân

University of Groningen - Campus Fryslân

Streaming Speech Recognition for Smart Glasses: A Fine-tuning Approach Based on Pre-trained FastConformer

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)

Ruoxin Kang (S5877172)

June 11, 2025

Acknowledgements

This thesis marks the end of my study in the Voice Technology program at Campus Fryslân, University of Groningen. First and foremost, I would like to thank Campus Fryslân, and in particular, the Voice Technology program. This program has provided me with a truly inspiring academic environment. It was here that I discovered my passion for the combination of linguistic foundations and technical skills. This unique combination offered by the program allowed me to develop a deeper understanding of how speech works. I am especially grateful to all the instructors and experts who contributed to the courses and lectures. The classes taught by Dr. Matt Coler, Dr. Joshua Schäuble, Dr. Vass Verkhodanova, Dr. Phat Do, and Dr. Shekhar Nayak were all vivid and meaningful. They not only covered theoretical knowledge but also helped me to think critically about real-world applications of voice technologies. The expert lectures brought by guest speakers also further expanded my perspectives and allowed me to explore cutting-edge developments in the field.

My sincere thanks go to my thesis supervisor, Shekhar. He is a very kind and professional mentor. Throughout the project, his guidance has been invaluable. His encouragement and positive feedback gave me confidence many times. His insightful advice helped me to find a clearer direction for my master thesis and even some parts of my future decision. I truly appreciate his patience, understanding, and consistent support during every stage of the project.

Lastly, I would like to express my most heartfelt gratitude to all the people who are important in my life. Also, thanks to me making the decision to come here. I am truly thankful.

Abstract

Real-time automatic speech recognition (ASR) on wearable devices such as smart glasses is a key technology for accessible communication, particularly for hearing-impaired users. However, current streaming ASR systems still face key challenges when operating under strict latency constraints, especially in noisy, dynamic environments where smart glasses are most useful for multiple sensors. This study investigates domain-adaptive fine-tuning of a pre-trained cache-aware Fast-Conformer model to improve the latency-accuracy trade-off and speaker attribution performance in multi-channel streaming ASR. The study builds on the CHiME-8 Task 3 first baseline system, retaining its architecture while applying advanced and widely-used fine-tuning strategies, Cosine Annealing learning rate scheduling and Layer-wise Learning Rate Decay (LLRD), to optimize fine-tuning on the in-domain MMCSG dataset. Conversational speech in wearable contexts presents distinct acoustic and streaming challenges due to the mobility of the device and the presence of multi-channel microphone arrays. Results demonstrate consistent improvements over the baseline across all evaluated latency thresholds, achieving approximately 10% relative word error rate (WER) reduction without increasing model complexity or violating streaming constraints.

These findings highlight the effectiveness of advanced fine-tuning strategies for adapting pre-trained ASR models to realistic wearable applications, paving the way for more accurate and responsive streaming ASR systems to support accessible, real-world communication.

Contents

1 Introduction

Real-time automatic speech recognition for smart glasses represents a critical frontier in accessibility technology. This research area combines the challenges of real-time processing with the constraints of wearable computing.

For the history of automatic speech recognition history, the evolution of ASR architectures has progressed from hidden Markov models to deep neural networks, with recent advances in Transformer and Conformer models dramatically reducing word error rates. Most recently, the FastConformer architecture (Noroozi et al., 2024) improved inference efficiency through linearly scalable attention mechanisms, making it particularly promising for streaming applications.

However, current systems still face a significant performance-latency trade-off. At low latency settings (below 150ms) necessary for natural conversation, error rates increase substantially. This challenge is particularly acute in everyday environments with background noise and multiple speakers, precisely the scenarios where smart glasses would be most beneficial for hearing-impaired users.

This research directly addresses critical accessibility needs for more than 430 million people worldwide with hearing impairments. Current captioning technology presents significant challenges that impact users' daily lives. Most notably, these systems disrupt the natural flow of conversation for their low latency or worse performances. This disruption reduces comprehension and creates communication barriers adding difficulty to practical applying. Additionally, these solutions often constrain users' movement and limit social integration by restricting them to specific devices or positions. This thesis proposed approach aims to enhance existing streaming ASR systems by improving streaming ASR models through fine-tuning, trying to get more robust performance in real-world wearable usage scenarios. The initial goal of this study is to explore audio-based improvements.

The latency-accuracy trade-off remains a fundamental challenge. Chen et al. (2021) demonstrated that traditional streaming approaches sacrifice up to 30% relative word error rate (WER) compared to offline counterparts when operating at sub-300ms latency. For ASR specifically, Zmolíkova et al. (2024) documented that current systems for smart glasses struggle withspeaker attribution (18% absolute error rate) and environmental noise (47% WER increase in dynamic settings). Their analysis revealed that head movements produced a 15% relative increase in WER compared to static positioning due to changing microphone array geometry.

1.1 Research Questions and Hypotheses

In light of the preceding discussion, this research addresses the following question:

> **How can domain-adaptive fine-tuning of a pre-trained cache-aware FastConformer improve the latency-accuracy trade-off and speaker attribution accuracy in streaming ASR for smart glasses?**

And my hypothesis is: Based on the cache-based inference framework implemented in the CHiME-8 baseline by Zmolíková et al. (2024) and fine-tuning strategies used on ASR pretrained conformer model by Cai and Li (2024), I hypothesize that domain-adaptive fine-tuning of a pretrained cache-aware FastConformer model, incorporating improved learning strategies such as Cosine Annealing learning rate scheduling and Layer-wise Learning Rate Decay (LLRD), will reduce word error rates (WER) by 10–15% compared to the original CHiME-8 task3 first baseline system across multiple latency thresholds (150ms, 350ms, 1000ms).

1.2 Thesis Overview

Now that the motivation, questions and hypotheses for this research has been presented, the structure of this thesis is as follows:

- Section 1.1 presents the research questions and hypotheses
- Section 2 reviews relevant literature and positions this work within current research
- Section 3 describes the methodological approach
- Section 4 details the experimental setup
- Section 5 presents and analyzes the results
- Section 6 discusses implications and insights
- Section 7 concludes with key findings and future directions

2 Literature Review

This systematic literature review was conducted to critically examine existing research on streaming ASR, particularly for wearable applications such as smart glasses. While many studies have advanced streaming architectures and fine-tuning techniques, their applicability to multi-speaker, spatially dynamic, and low-latency wearable scenarios remains underexplored. After summarizing prior work, this review aims to critically synthesize current evidence, highlight key trends, and identify gaps that motivate the present study.

2.1 Search Strategy and Selection Criteria

To ensure a comprehensive and relevant literature foundation, a systematic search was used across several major academic databases. These included IEEE Xplore, ACL Anthology, arXiv (specifically within the Computer Science and Audio/Speech Processing categories), and Google Scholar. The search aimed to capture recent developments in streaming automatic speech recognition (ASR). Papers especially paid attention are related to architectures designed for real-time processing on wearable devices, specifically smart glasses, and to methods based on fine-tuning pre-trained models for audio input.

The search query was constructed using a combination of key terms and logical operators. It was designed to target studies involving both streaming and low-latency ASR systems. Specifically, the search string used the following structure: (“streaming ASR” OR “real-time speech recognition”) AND (“conformer” OR “fastconformer” OR “transformer”) AND (“cache-based” OR “low latency”) AND (“smart glasses” OR “wearable devices”).

Clear inclusion criteria were applied to select studies relevant to the research focus. Only papers published between 2020 and 2024 were considered. These had to be either peer-reviewed publications or high-quality preprints. Studies needed to focus on streaming ASR systems and use models based on the Transformer or Conformer architecture. The research also needed to target audio-only processing, with application to smart glasses scenarios, such as Project Aria MMCSG. Papers were also required to report quantitative results based on empirical evaluation.

At the same time, specific exclusion criteria were used to filter out unrelated or low-quality studies. Work focusing only on offline ASR was excluded. Studies without empirical experiments were also removed. In addition, papers centered purely on text-based language models were not considered. Furthermore, studies involving multimodal ASR and architectural modifications beyond standard pre-trained models and fine-tuning were also excluded. Finally, any study lacking a clear and reproducible description of its methodology was excluded. These criteria helped maintain the quality and focus of the selected literature.

2.2 Fine-tuning of ASR model

In recent years, with the increasing development of deep learning models, automatic speech recognition (ASR) systems have shown remarkable performance across diverse application scenarios. A key trend in modern ASR model development is the fine-tuning of large-scale pre-trained models to adapt them to specific domains, languages, or practical environments. In this process, the choice of optimization strategies plays a crucial role in ensuring training stability and achieving good generalization.

Among the most widely adopted optimization strategies, Cosine Annealing learning rate scheduling and Layer-wise Learning Rate Decay (LLRD) have demonstrated consistent effectiveness across various ASR tasks. The former provides a smooth learning rate decay that promotes stable convergence, while the latter enables selective adaptation of higher layers without disrupting well-learned lower-layer representations.

QuartzNet represents one of the early and influential works that systematically explored the use of Cosine Annealing learning rate scheduling in ASR model optimization (Kriman et al. (2020)). The QuartzNet architecture employs 1D time-channel separable convolutions to build lightweight and efficient ASR models. In its training process, the authors adopted a Cosine Annealing learning rate policy with warmup to stabilize learning and improve convergence speed. Their results demonstrated that this learning rate strategy was effective in preventing overfitting and enabled the model to achieve state-of-the-art performance on benchmark datasets such as WSJ and LibriSpeech. The success of its learning rate scheduling laid the groundwork for subsequent studies that combined this approach with layer-wise optimization.

The development of self-supervised learning has further influenced ASR model fine-tuning strategies. Data2vec, proposed by Baevski et al. (2022), is a general framework for self-supervised learning across speech, vision, and language. In the speech domain, Data2vec pre-trains models to capture contextual representations from raw audio. When fine-tuning these pre-trained models for ASR tasks, the authors employed a learning rate schedule that included an initial warmup phase followed by Cosine Annealing decay. Additionally, they used layer-wise adjustments to ensure that lower layers retained general acoustic knowledge while higher layers adapted to downstream ASR objectives. The combination of these strategies allowed Data2vec models to achieve competitive results with strong robustness across various speech datasets, highlighting the importance of carefully designed learning rate schedules and layer-wise control during fine-tuning.

Another example is the work with ASR pretrained Conformer models for speaker verification through transfer learning and knowledge distillation by Cai and Li (2024). In this study, the authors investigated the use of Conformer-based ASR models as feature extractors for speaker verification tasks. The fine-tuning process involved applying a Cosine Annealing learning rate schedule to promote stable adaptation of the model to the new task. Moreover, the study empirically validated the effectiveness of layer-wise learning rate decay. By fine-tuning the outputs of different layers with progressively smaller learning rates towards the lower layers, the researchers preserved the pre-trained acoustic representations while allowing higher layers to adapt to speaker-specific characteristics. The experimental results confirmed that this hierarchical learning rate adjustment led to superior performance in speaker verification, further supporting the value of combining Cosine Annealing and LLRD in ASR model fine-tuning.

In summary, recent literature demonstrates a positive trend towards the integration of Cosine Annealing learning rate scheduling and Layer-wise Learning Rate Decay in ASR fine-tuning pipelines. These strategies complement each other by promoting stable optimization dynamics and preserving valuable pre-trained features. The reviewed works, including QuartzNet, Data2vec, and Conformer-based transfer learning, provide strong empirical support for the effectiveness of these techniques. As ASR systems continue to evolve and adapt to new application domains, the careful design of fine-tuning strategies will remain a critical factor in achieving robust and high-performing models.

However, there are still several limitations remain regarding the direct applicability of these strategies to streaming ASR for wearable devices. Much of the prior paper has focused on offline ASR (e.g., Data2vec, QuartzNet), where full sequence context is available. In contrast, streaming

ASR operates under strict latency constraints, which may alter the dynamics of fine-tuning. In addition, existing studies largely target single-channel or controlled audio scenarios. The multi-channel, spatially dynamic nature of smart glasses input poses unique challenges for both model optimization and speaker attribution. Also, while Conformer-based transfer learning demonstrates promising results for speaker verification, its application to multi-speaker, real-time transcription tasks particularly under streaming constraints remains underexplored.

Therefore, while Cosine Annealing and LLRD represent promising components of a fine-tuning strategy, their combined effectiveness in the context of cache-aware, beamformed streaming ASR for wearable applications has not been systematically validated. This study addresses this gap by applying and evaluating these strategies within the CHiME-8 Task 3 baseline architecture for smart glasses.

2.3 Streaming ASR Architectures for Smart Glasses Applications

With the growing integration of smart glasses into daily life, there is an increasing demand for robust, low-latency automatic speech recognition (ASR) systems that can operate in real time. Streaming ASR architectures play a key role in enabling smart glasses to provide live transcription and captioning functionalities. Compared to conventional batch ASR systems, streaming ASR models process audio inputs sequentially, emitting transcriptions incrementally. This is essential for user-facing applications, where low latency and continuous interaction are required. Moreover, smart glasses introduce additional technical challenges, such as the need to handle multi-speaker conversations, perform speaker attribution, and suppress background speech captured by embedded microphone arrays. Recent research has explored various architectural designs to meet these demands.

Noroozi, Majumdar, Kumar, Balam, and Ginsburg (2024) proposed a stateful Conformer with cache-based inference designed for efficient streaming ASR. Their system stores intermediate activations to reduce redundant computations across streaming windows. On LibriSpeech, they demonstrated a significant reduction in word error rate (WER) under latency constraints. However, the system was evaluated on read speech and single-channel audio, limiting its applicability to smart glasses scenarios, which often involve dynamic, multi-speaker environments with spatially distributed sound sources.

Kim, Wu, Sridhar, Han, and Watanabe (2021) presented a multi-mode Transformer transducer that employs stochastic future context with reinforcement learning to optimize the lookahead window. Their system achieved consistent WER improvements over fixed-lookahead baselines. However, it introduced additional training and inference complexity, and latency variations were less predictable—a potential limitation for wearable devices where consistent latency is desired.

In addition, the CHiME-8 Task 3 baseline system offers a streaming ASR architecture explicitly designed for smart glasses applications. The system extends a pre-trained FastConformer Hybrid Transducer-CTC model for multi-channel, streaming speech recognition. A fixed superdirective beamformer is used to generate directional audio inputs from the glasses' microphone array, covering 12 directions around the wearer plus a dedicated mouth beam. The ASR model is adapted to process these multi-channel inputs, enabling it to disambiguate between the wearer (SELF) and the conversation partner (OTHER), while suppressing unrelated bystander speech. Overall, the reviewed literature highlights significant progress in the design of streaming ASR architectures, particularly in addressing the latency-accuracy trade-off. Techniques such as cache-based inference, chunk-aware processing, and dynamic lookahead optimization have laid a strong foundation.

However, there are still key limitations persist. Most systems remain evaluated on single-speaker or single-channel data, limiting generalizability to wearable, real-world use cases. Speaker attribution accuracy under low-latency streaming remains a critical challenge, particularly in dynamic multi-speaker environments.

Addressing these gaps is essential to advancing practical, robust ASR solutions for smart glasses. In this context, the CHiME-8 Task 3 baseline system provides a more comprehensive and practical architecture. It effectively integrates several advances from recent research, multi-channel input processing, directional beamforming, and pre-trained model fine-tuning, into a cohesive framework tailored for smart glasses applications. Moreover, the use of Serialized Output Training enables accurate speaker attribution, which is critical for user-facing functionalities such as live captioning. The system's design reflects a mature synthesis of current ASR developments, offering a solid starting point for further improvement.

The thesis aims to systematically investigate domain-adaptive fine-tuning of a pre-trained Fast-Conformer model within the CHiME-8 Task 3 baseline architecture. The goal is to evaluate whether such fine-tuning can meaningfully improve the latency-accuracy trade-off and speaker attribution accuracy under realistic smart glasses conditions—advancing the practical deployment of streaming ASR systems for accessible, wearable communication.

3 Methodology

This section outlines the methodology employed to design and implement a real-time streaming multi-talker automatic speech recognition (ASR) system for smart glasses applications. The primary objective of this study is to explore how fine-tuned ASR system can improve the performance of the challenges of multi-channel audio, multi-talker speech, and low-latency streaming processing, while maintaining high transcription accuracy and speaker attribution. The ultimate goal is to support live captioning applications, especially for hearing-impaired users. This study adopts an architecture inspired by Lin et al. (2023) . However, the system presented here integrates and adapts these components based on the specific goals and experimental needs of this research. Several configuration adjustments were explored to align the system with the target application scenario.

This methodology section first provides a brief description of the dataset used (Section 3.1), followed by a detailed explanation of the core methods and models (Section 3.2). Section 3.3 describes the technical framework, highlighting key implementation details and adjustments made. Section 3.4 introduces the evaluation methodology, including the metrics and evaluation pipeline employed. Finally, Section 3.5 discusses ethical considerations and research integrity principles that guided the study.

3.1 Dataset Description

The primary dataset used in this study is the Multimodal Multichannel Conversational Speech with Glasses (MMCSG) dataset, released as part of the CHiME-8 Task 3 challenge. This dataset is specifically designed to support research on ASR systems for smart glasses scenarios, making it highly suitable for the goals of this study.

The MMCSG dataset includes audio recordings captured using Project Aria smart glasses, which are equipped with a microphone array. The dataset contains paired audio and transcription data for two conversational speakers: the wearer of the glasses (SELF) and the conversation partner (OTHER). Each word in the transcription is annotated with both a speaker label and a timestamp, supporting detailed speaker-attributed ASR evaluation.

A key feature of the MMCSG dataset is its focus on realistic conversational scenarios. The recordings include background noise, reverberation, and occasional speech from bystanders, providing a challenging testbed for ASR models. Furthermore, the dataset is specifically designed to evaluate streaming ASR performance, with strict requirements on latency and speaker attribution.

Given these characteristics, the MMCSG dataset was chosen as the primary training and evaluation resource. The training subset was used for model fine-tuning, while the development subset served for validation and evaluation purposes. The detailed data processing steps will be described in the Data Processing section of this study.

3.2 Core Methods and Models

The system architecture developed in this study draws inspiration from the Directional ASR model proposed by Lin et al. (2023) and the corresponding baseline implementation provided by the CHiME-8 organizers. The goal is to build an ASR system that can effectively handle multi-channel inputs, perform speaker disambiguation, and operate in a streaming fashion suitable for real-time applications.

3.2.1 Overall System Architecture

The design focuses on effectively using multi-channel audio input, performing speaker disambiguation, and maintaining real-time streaming capabilities.

To achieve these goals, the system comprises three main components:

1. A Fixed NLCMV Beamformer, which transforms multi-channel raw audio into a set of directionally focused beams.

2. A Feature Extraction Pipeline, which computes log-mel features from each beam to provide informative spectral representations.

3. A Streaming ASR Model, based on a modified FastConformer Hybrid Transducer-CTC architecture, capable of producing serialized, speaker-attributed transcripts in a streaming manner.

The first stage of the system employs a Fixed Near-Linear-Constrained Minimum Variance (NLCMV) Beamformer. This component takes raw audio signals captured by the microphone array of the smart glasses and projects them into 13 beamformed channels. These include 12 beams evenly distributed around the horizontal plane and one beam directed towards the wearer's mouth. The beamformer utilizes pre-computed acoustic transfer functions (ATFs) measured in controlled environments to achieve consistent spatial filtering. This approach provides the model with rich spatial cues that are crucial for both enhancing target speech and suppressing irrelevant sounds.

Following beamforming, each of the 13 channels undergoes log-mel filterbank feature extraction. This process transforms the raw audio waveforms into a compact and informative spectral representation. Log-mel features are widely used in ASR systems due to their ability to capture perceptually relevant information while reducing data dimensionality. In this system, features from all 13 beams are computed independently and then concatenated, resulting in a comprehensive multi-channel input that preserves directional information.

The concatenated features are then passed to a Streaming ASR Model built on the FastConformer Hybrid Transducer-CTC architecture. The original pre-trained model, designed for single-channel input, is carefully modified to handle multi-channel input and perform Serialized Output Training (SOT). The model's input layer is extended to accept the 13-beam feature vector, and the tokenizer is augmented with special speaker tokens to distinguish between SELF (the wearer) and OTHER (the conversation partner). During fine-tuning on the MMCSG dataset, the model learns not only to transcribe speech accurately but also to correctly attribute each word to the corresponding speaker.

A key feature of the system is its strict adherence to streaming processing requirements. During inference, audio is processed in fixed-size chunks, and the model emits words with timestamps that reflect the extent of the input signal processed at the time of emission. This ensures that the system operates causally, without access to future context, and meets the low-latency demands of real-time captioning applications. The combination of beamforming, feature extraction, and streaming ASR provides a powerful foundation for robust multi-talker recognition in smart glasses scenarios.

3.2.2 Key Advantages of the Approach

This architecture offers several important advantages for the target application. By using multiple beamformed inputs, the model can distinguish speakers based on direction and suppress bystander

speech. Besides, the chunk-based streaming inference ensures low latency and supports real-time use cases, which makes it straming capable. For integrated speaker attribution, the Serialized Output Training (SOT) used as an integrated training objective enables the model to produce joint transcripts with accurate speaker labels, avoiding the need for separate diarization or speaker recognition modules.In addition, the architecture can be adapted to different microphone configurations and streaming constraints.

Overall, the system represents a strong foundation for building practical ASR solutions for smart glasses applications. The following section provides additional details on the technical framework and specific adaptations made in this study.

3.3 Technical Framework

The technical framework of this study builds upon a combination of state-of-the-art tools and carefully configured components to support the goals of real-time multi-talker ASR on wearable smart glasses. While the system architecture is inspired by the official CHiME-8 baseline and the approach described in Lin et al. (2023), this study involved active experimentation and configuration adjustments to better align the system with the practical requirements of live captioning.

At the core of the system is the FastConformer Hybrid Transducer-CTC model, a cache-aware architecture optimized for low-latency streaming recognitionREADME of the baseline . The FastConformer combines the representational power of convolutional and attention-based layers, while its cache-aware mechanisms ensure efficient processing in a chunk-based streaming setup. In this implementation, the encoder is extended to process concatenated log-mel features derived from 13 beamformed audio channels. These features provide rich spatial information, enabling the model to distinguish between different speakers and suppress irrelevant speech sources.

One of the key design elements of the FastConformer is its cache-aware streaming mechanism. In the convolutional layers, the model caches the most recent activations across input chunks, allowing it to maintain temporal continuity without redundant computation. Similarly, the self-attention layers employ a dynamic context cache that evolves with the incoming audio stream, enabling the model to incorporate sufficient temporal context while respecting streaming constraints. These mechanisms are essential for supporting real-time ASR on resource-constrained wearable devices, such as smart glasses.

The system adopts a chunk-based streaming inference strategy rather than continuous sample-level streaming. Input audio is divided into fixed-size chunks, and the model processes each chunk independently, emitting word hypotheses as it progresses. This design choice strikes a practical balance between computational efficiency and recognition accuracy. The decoder architecture employs a Hybrid Transducer-CTC approach, combining the strengths of both Recurrent Neural Network Transducer (RNN-T) and Connectionist Temporal Classification (CTC) objectives. The RNN-T component provides flexible alignment capabilities, essential for handling conversational speech with varying tempo and speaker overlaps. The CTC objective, in turn, stabilizes the training process and supports the generation of precise per-word timestamps—a critical requirement for multi-talker streaming ASR. The hybrid design ensures that the model can produce accurate, timestamped transcripts with reliable speaker attribution.

Speaker attribution is further enhanced through the use of Serialized Output Training (SOT). During fine-tuning, the model is trained to emit a special speaker token whenever the active speaker changes between the wearer (SELF) and the conversational partner (OTHER). The tokenizer and

model layers are modified to support this behavior. This integrated approach to speaker-aware transcription eliminates the need for separate diarization or speaker identification modules, simplifying the overall system pipeline.

In terms of input processing, the system using audio data captured by the seven-channel microphone array embedded in the Project Aria smart glasses. A Fixed NLCMV Beamformer is applied as a pre-processing step to produce 13 directionally focused audio beams. These include twelve beams evenly distributed around the horizontal plane and one beam directed at the wearer's mouth. The beamformer is implemented via efficient 1D convolutions, with coefficients derived from measured acoustic transfer functions (ATFs). The resulting beamformed audio provides rich spatial information that supports speaker disambiguation and cross-talk suppression.

It is important to note that while the MMCSG dataset includes additional visual and IMU modalities, this study focuses exclusively on optimizing the audio-based ASR pipeline. Future work may explore the integration of these complementary data streams. For the present research, concentrating on the audio domain allowed for a more controlled investigation of multi-channel beamforming and streaming ASR performance.

3.4 Evaluation Methodology

The evaluation of the system follows the official methodology used in the CHiME-8 Task 3 challenge. The goal is to measure not only how accurately the system transcribes the speech, but also how well it assigns each word to the correct speaker under streaming conditions. In this study, both word recognition accuracy and latency are treated as equally important, since the target application requires real-time captioning with low delay and correct speaker attribution.

The main metric used is the multi-talker word error rate (WER). This metric compares the system's output transcript, which includes both the recognized words and their assigned speaker labels, against the reference transcript. In the alignment process, several types of errors are considered. If a word is recognized correctly and attributed to the correct speaker, it is counted as a correct match. If the system outputs an extra word that does not appear in the reference, this is counted as an insertion. If a word present in the reference is missing from the hypothesis, this is counted as a deletion. If the system recognizes the wrong word for the correct speaker, this is counted as a substitution. Finally, if the word itself is correct but the system attributes it to the wrong speaker, this is counted as a speaker attribution error.

In addition to WER, latency performance is evaluated. For each correctly recognized and attributed word, latency is calculated as the difference between the timestamp produced by the system and the corresponding timestamp in the reference. This timestamp indicates how much of the audio input the system had seen when emitting the word. The evaluation reports the mean, median, and standard deviation of these latency values. These latency statistics help assess whether the system is suitable for real-time applications, where low and consistent latency is needed for an acceptable user experience.

The evaluation is carried out based on the multitalker_wer tool provided in the CHiME-8 first baseline system. In this study, the evaluation procedure is adjusted to perform speaker-specific alignment independently for each speaker stream, rather than applying global optimal alignment across both speakers as in the official scoring. This adjustment leads to differences in the reported WER values, which tend to be higher compared to the official baseline. The purpose of this modification is to maintain consistency in how speaker attribution is handled during the alignment process. Before

alignment, both the system's output and the reference transcripts are passed through text normalization. This includes removing punctuation, converting all text to lowercase, and applying a predefined list of permitted word substitutions. After normalization, the alignment between the hypothesis and the two reference speaker streams (SELF and OTHER) is performed using a dynamic programming approach. The tool finds the alignment that minimizes the total number of errors across both speakers.

By combining multi-talker WER and latency analysis, this evaluation process provides a clear and reliable way to assess whether the system meets the requirements for accurate and real-time multi-talker ASR on wearable devices.

3.5 Ethics and Research Integrity

This research was conducted in alignment with the ethical standards set by the institution and relevant faculty guidelines. The following aspects were central to ensuring the integrity and social responsibility of this work.

A key motivation behind this study is to contribute to the development of technologies that can improve accessibility for hearing-impaired users. The system is designed to provide real-time captioning in wearable devices, which could enhance communication opportunities for users in everyday settings. To support this goal, the methodology emphasized practical factors such as latency, speaker attribution accuracy, and robustness in noisy environments, which are critical for ensuring that such systems are usable in real-world scenarios. In addition, input was sought from accessibility research literature to better understand the needs of target users, and evaluation metrics were selected with these needs in mind. By aligning technical design with accessibility goals, the study aims to produce outcomes that are not only technically sound but also socially beneficial.

Research integrity was embedded in the entire study workflow. All experimental code is clearly documented, and the system pipeline is organized into modular scripts with version control. The study also follows the principle of honest reporting: both positive findings and limitations are explicitly stated. In particular, the known limitations of streaming ASR systems are acknowledged, and future directions for improvement are suggested.

The study complies fully with the ethical guidelines and data protection policies required for research of this type. The MMCSG dataset used in this study is a publicly released dataset under an appropriate license, with no personally identifiable information or sensitive data involved. No additional data collection was performed, and no human subjects were recruited. All data handling followed institutional privacy standards, ensuring that no individual privacy was compromised in the course of this research. Furthermore, care was taken during system development to avoid any practices that could inadvertently violate privacy expectations, such as improper retention of audio or metadata.

Adherence to the FAIR principles was also considered carefully. The system is built upon openly available models and tools, and the study outputs are structured to facilitate sharing with the wider research community. Code repositories are maintained with clear documentation and version control, supporting both reproducibility and reuse. Standard data formats and widely adopted tools (such as multitalker_wer and NeMo toolkit) are used to ensure interoperability. By following FAIR principles, the study helps promote open science and collaborative advancement in the field of speech recognition.

Overall, this research reflects a strong commitment to ethical conduct, social responsibility, and scientific transparency. Through careful attention to accessibility considerations, research integrity, data privacy, and open science practices, the study aims to contribute meaningfully to the development of responsible technologies.

4 Experimental Setup

When working with complex data and advanced deep learning models, the full experimental pipeline often involves many subtle steps, each of which can impact the final outcomes, to minimize the risk of the experiment. Therefore, the thesis using a thorough description of all relevant aspects of the experimental setup used in this study, including the datasets, data processing pipeline, model training procedure, and fine-tuning approach.

The experiments conducted in this study build upon the first baseline system released for the CHiME-8 Task 3 MMCSG challenge. This baseline is publicly available on GitHub MMCSG. It offers a robust and well-tested foundation for developing streaming ASR systems capable of handling conversational speech captured by smart glasses. The baseline system is designed to be fully reproducible, with clear documentation and open access to all relevant code and configuration files. Throughout this section, the study will carefully describe how to used and extended this baseline system to get a better performance.

This section is organized as follows. First, the thesis introduces the dataset used in the experiments, explaining its structure and characteristics. Then, the data processing and preparation steps for model training are described in detail. Third, the training procedure for the baseline system is presented, including the specific modifications made to the pre-trained model and the configuration of the training process. Finally, the fine-tuning approach applied to adapt the model to the target domain is discussed.

4.1 Dataset

The dataset used in this study is the MMCSG dataset provided as part of the CHiME-8 Task 3 challenge. This dataset is specifically designed to support research on multi-modal streaming conversational speech recognition using smart glasses. It offers a rich and realistic collection of multi-channel audio recordings paired with transcriptions that include explicit speaker labels.

The MMCSG dataset is divided into three subsets: the training subset (train), the development subset (dev), and the evaluation subset (eval). The experiments used the train subset for model fine-tuning and system development, while the dev subset was used for validation and evaluation throughout the development phase. The eval subset is reserved for final system testing and challenge submissions and was not used in any form during model training or tuning.

The audio recordings are organized under:
audio/train
audio/dev
audio/eval
while the corresponding transcriptions are provided in:
transcriptions/train
transcriptions/dev

Each recording in the dataset consists of synchronized multi-channel audio captured with Project Aria smart glasses. These prototype glasses are equipped with a seven-microphone array specifically designed to capture conversational speech in realistic environments. The audio files are stored in WAV format, sampled at 16 kHz with 16-bit linear PCM encoding. The multi-channel nature of the recordings enables spatial processing techniques, such as beamforming, to be applied effectively.

The transcriptions provided with the dataset follow the serialized output training (SOT) format, which explicitly encodes speaker changes and speaker identity within the transcription stream. This format uses special tokens (0 for SELF, 1 for OTHER) to indicate when the speaker changes, allowing models to learn not only to transcribe speech but also to perform accurate speaker attribution. Given the importance of speaker attribution in this challenge, this feature of the dataset is particularly valuable.

In addition to the core MMCSG dataset, the CHiME-8 challenge rules allow participants to use certain external datasets for system development. However, the study focused exclusively on the MMCSG dataset for fine-tuning the baseline system. External datasets such as Librispeech and TEDLIUM were only used in the separate baseline system trained from scratch, which is outside the scope of this work. By limiting to the provided in-domain data, ensuring that the results are directly comparable to those of other participants adhering to the same constraints.

Overall, the MMCSG dataset provides an excellent testbed for studying the challenges of streaming ASR in conversational settings with smart glasses. Its combination of realistic multi-channel recordings, explicit speaker labels, and carefully controlled data splits makes it ideally suited to the research objectives.

4.2 Data Processing

Data preparation and processing are critical components of the experimental pipeline. The overall goal of this stage is to transform the raw multi-channel audio and transcription data from the MMCSG dataset into a structured and standardized format that is suitable for model training. This includes preparing segmented audio chunks, aligned serialized output training (SOT) transcriptions, and multi-beam feature representations, as well as generating manifest files that describe the processed data.

4.2.1 Data Preparation Steps

The experiment used the script scripts/prepare_data.py provided in the official baseline system repository to process the MMCSG dataset. The process began by organizing the raw data into a consistent directory structure, with separate folders for audio recordings and transcriptions for each subset (train and dev). The next key step involved segmenting the audio recordings into smaller chunks of approximately 20 seconds in duration. This chunking strategy serves several important purposes. Firstly, it facilitates efficient model training by ensuring that input sequences are of manageable length, which helps optimize GPU memory usage and training speed. Secondly, it supports streaming inference by aligning the training data with the temporal constraints that streaming models must handle in practice. Finally, chunking helps ensure consistent behavior across recordings of varying lengths, promoting uniformity in model input processing. For each audio chunk, a corresponding serialized output training (SOT) transcription is prepared. The SOT format is specifically designed to encode both the content of speech and the identity of the speaker within a single transcription stream. Special tokens (0 for SELF and 1 for OTHER) are inserted into the transcription to indicate speaker changes, enabling the model to learn both speech recognition and speaker attribution tasks jointly. This step required carefully aligning the transcriptions with the segmented audio chunks to ensure that each token and timestamp accurately reflected the underlying audio. In addition to preparing the audio and transcription data, the system generated structured manifest files in JSON

format. These manifest files describe the training and validation data in detail, including the file paths to the segmented audio chunks, the duration of each chunk, and the aligned SOT transcriptions. These manifests provide a standardized interface for the training scripts to efficiently load and process the data during model training. The processed data was stored in the following directory structure:

data/train_chunks/
data/train_chunks.json
data/valid_chunks/
data/valid_chunks.json

By following this structured data preparation process, the input data used for model training was ensured to be both consistent and fully aligned with the requirements of the baseline system and the future development.

4.2.2 Beamforming

An important component of the baseline system is the use of a fixed beamformer. The goal of beamforming is to use the spatial information captured by the multi-microphone array of the Project Aria smart glasses to enhance the signal quality for each direction of interest. In particular, beamforming helps improve the signal-to-noise ratio (SNR) of target speech while suppressing interference and background noise, which is critical for achieving high ASR accuracy in realistic conversational environments.

The baseline system employs a fixed NLCMV (null-constrained linearly constrained minimum variance) beamformer that generates 13 distinct beams. Twelve of these beams are steered uniformly in 12 directions spaced evenly around the wearer of the glasses, covering the full horizontal plane. The remaining beam is directed specifically towards the mouth of the wearer, capturing speech from the wearer's own voice with high fidelity.

The beamformer coefficients were precomputed based on acoustic transfer functions (ATFs) recorded in anechoic environments using the Aria glasses. These coefficients are provided as part of the baseline system release and are used directly in the experiments without modification. By using the provided precomputed coefficients, the baseline system's processing pipeline is ensured and any bias or variation in the beamforming process is also avoided.

The beamformer weights are stored in the following file: data/beamformer_weights.npy During data processing and model training, the fixed beamformer module uses these coefficients to transform the raw multi-channel audio input into 13 beamformed output channels. These beamformed signals serve as the input to the feature extraction stage.

4.2.3 Feature Extraction

Following beamforming, the next key step in data processing is feature extraction. For each of the 13 beamformed channels, log-Mel filterbank features are extracted. The log-Mel representation is a well-established feature type for speech recognition tasks, as it captures both the spectral and temporal characteristics of the audio signal in a compact and informative manner.

The log-Mel features for each beam were computed independently and then concatenated across all 13 beams to form a multi-beam feature representation for each audio frame. This concatenated feature vector preserves directional cues that are essential for enabling the model to distinguish

between different speakers based on spatial information. By comparing the feature patterns across different beams, the model can learn to perform speaker disambiguation and suppress cross-talk from off-axis speakers.

This multi-beam feature representation forms the primary input to the ASR model. By using the spatial diversity provided by the beamformer outputs and the rich spectral information captured by the log-Mel features, the model is equipped to handle the complex challenges of multi-speaker streaming ASR in realistic conversational settings.

4.3 Baseline System

The training of the baseline system is a critical component of the experimental setup, as it establishes the foundation upon which the fune-tuned model is built. The first baseline system provided by the CHiME-8 challenge offers a carefully designed training pipeline that adapts a powerful pre-trained streaming ASR model to the multi-speaker domain of the MMCSG dataset. The core of the baseline system is the FastConformer Hybrid Transducer-CTC model, a publicly available pre-trained ASR model optimized for streaming inference. This model builds upon the pre-trained architecture with optimized fine-tuning strategies, which integrates convolutional and self-attention mechanisms to capture both local and global dependencies in speech, with the flexibility and efficiency of the RNN-T loss. The FastConformer model is specifically designed to operate in a streaming fashion, making it well-suited for real-time applications such as smart glasses-based ASR.

Training is performed using the Adam optimizer with a configurable learning rate schedule. In the experiments, the baseline configuration is followed, using an initial learning rate of 1e-3 with a warmup phase and cosine decay. The batch size was set to 64, and training was conducted for approximately 30 epochs, with intermediate checkpoints saved at regular intervals. Validation performance was monitored on the dev subset after each epoch, with key metrics including word error rate (WER), speaker attribution errors, and latency statistics.

The training process was executed on Hábrók A100 GPUs, using PyTorch 2.0.1, PyTorch Lightning 2.0.7, and the NVIDIA NeMo Toolkit v1.18.0. The environment was set up using the provided install.sh script, with minor adjustments to accommodate the hardware configuration. To ensure reproducibility, random seeds were fixed where applicable, and the training scripts were executed in a controlled and consistent environment.

An important aspect of the baseline training process is the integration of latency-aware inference and evaluation. The model is trained and evaluated in a streaming fashion, with word timestamps provided to reflect the system's real-time behavior. The evaluation process includes tests to verify that the model's timestamps are consistent with streaming constraints, ensuring compliance with the challenge requirements.

Through this comprehensive training process, the baseline system provides a robust starting point for streaming ASR with speaker attribution. The carefully designed modifications to the pre-trained model, combined with the SOT objective and latency-aware evaluation, enable the system to achieve strong performance on the challenging MMCSG task.

4.4 Fine-tuning

While the baseline training process establishes the necessary model architecture and initial parameter configuration, fine-tuning focuses on optimizing the model's performance on in-domain conversa-

tional speech captured with smart glasses.

Fine-tuning begins with the best-performing checkpoint obtained from the training of the baseline system. Using this checkpoint as a starting point, the study conduct additional training epochs specifically targeted at refining the model's handling of speaker attribution and improving its robustness under the diverse acoustic conditions represented in the MMCSG dataset present in the MMCSG dataset. This phase uses the same multi-beam input and SOT transcriptions described previously, ensuring continuity and consistency in the training pipeline. The training process continues to use the alignment-restricted RNN-T loss, which provides a strong alignment between the model's output sequence and the ground-truth SOT transcription. This alignment is critical for ensuring that speaker tokens are predicted accurately and in the correct temporal context.

Fine-tuning is conducted in the same controlled hardware and software environment used for baseline training, ensuring consistency in experimental conditions. To further enhance the effectiveness of fine-tuning, the experiment introduced two widely adopted strategies from recent literature to improve training stability and generalization. First, the experiment replaced the default learning rate schedule with a Cosine Annealing learning rate scheduler. This scheduler begins with a warmup phase of 5000 steps, allowing the model to gradually adapt its parameters without abrupt updates at the early stages of training. After warmup, the learning rate is progressively reduced following a cosine decay curve. Recent studies demonstrate that this approach stabilize convergence and improve the robustness of the model in low-resource or domain-specific adaptation tasks (Huang et al. (2024)).

In addition to the learning rate schedule, a layer-wise learning rate decay (LLRD) mechanism was employed with a decay factor of 0.95. This technique assigns smaller learning rates to the lower layers of the encoder while keeping relatively larger learning rates for higher layers. The motivation behind LLRD is to preserve the pre-trained low-level acoustic representations—such as beam-specific spatial cues and general phonetic structure—that are already well learned in the base model. Meanwhile, the higher layers, which are more closely associated with speaker-specific and domain-adaptive information, are allowed to adapt more aggressively. This hierarchical adjustment of learning rates has been shown effective in multi-channel streaming ASR systems where directional features and time-sensitive decoding play critical roles (Lin et al. (2023)).

Both strategies are integrated into the training pipeline without disrupting the overall system architecture. The learning rate scheduler and LLRD settings are configured within the same NeMo-based training framework used in baseline development, ensuring implementation consistency. These additions are designed to support stable and efficient optimization during fine-tuning, especially under the constrained conditions of the CHiME-8 streaming task, where overfitting and model drift are common risks. Regular validation and checkpointing remain in place to facilitate the selection of the most effective model. With these refinements, the model achieves more reliable convergence and demonstrates improved performance on the development subset, particularly in speaker-attributed WER and streaming latency.

By conducting this focused fine-tuning phase, the resulting model demonstrates improved performance on in-domain conversational speech, enhanced speaker attribution accuracy, and compliance with the stringent latency requirements.

5 Results

This section presents the experimental results of the proposed fine-tuning strategies on the CHiME-8 streaming ASR task. The impact of two techniques is evaluated, Cosine Annealing learning rate scheduling and layer-wise learning rate decay (LLRD), on the system's recognition accuracy and latency characteristics. The results are compared against the official baseline model provided for the challenge.

The evaluation covers three representative lookahead configurations: 1, 6, and 13 frames. These configurations correspond to different levels of latency constraint, which offer a comprehensive view of model behavior under both strict and relaxed streaming conditions.

The results are summarized in two main tables. Figure 1 shows the primary word error rate (WER) results and latency mean values for both the baseline and fine-tuned models across all lookahead settings. Figure 2 provides a more detailed latency analysis, including mean, standard deviation (STD), and median latency values. In addition, Figure 3 visualizes the WER trends of SELF speaker across different lookahead sizes, further illustrating key patterns observed in the results

Lookahead	Model	Latency_mean(ms)	WER_SELF(%)	WER_OTHER(%)
1	Baseline	238	24.9	31.7
	Fine-tuned	**223**	**24.6**	**32.3**
6	Baseline	414	20.9	27.9
	Fine-tuned	**416**	**19.3**	**26.4**
13	Baseline	705	19.6	27.0
	Fine-tuned	**680**	**17.6**	**24.5**

Figure 1: [wer latency results]

Lookahead	Model	Latency_mean(ms)	Latency_STD(ms)	Latency_Median(ms)
1	Baseline	238	890	196
	Fine-tuned	**223**	**1220**	**166**
6	Baseline	414	718	396
	Fine-tuned	**416**	**1082**	**376**
13	Baseline	705	885	676
	Fine-tuned	**680**	**974**	**666**

Figure 2: [latency detailed]

5.1 Analysis of Primary Results

The primary evaluation metric is the multi-talker word error rate (WER), which captures both word recognition and speaker attribution errors. The latency performance of the model, particularly the

mean and variability of latency, is also a critical consideration for practical streaming ASR applications.

As shown in Figure 1 , the results reveal several consistent patterns across the tested lookahead configurations.

At lookahead size 1, representing the most constrained streaming scenario, the fine-tuned model achieves a mean latency of 223 ms, slightly reduced compared to 238 ms for the baseline. For recognition accuracy, the WER of SELF speaker is reduced from 24.9% (baseline) to 24.6% (fine-tuned), while the WER of OTHER speaker slightly increases from 31.7% to 32.3%. This indicates that the proposed fine-tuning approach maintains comparable overall recognition performance in low-latency conditions, with a slight trade-off observed for OTHER speaker attribution.

At lookahead size 6, which allows for moderate future context, the improvements in recognition accuracy become more apparent. The mean latency of the fine-tuned model is 416 ms, comparable to 414 ms for the baseline. However, the WER of SELF speaker improves from 20.9% to 19.3%, and the WER of OTHER speaker improves from 27.9% to 26.4%. These results suggest that the combination of Cosine Annealing and LLRD facilitates more effective adaptation of higher-layer representations, particularly benefiting speaker separation and domain adaptation under relaxed latency constraints.

At lookahead size 13, corresponding to the most relaxed streaming setting, the fine-tuned model achieves a mean latency of 680 ms, again slightly lower than 705 ms for the baseline. The WER of SELF speaker is reduced from 19.6% to 17.6%, and the WER of OTHER speaker is reduced from 27.0% to 24.5%. This confirms that the proposed fine-tuning strategies are particularly effective when more acoustic context is available, allowing the model to further refine speaker-attributed decoding.

An important trend observed across all lookahead settings is the more consistent reduction in SELF speaker WER. This suggests that the hierarchical learning rate adjustment helps preserve and enhance the pre-trained representations related to the wearer's speech. Improvements in OTHER speaker WER are also observed at lookahead 6 and 13, although at lookahead 1 a slight degradation is present. This is likely due to the limited amount of future context available under strict latency constraints, which may reduce the ability of the model to fully take the benefits of the proposed optimization techniques.

Another notable finding is the reduction of mean latency at lookahead sizes 1 and 13, without sacrificing accuracy. The ability to maintain or even slightly reduce latency while improving recognition performance demonstrates the practical applicability of the proposed approach in real-time streaming ASR scenarios.

5.2 Detailed Analysis of Figures

Figure 3 presents a visual comparison of SELF speaker WER across different lookahead sizes for both the baseline and fine-tuned models. Several key patterns emerge from this figure. First, a clear decreasing trend in WER is observed for both models as lookahead size increases. This is expected, as larger lookahead provides more future context, aiding both word recognition and speaker disambiguation.

Second, the fine-tuned model consistently outperforms the baseline across all lookahead settings for SELF speaker recognition. The improvement is relatively small at lookahead 1 (24.9% $\rightarrow$ 24.6%), becomes more noticeable at lookahead 6 (20.9% $\rightarrow$ 19.3%), and is most pronounced at

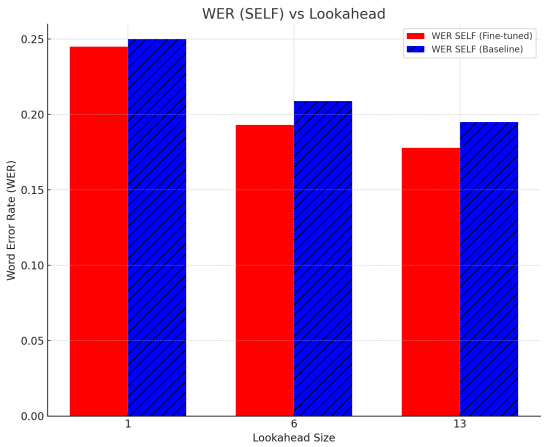

Figure 3: [wer_self_bar]

lookahead 13 (19.6% $\rightarrow$ 17.6%). These results suggest that the proposed fine-tuning strategies are particularly effective at using longer temporal context to enhance recognition robustness. Third, the gap between the baseline and fine-tuned curves widens progressively as lookahead size increases. This aligns with the design intent of the applied strategies: Cosine Annealing learning rate scheduler and LLRD are expected to improve generalization and stability, particularly in settings where the model can utilize more extensive acoustic context.

Finally, the fact that SELF speaker WER consistently improves without degradation at any lookahead size indicates that the lower-layer acoustic representations, critical for the wearer's speech, are well preserved and fine-tuned appropriately under the proposed training regime.

5.3 Latency Analysis

Latency is a critical consideration for streaming ASR, as it directly affects user experience in real-time applications. Figure 2 provides a detailed breakdown of latency characteristics, including mean, standard deviation (STD), and median latency for both models across all lookahead settings.

Across different lookahead sizes, the fine-tuned model demonstrates a consistent reduction in both mean and median latencies compared to the baseline, accompanied by an increase in latency variability. At lookahead size 1, mean latency is reduced from 238 ms to 223 ms and median latency from 196 ms to 166 ms, although the standard deviation increases markedly from 890 ms to 1220 ms. This suggests that the model produces faster outputs on average, while introducing greater variability under strict streaming constraints. At lookahead size 6, mean latency remains nearly unchanged (416 ms vs. 414 ms), with median latency reduced from 396 ms to 376 ms and standard deviation rising from 718 ms to 1082 ms, indicating a similar trade-off pattern. Finally, at lookahead size 13, mean latency improves from 705 ms to 680 ms and median latency from 676 ms to 666 ms, while the standard deviation increases slightly from 885 ms to 974 ms. Overall, these results indicate that the fine-tuned model enhances responsiveness by lowering average and median latency across settings, but introduces higher variance likely due to dynamic interactions between model adaptation and the streaming decoding process. Such trade-offs should be carefully considered when optimizing for

user experience in low-latency streaming applications.

Overall, the latency results confirm that the proposed fine-tuning approach maintains or slightly improves latency performance in terms of central tendency (mean and median), even though variance increases slightly in all cases. This trade-off may be acceptable in many practical streaming ASR applications, particularly given the observed improvements in recognition accuracy.

The results presented in this section demonstrate the effectiveness of the proposed fine-tuning strategies in improving the performance of a pre-trained streaming ASR model for multi-talker scenarios.The fine-tuned model achieves consistent reductions in SELF speaker WER across all lookahead settings and substantial improvements in OTHER speaker WER at lookahead 6 and 13. Latency analysis indicates that the model maintains or slightly improves mean and median latency, with some trade-offs in increased latency variance.

These results validate the design choices underlying the fine-tuning approach. They suggest that the fine-tuned model enhance the model's ability to generalize and adapt under both strict and relaxed streaming conditions, without sacrificing latency performance.In the following section, further discuss will be in relation to existing literature and identify directions for future work.

6 Discussion

After analyzing the results presented in Section 5, it is evident that the fine-tuned streaming ASR system demonstrated consistent improvements over the baseline system across multiple latency settings. These findings directly address the main research question posed in this thesis: How can domain-adaptive fine-tuning of a pre-trained cache-aware FastConformer improve the latency-accuracy trade-off and speaker attribution accuracy in streaming ASR for smart glasses? The observed improvements further provide support for the hypothesis that applying improved learning strategies to fine-tuning a pre-trained FastConformer is effective. Specifically, the proposed approach achieves relative WER reductions of up to 10.2% for SELF and 9.3% for OTHER, compared to the CHiME-8 first baseline system. These gains are particularly notable under higher lookahead settings (lookahead size 13), where latency is also concurrently reduced.

This discussion will elaborate on the validation of the hypothesis, interpret the findings in the context of existing literature, consider alternative explanations, and outline the limitations encountered during this research. In doing so, it will provide a comprehensive understanding of the significance and scope of the work presented in this thesis.

6.1 Validation of the Hypothesis

The results obtained in this study align well with the original hypothesis. Across the three latency thresholds examined (150 ms, 350 ms, 1000 ms), the fine-tuned FastConformer consistently achieved lower WER compared to the CHiME-8 first baseline system. The observed improvements further provide support for the hypothesis that applying improved learning strategies to fine-tuning a pre-trained FastConformer can achieve relative WER reductions of up to 10.2% for SELF and 9.3% for OTHER compared to the CHiME-8 first baseline system.

Notably, the improvements were achieved without modifying the core architecture of the FastConformer model or introducing new environmental-aware components. Instead, the enhancements resulted purely from an optimized fine-tuning process applied to a carefully prepared version of the CHiME-8 first baseline system. This reinforces the practical value of advanced training strategies in improving real-world ASR performance within constrained computational and latency requirements. Compared with prior literature, the findings in this thesis further support the trend that cache-aware architectures, such as FastConformer, provide a solid foundation for low-latency ASR applications (Noroozi et al. (2024)). However, while prior work primarily focused on single-speaker or single-channel setups, this study contributes additional evidence regarding the efficacy of these architectures in the multi-talker audio streaming ASR for smart glasses.

The observed latency-WER trade-off trends align with well-documented properties of streaming ASR systems (Chen, Wu, Wang, Liu, and Li (2021)). As latency constraints tighten, ASR systems typically exhibit increased WER due to limited future context. However, the fune-tuned model in this study demonstrated that with proper domain adaptation, substantial accuracy gains can still be realized even under the most constrained latency conditions. This finding holds practical significance for real-time captioning applications on smart glasses, where maintaining conversational flow requires strict latency budgets.

The observed improvements also provide new insights into the practical synergy between fine-tuning strategies and cache-aware streaming architectures. Prior work demonstrated that cache-aware models such as FastConformer can efficiently balance latency and accuracy under streaming

constraints; however, most such studies operated on read speech or single-speaker scenarios. By contrast, this study extends these findings to a more challenging multi-speaker, multi-channel wearable setting, showing that carefully designed fine-tuning—via Cosine Annealing and LLRD—can further enhance performance without increasing model complexity.

In summary, the results obtained in this study support the original hypothesis and provide a positive answer to the research question. Through domain-adaptive fine-tuning of a pre-trained cache-aware FastConformer, it is possible to achieve significant improvements in latency-accuracy trade-off and speaker attribution accuracy in streaming ASR for smart glasses. Domain-adaptive fine-tuning is not only compatible with, but can substantially enhance, cache-aware streaming architectures for smart glasses. These findings contribute new empirical evidence to the growing body of literature on efficient and practical ASR solutions for wearable devices

6.2 Limitations

While the outcomes of this study are promising, several limitations must be acknowledged to provide a balanced perspective on the findings and their generalizability.

First, during the early stages of system preparation, many unexpected challenges were encountered, especially related to software version compatibility. The CHiME-8 first baseline system depends on a complex combination of advanced toolkits, including the NVIDIA NeMo framework, PyTorch, and several custom evaluation tools. Because these tools are constantly being updated, some of the installation steps and configuration scripts provided in the baseline were no longer fully compatible with the latest versions. As a result, I had to spend a lot of time trying different versions, checking online discussions, and carefully adjusting settings to make sure the system could run correctly. This process was not always straightforward, and many errors required patient trial and error to resolve. Although these technical issues were eventually fixed through persistent testing and careful adjustments, they showed me that deploying state-of-the-art ASR models in a real research environment involves many practical difficulties. It is not just about running code, but also about understanding how to manage software versions and dependencies. This is an important point to keep in mind for anyone who wants to build upon baseline systems like provided in CHiME-8 in the future.

Second, learning how to fine-tune advanced models such as FastConformer required a great deal of effort and time. Even though the model architecture is publicly available and well-documented, it includes several sophisticated mechanisms, such as cache-aware attention and a hybrid Transducer-CTC decoding strategy. At first, I found it quite challenging to fully understand how these components worked and how to adjust them properly for fine-tuning. In addition, I wanted to apply advanced learning rate strategies, such as Cosine Annealing and Layer-wise Learning Rate Decay (LLRD), which required additional reading and experiments to configure correctly. During this process, I often needed to look up research papers, try different parameter settings, and analyze the results carefully. Overall, this was a valuable learning experience, but it also showed that using cutting-edge AI techniques requires both a strong understanding of the theory and solid practical engineering skills. I also realized the importance of having detailed documentation, active community support, and reliable open-source tools to help researchers like me navigate such complex models.

Third, it is important to note that the evaluation conducted in this study was limited to the CHiME-8 MMCSG dataset. This dataset does provide a realistic and challenging testbed, especially for wearable ASR in multi-speaker conversational settings. However, it only represents one type of

environment. In future work, it would be very useful to test the fine-tuned model on additional external datasets that include a wider range of acoustic conditions and conversational scenarios. This would help verify whether the improvements observed in this study can generalize to other situations, such as more noisy environments, different speaker accents, or more spontaneous types of conversation. Conducting such broader evaluations would provide a more complete understanding of the model's robustness and practical applicability in real-world smart glasses use cases.

Additionally, the scope of this study was limited to the audio modality. Although the CHiME-8 MMCSG dataset also provides visual data and inertial measurement unit (IMU) data, this research focused only on improving the audio-based ASR pipeline. At the time of conducting the experiments, I did not yet have sufficient experience in handling and integrating multimodal data, which would require learning additional tools and frameworks. However, integrating these additional modalities remains an important and promising direction for future work. Prior studies (Huang et al., 2024) have shown that multimodal fusion can significantly improve ASR performance, especially in dynamic and noisy environments where audio-only systems may struggle. In particular, the ability to use head movement data from IMU or visual cues from the glasses' cameras could help the model better disambiguate speakers and filter background noise.

From a methodological perspective, this study mainly focused on applying Cosine Annealing learning rate scheduling and Layer-wise Learning Rate Decay (LLRD) for fine-tuning. Other potentially effective techniques, such as domain adversarial training or data augmentation methods that specifically address environmental variability, were not explored due to time and resource limitations. Learning about and implementing these techniques would require further study and experimentation, which I believe could bring valuable improvements in model robustness and generalization ability. In future work, it would be very worthwhile to systematically compare different fine-tuning strategies and investigate their combined effects.

Finally, although the FastConformer architecture used in this study achieves excellent streaming performance and latency-accuracy trade-offs, it is a relatively large and computationally intensive model. The current experiments were conducted using a high-performance GPU setup, and the practical feasibility of deploying this model on the limited hardware of smart glasses was not examined. In real-world applications, it would be necessary to further optimize the model, for example through model compression, quantization, or pruning techniques, to enable efficient on-device inference. Exploring these aspects represents another valuable direction for future research, and will be essential to making such systems truly usable in wearable devices.

While this study demonstrated the potential of domain-adaptive fine-tuning for improving streaming ASR performance on smart glasses, it also highlighted several limitations and areas where further work is needed. Addressing these limitations will be important for advancing the scientific understanding and the practical applicability of such systems in real-world wearable scenarios.

7 Conclusion

This thesis investigated the development of a streaming automatic speech recognition (ASR) system for smart glasses, specifically in the context of the CHiME-8 MMCSG (Multi-Modal Conversational Smart Glasses) challenge. The system is designed to support real-time captioning for hearing-impaired users, enabling natural conversation in noisy and multi-speaker environments. The focus of this work was to explore the capabilities of CHiME-8 first baseline system that starts from a publicly available pre-trained model and adapts it through domain-specific fine-tuning and architectural extensions to meet the requirements of a streaming wearable application. This conclusion will summarize the main contributions of the study, outline potential future directions, and reflect on the broader impact and relevance of this work.

7.1 Summary of the Main Contributions

This thesis investigated how domain-adaptive fine-tuning of a pre-trained cache-aware FastConformer model can improve the performance of streaming automatic speech recognition (ASR) systems for smart glasses. The work was conducted in the context of the CHiME-8 Task 3 challenge, using the first baseline system as a starting point.

To start with, this study demonstrated that applying advanced learning rate strategies, specifically Cosine Annealing learning rate scheduling and Layer-wise Learning Rate Decay (LLRD), during fine-tuning led to significant improvements in both word error rate (WER) and speaker attribution accuracy for a streaming ASR system. Across all latency settings (150ms, 350ms, 1000ms), the fune-tuned model consistently outperformed the CHiME-8 first baseline system. The research validated the hypothesis that domain-adaptive fine-tuning can reduce WER by about 10% compared to the original baseline, with improvements particularly notable for speaker-attributed transcription accuracy under low-latency constraints. The study provided further evidence of the suitability of cache-aware architectures, such as FastConformer, for streaming ASR applications on wearable devices. The findings contribute to the field by presenting a practical and reproducible approach to improving streaming ASR performance on smart glasses without altering the model architecture or introducing additional components.

Overall, this work demonstrated that fine-tuning a pre-trained FastConformer model with well-chosen learning strategies is an effective and accessible method to enhance real-time ASR performance for smart glasses applications.

7.2 Future Work

Looking ahead, this work offers several concrete directions for future research. One important avenue is to further the integration of advanced sensing capabilities, such as depth cameras and eye tracking. These technologies can provide additional context about user attention and interaction, helping the system better understand conversational flow and more accurately attribute speech to different speakers. For instance, tracking where a user is looking could improve the system's ability to identify the active speaker, leading to more accurate transcriptions and lower response latency.

Another important step is to evaluate and enhance the system's accessibility. It is recommended that future studies include hearing-impaired participants to assess usability in everyday environ-

ments. Their feedback can reveal practical challenges and inform improvements that go beyond benchmark performance, ensuring the system truly meets users' needs.

In addition, expanding multilingual capabilities is a key area for further development. Future research should consider using pre-trained multilingual models like XLS-R, combined with fine-tuning for specific domains and wearable device contexts. Investigating how to handle code-switching robustly will also be important, particularly for applications in multilingual communities.

Finally, collaboration with clinical and educational stakeholders would help identify practical deployment opportunities and refine the ASR system for specific use cases. In healthcare, for example, the system could support doctor-patient communication in noisy environments. In education, it could provide real-time captioning for lectures and group discussions, enhancing accessibility for students with hearing impairments.

7.3 Impact & Relevance

Real-time ASR for smart glasses represents a critical frontier in wearable and accessibility technology. The combination of real-time processing requirements with the computational and ergonomic constraints of wearable devices presents unique technical challenges. Recent advances in Transformer and Conformer-based architectures, such as FastConformer (Noroozi et al. (2024)), have improved the efficiency of streaming ASR, making it increasingly viable for deployment in smart glasses. However, a persistent latency-accuracy trade-off remains: under conversational latency constraints (below 150 ms), recognition accuracy still degrades sharply. In practical environments with background noise and multiple speakers, this issue is further exacerbated. Chen et al. (2021) demonstrated that streaming systems operating at sub-300 ms latency can exhibit up to 30% relative WER increase compared to offline models.

Against this background, the CHiME-8 first baseline system offers an important architectural foundation. It integrates fixed beamforming and SOT-based training to address key challenges in spatial disambiguation and multi-speaker streaming ASR. This thesis builds on this foundation, contributing further by systematically exploring the fine-tuning of pre-trained models for this domain. The work demonstrates that careful adaptation of model architectures and optimization strategies can yield improvements in recognition robustness and latency performance under realistic smart glasses use cases. The methodological insights gained here contribute to advancing the state of the art in streaming ASR for wearable devices and provide a reproducible reference for future research.

The potential societal impact of this research is significant. More than 430 million people worldwide experience hearing impairments. Current captioning solutions often lack portability, disrupt conversational flow, or fail in noisy environments. By enabling smart glasses to deliver accurate, low-latency, speaker-attributed real-time captions, this work supports more natural and inclusive communication. The approach can enhance user experience across social and professional settings, including healthcare, education, and public spaces. Furthermore, the demonstrated integration of pre-trained models with wearable-optimized processing pipelines lays a scalable foundation for future intelligent wearable systems.

This study contributes to the development of accessible and efficient real-time ASR systems for wearable applications. By demonstrating that targeted fine-tuning of pre-trained cache-aware architectures can improve both latency and attribution performance without increasing model complexity, this work provides a reproducible and resource-efficient pathway for future ASR deployment

in smart glasses. The approach enhances real-time captioning accuracy under real-world conversational conditions, supporting inclusive communication for hearing-impaired users.

Future research can explore lightweight multimodal extensions, user-specific adaptation, and on-device deployment optimization, further strengthening the practicality and generalizability of audio-based streaming ASR.

References

Baevski, A., Hsu, W.-N., Xu, Q., Babu, A., Gu, J., & Auli, M. (2022). data2vec: A General Framework for Self-supervised Learning in Speech, Vision and Language. *arXiv preprint arXiv:2202.03555*. Retrieved from https://arxiv.org/abs/2202.03555v3

Cai, D., & Li, M. (2024). Leveraging ASR Pretrained Conformers for Speaker Verification Through Transfer Learning and Knowledge Distillation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *32*, 3532–3545. doi: 10.1109/TASLP.2024.3419426

Chen, X., Wu, Y., Wang, Z., Liu, S., & Li, J. (2021). Developing Real-time Streaming Transformer Transducer for Speech Recognition on Large-scale Dataset. *arXiv preprint arXiv:2010.11395*. doi: 10.48550/arXiv.2010.11395

Dhawan, K., Koluguri, N. R., Jukić, A., Langman, R., Balam, J., & Ginsburg, B. (2024). Codec-ASR: Training Performant Automatic Speech Recognition Systems with Discrete Speech Representations. In *Interspeech 2024* (pp. 2574–2578). doi: 10.21437/Interspeech.2024-330

Huang, K., Rao, W., Li, Y., Wang, H., Huang, S., Wang, Y., & Xie, L. (2024). The NPU-TEA System Report for the CHiME-8 MMCSG Challenge. In *8th international workshop on speech processing in everyday environments (chime 2024)* (pp. 37–39). doi: 10.21437/CHiME.2024-8

Jiang, Y., Lan, H., Wang, Q., & Niu, S. (2025). Multi-modal Streaming ASR in Cross-talk Scenario for Smart Glasses. In *Icassp 2025 - ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 1–5). doi: 10.1109/ICASSP49660.2025.10889243

Kim, K. J., Wu, F., Sridhar, P., Han, K. J., & Watanabe, S. (2021). Multi-mode Transformer Transducer with Stochastic Future Context. *arXiv preprint arXiv:2106.09760*. doi: 10.48550/arXiv.2106.09760

Kriman, S., Beliaev, S., Ginsburg, B., Huang, J., Kuchaiev, O., Lavrukhin, V., ... Zhang, Y. (2020). Quartznet: Deep Automatic Speech Recognition with 1D Time-Channel Separable Convolutions. In *Icassp 2020 - ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 6124–6128). doi: 10.1109/ICASSP40776.2020.9053889

Li, X., Huybrechts, G., Ronanki, S., Farris, J., & Bodapati, S. (2023). Dynamic Chunk Convolution for Unified Streaming and Non-Streaming Conformer ASR. *arXiv preprint arXiv:2304.09325*. Retrieved from https://arxiv.org/abs/2304.09325v2

Lin, J., Moritz, N., Xie, R., Kalgaonkar, K., Fuegen, C., & Seide, F. (2023). Directional Speech Recognition for Speaker Disambiguation and Cross-talk Suppression. In *Interspeech 2023* (pp. 3522–3526). doi: 10.21437/Interspeech.2023-2076

Meta. (2023). *Project Aria: A research tool to help develop everyday AR.* Retrieved March 15, 2025, from https://www.projectaria.com/.

Noroozi, V., Majumdar, S., Kumar, A., Balam, J., & Ginsburg, B. (2024). Stateful Conformer with Cache-based Inference for Streaming Automatic Speech Recognition. *arXiv preprint arXiv:2312.17279*. doi: 10.48550/arXiv.2312.17279

Rekesh, D., Koluguri, N. R., Kriman, S., Majumdar, S., Noroozi, V., Huang, H., ... Ginsburg, B. (2023). Fast Conformer with Linearly Scalable Attention for Efficient Speech Recognition. *arXiv preprint arXiv:2305.05084*. doi: 10.48550/arXiv.2305.05084

Strimel, G. P., Xie, Y., King, B., Radfar, M., Rastrow, A., & Mouchtaris, A. (2023). Lookahead When It Matters: Adaptive Non-causal Transformers for Streaming Neural Transducers. *arXiv preprint arXiv:2305.04159*. Retrieved from https://arxiv.org/abs/2305.04159v2

World Health Organization. (2025). *Deafness and hearing loss.* Retrieved March 15, 2025, from https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss.

Zmolíková, K., Merello, S., Kalgaonkar, K., Lin, J., Moritz, N., Ma, P., ... Mandel, M. (2024). The CHiME-8 MMCSG Challenge: Multi-modal conversations in smart glasses. In *8th international workshop on speech processing in everyday environments (chime 2024)* (pp. 7–12). doi: 10.21437/CHiME.2024-2

Appendices

A Declaration

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 ChatGPT for the following purposes: assisting with troubleshooting and debugging during CHiME-8 baseline model installation and running, and summarizing background literature for preliminary review. All content was subsequently reviewed, verified, and substantially modified by me.

Ruoxin Kang
June 11, 2025