MULTILINGUAL SPEECH RECOGNITION VIA ATTENTION ENCODER DECODER

THESIS

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE AWARD OF THE DEGREE
OF

MASTER OF TECHNOLOGY
IN
ARTIFICIAL INTELLIGENCE

Submitted by

ABHISHEK DUBEY (2K23/AFI/09)

Under the supervision of

DR. MANOJ KUMAR

DEPARTMENT OF COMPUTER SCIENCE ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi 110042

MAY, 2025

DEPARTMENT OF COMPUTER SCIENCE ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

CANDIDATE'S DECLARATION

I, **ABHISHEK DUBEY**, Roll No – **2K23/AFI/09** students of M.Tech (COMPUTER SCIENCE ENGINEERING),hereby declare that the project Dissertation titled "**MULTILINGUAL SPEECH RECOGNITION VIA ATTENTION ENCODER DECODER**" which is submitted by me to the computer science engineering, Delhi Technological University, Delhi in partial fulfilment of the requirement for the award of degree of Master of Technology, is original and not copied from any source without proper citation. This work has not previously formed the basis for the award of any Degree, Diploma Associateship, Fellowship or other similar title or recognition.

Place: Delhi ABHISHEK DUBEY

Date: 30/05/2025

DEPARTMENT OF COMPUTER SCIENCE ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

CERTIFICATE

I hereby certify that the Project Dissertation titled "**MULTILINGUAL SPEECH RECOGNITION VIA ATTENTION ENCODER DECODER**" which is submitted by ABHISHEK DUBEY, Roll No – 2K23/AFI/09, Computer Science Engineering ,Delhi Technological University, Delhi in partial fulfilment of the requirement for the award of the degree of Master of Technology, is a record of the project work carried out by the students under my supervision. To the best of my knowledge this work has not been submitted in part or full for any Degree or Diploma to this University or elsewhere.

Place: Delhi DR. MANOJ KUMAR

Date: 30/05/2025 **SUPERVISOR**

DEPARTMENT OF COMPUTER SCIENCE ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

ACKNOWLEDGEMENT

I wish to express our sincerest gratitude to Dr Manoj Kumar for his continuous guidance and mentorship that he provided us during the project. He showed us the path to achieve our targets by explaining all the tasks to be done and explained to us the importance of this project as well as its industrial relevance. He was always ready to help us and clear our doubts regarding any hurdles in this project. Without his constant support and motivation, this project would not have been successful.

Place: Delhi Abhishek Dubey

Date: 30/05/2025

Abstract

Speech-based interfaces have become a revolutionary way to enhance clinical documentation, telemedicine accessibility, and doctor-patient communication in the rapidly changing field of healthcare technology. Nevertheless, the majority of current Automatic Speech Recognition (ASR) systems cover monolingual scenarios and are frequently designed for general-purpose jobs. This significantly reduces their suitability for use in actual healthcare settings where multilingual and accent-diverse communication is commonplace. In order to fill this void, this thesis presents MultiMed, a comprehensive, multilingual dataset created especially for medical speech recognition in five different languages: Mandarin Chinese, English, German, French, and Vietnamese. More than 150 hours of annotated clinical speech that was gathered from actual healthcare situations and enhanced with linguistic, demographic, and acoustic diversity make up the dataset. The paper investigates and assesses cutting-edge ASR architectures built on the Attention Encoder-Decoder (AED) framework in order to make efficient use of this dataset. It specifically optimizes several Whisper model variations (Tiny, Base, Small, Medium), which were first created by OpenAI, in both monolingual and multilingual training environments. In order to assess the accuracy and efficiency of the architecture, comparative tests are also conducted against Hybrid ASR systems, such as wav2vec 2.0 with shallow-fusion language models. Additionally, the thesis examines two different fine-tuning techniques that aim to strike a balance between recognition performance and computational efficiency: Decoder-Only Fine-Tuning and Full Encoder-Decoder Training.

Contents

List of Tables

List of Figures

Chapter 1

INTRODUCTION

1.1 Background

Automatic Speech Recognition (ASR) is now an important part of how people and computers talk to each other, especially in areas where voice input needs to be fast and accurate. The medical field is one of these areas where ASR systems could have a big impact. In clinical settings, good speech-to-text systems can help with real-time transcription of conversations between patients and doctors, make it easier for healthcare workers to keep track of their work, and make healthcare services more accessible, especially in places where there are not many resources and people speak more than one language. The need for multilingual support in ASR systems is growing because healthcare delivery around the world is becoming more and more diverse. Most ASR models only work well in situations where only one language is spoken. This makes them less useful in real-world medical settings where many languages, accents, and dialects are spoken.

1.2 Problem Statement

There are a lot of ASR systems that can recognize speech in general, but not many that can handle the unique problems that come with multilingual medical speech. Some of these problems are domain-specific vocabulary, different accents, background noise in clinical settings, and the lack of labelled datasets. Also, traditional ASR models often do not find the right balance between accuracy and speed, which makes them hard to use in places with limited resources, like rural clinics. This thesis fills in these gaps by introducing and testing MultiMed, a new multilingual ASR dataset and benchmark made for the medical field. It also looks into how well different ASR architectures work, especially the Attention Encoder-Decoder (AED)models.

1.3 Objectives

To publish MultiMed, a high-quality multilingual medical speech recognition corpus for Vietnamese, English, German, French, and Mandarin Chinese at scale. To study the Attention Encoder-Decoder (AED) model and compare its performance with Hybrid ASR models. To investigate the effect of monolingual versus multilingual fine-tuning on model performance. In order to conduct a layer-by-layer ablation experiment of AED models to investigate their performance and computational trade-offs. In order to perform an error analysis for learning the limitations of multilingual medical ASR systems from both the model-design and linguistic perspectives.

1.4 Motivation

The motivation behind this work stems from the urgent necessity to improve communication between healthcare professionals and patients in multilingual regions. Through the establishment of robust multilingual ASR systems specific to the healthcare sector, we can enable the removal of language barriers in medicine and maximize diagnostic precision and workflow effectiveness together. Furthermore, this effort helps the open science community by releasing all relevant code, models, and data openly to promote transparency, reproducibility, and additional innovation in this important field.

1.5 Scope Of Work

This thesis responds to the following: Collection and quality control of a diversified multilingual medical speech data set. Training and fine-tuning Whisper ASR models (Tiny, Base, Small, Medium) on decoder-only and encoder-decoder full methods. Evaluating based on reference measures like Word Error Rate (WER) and Character Error Rate (CER). Comparative analysis of Hybrid and AED architectures. Analysis of training methods, computational trade-offs, and deployability feasibility in real-world settings.

Chapter 2

LITERATURE REVIEW

Automatic Speech Recognition (ASR) underwent a significant transformation in the last decade, with developments in deep learning and large-scale corpora availability playing a central role. The early ASR systems, especially in the general domain, were based on Hybrid models of Hidden Markov Models (HMM) and Gaussian Mixture Models (GMM), followed by more capable Deep Neural Networks (DNNs). With the newer architectures of Convolutional Neural Networks (CNNs), Long Short-Term Memory networks (LSTMs), and Transformer-based models entering the scene, the domain of ASR saw the move towards the end-to-end learning paradigms. In the medical domain, however, ASR finds itself grossly underutilized due to the sensitive nature of health-related information and absence of access to publicly usable annotated corpora. Research studies like those of Hodgson and Coiera (2016) indicated the potential of ASR in automating clinical documentation and enhancing the efficiency of doctors, particularly during peak pressure moments like pandemic outbreaks. However, access to multilingual and medically relevant speech corpora remained limited. For instance, Fareez et al. (2022) suggested an English medical ASR dataset based on simulated data but limited to respiratory disease scenarios and English-speaking populations from the West England population, thus limiting the generalizability scope. PriMock57 (Korfiatis et al., 2022) and myMediCon (Htun et al., 2024) provided small-scale simulated corpora, thus further reflecting the absence of access to real-world, multilingual, medical-grade speech corpora.

Multilingual ASR systems have been an essential need in the last couple of years. Work such as Baevski et al. (2020) and Conneau et al. (2021) introduced self-supervised models such as wav2vec 2.0 and XLSR-53, pre-trained on multilingual, large-scale unlabeled audio and fine-tuned for target languages or tasks. While such approaches improve low-resource language performance, they are not applied in domain-specific tasks such as medical transcription. Whisper architecture (Radford et al., 2023) by OpenAI is another primary contribution. It is pre-trained on 680,000 hours of labeled multilingual audio and is a general-purpose base for downstream ASR tasks for languages. Not much work is, however, done in fine-tuning and testing such models for medical use cases. Hybrid ASR models that include pretrained acoustic encoders with shallow or deep fusion language models remain data-efficient and robust, as indicated in work by Lüscher et al. (2019) and Zeyer et al. (2019), but yet to be thoroughly tested on multilingual medical use cases.

In general-domain multilingual ASR research, most of the effort has been given to general-domain data, with little evidence for the effect of linguistic features such as tones, accents, and speech roles on performance in the clinical domain. While code-switching, accent switching, and phonological ambiguity have been explored in general ASR, these features have not been studied in medical speech with regularity. Furthermore, benchmarks used are typically not diverse in recording conditions, speaker populations, and medical vocabulary coverage. This thesis bridges these gaps by capitalizing on recent progress in multilingual speech recognition and applying it to the medical domain. Specifically, it generalizes comparative evaluation of AED and Hybrid architectures, provides reproducible baselines, and reports empirical studies of multilinguality and monolinguality trade-offs on real-world data.

The novelty of this paper is not just the introduction of the first large-scale multilingual medical dataset for ASR—MultiMed—but also the first systematic benchmarking of AED against Hybrid models in medicine on five linguistically diverse languages. A systematic layer-wise ablation study and linguistic error analysis, part of this research, distinguishes it further, shedding light into the optimization of multilingual ASR models for healthcare applications. By leveraging the synergy between state-of-the-art deep learning architectures and practical evaluation methods, this literature review establishes the context and motivation for the following methodology and experiments.

Chapter 3

DATASET

Representativeness, diversity, and quality of the training and test dataset employed do have a significant impact on the performance of any Automatic Speech Recognition (ASR) system. We make use of and expand upon the MultiMed dataset—a medically domain-oriented, multilingual ASR dataset of five languages: Vietnamese, English, German, French, and Mandarin Chinese—through this thesis. The dataset has been created with the aim of facilitating large-scale, realistic experimentation in multilingual medical ASR research.

3.1 Dataset Overview

Data was collected from actual medical audio recordings, i.e., actual medical YouTube channels, to maintain domain applicability. Recordings contain doctor-patient conversations, medical consulting, healthcare interviews, and narrations. The transcripts were prepared by human annotators, which were then checked by domain experts to maintain quality.

Attribute	**value**
Total duration	150 hours
Languages	Vietnamese, English, German, French, Chinese
Recording environments	10 distinct acoustic conditions
Speaker accents	10 distinct acoustic conditions
Speaking roles	6 roles (Doctor, Patient, Nurse, Narrator, etc.)
Transcript verification	Human annotated + Medical expert reviewed
Data distribution	Train / Validation / Test (non-overlapping speakers)

Table 3.1: Sample of data collected

3.2 Sample Distribution By Language

The MultiMed dataset was purposefully created to offer multilingual, balanced coverage of spoken information that is pertinent to medicine. Vietnamese, English, German, French, and Mandarin Chinese are the five linguistically varied languages that make up the collection. Each of these languages contributes a significant number of audio samples and hours of cumulative time. In addition to reflecting the diversity of clinical communication around the world, this multilingual structure facilitates cross-lingual study in medical

Language	**Duration**	**No of samples**	**Dev/Test**
Vietnamese	30	6000	20.05/25.43
English	40	8200	19.01/19.41
German	28	5800	18.90/17.92
French	25	5100	34.89/31.05
Chinese	27	5400	23.95/34.28
Total	**150**	**30,500**	

Table 3.2: Sample distribution by language

One of the main low-resource languages in this dataset is Vietnamese, which has 6,000 samples and 30 hours of audio. In multilingual training conditions, Vietnamese demonstrated a significant gain in performance despite having fewer public ASR resources than English or German, indicating the potent advantages of cross-lingual transfer learning. The significance of linguistic context-sharing was highlighted by the best test WER for Vietnamese, which was reported at 25.43 percent and further improved to 20.05 percent under multilingual fine-tuning.

With 40 hours and 8,200 utterances, English is the language with the greatest data in the dataset. With a test WER of 19.41 percent and a somewhat better performance of 19.01 percent under monolingual fine-tuning, the system attained the highest overall accuracy here, thanks to the Whisper model's initial pretraining on a sizable corpus of English data. This demonstrates that even in situations with abundant resources, domain-specific adaptation is effective.

Together, German and French provide more than 50 hours of content. Out of all the languages in the test set, German produced the lowest WER (17.92 percent) after 28 hours and 5,800 utterances, demonstrating the critical roles that both model pretraining and high-quality transcribing consistency have in performance. French showed greater error rates (WER: 31.05 percent), although having comparable quantities (25 hours, 5,100 samples).

Because of the logographic structure of its writing system, Mandarin Chinese is unique in that its performance is assessed using Character Error Rate (CER) rather than WER. Mandarin obtained a CER of 34.28 percent in the test set after 27 hours and 5,400 samples. Despite being somewhat high, this result is in line with the difficulties caused by character ambiguity, homophones, and tone-dependent recognition. The model achieved a CER of 30.88 percent under multilingual training, underscoring once more the significance of shared feature representation in cross-lingual ASR.

3.3 Data Quality Control

During the collection and annotation stages, a strict data quality control procedure was put in place to guarantee that the MultiMed dataset offers a solid basis for training high-performance medical ASR models. A major worry was ensuring correctness, consistency, and domain relevance across all transcripts due to the delicate and specialized nature of clinical speech. Although the original audio recordings came from reputable medical

YouTube channels and health-related speech segments, the raw recordings frequently contained filler information, background noise, overlapping speech, or off-topic conversation. These artifacts were meticulously removed using a pipeline of preprocessing and manual examination.

To guarantee that every sample recorded a unique medical context or phrase, each audio file in the dataset was divided into utterances according to speaker turns and semantic bounds. Professional annotators prepared verbatim transcripts in the speaker's native tongue as part of the transcribing process. After that, multilingual medical professionals who were proficient in the relevant languages manually reviewed each transcript to verify medical terminology, fix transcription errors, and harmonize terminology across dialects and geographical areas.

A number of normalizing procedures were used to improve model compatibility and transcription uniformity. Among these were lowercasing, enlarging contractions, eliminating punctuation (unless it was medically necessary), and standardizing abbreviations, drug names, and units of measurement. Furthermore, a thorough anonymization procedure was carried out to exclude any personally identifying information (PII), including patient names, addresses, and contact information, from the audio and transcripts. This guaranteed adherence to institutional review guidelines and ethical data handling procedures.Acoustic feature thresholds were used to examine the audio's noise levels, and samples with a lot of static, reverberation, or ambiguous speech were not included. Additionally, utterances with unclear or insufficient speech were eliminated in order to preserve the quality of the transcript. Weighted sampling was used to further balance the dataset across languages and speaking roles, preventing model bias toward overrepresented classes.

Last but not least, the consistency and completeness of the metadata linked to each sample—such as speaker gender, language, accent region, and recording environment—were verified twice. Later phases of the project were made possible by the fine-grained assessments and demographic-specific analyses made possible by this metadata. The MultiMed dataset is a reliable source for training strong, multilingual ASR systems for practical healthcare applications because it meets a high standard of linguistic, acoustic, and clinical relevance through this multi-stage quality assurance pipeline.

3.4 Preprocessing Pipeline Data

A structured preprocessing pipeline was created to clean, normalize, and standardize the audio and transcript data in order to get the MultiMed dataset ready for training reliable multilingual ASR models. Using speaker pauses and silence detection, lengthy medical recordings were first divided into more manageable, meaningful utterances. This guarantees that every audio sample is a logical speech unit that can be consumed by models. To keep the quality of the audio consistent across different recording environments and acoustic conditions, it was normalized by changing the sample rate to 16 kHz, adjusting the volume, and reducing background noise.

To guarantee alignment with the matching audio, a rigorous cleaning procedure was used on the transcript side. This involved removing unnecessary characters or filler words, expanding abbreviations, normalizing text by changing it to lowercase, and making sure that formatting was consistent across languages. Byte Pair Encoding (BPE), which divides text into subword units to facilitate handling of multilingual vocabulary, uncommon medical terms, and spelling variations, was used for tokenization. Furthermore, anonymization was done to eliminate any personally identifiable information, guaranteeing data privacy and ethical compliance. Finally, the dataset was split into training, validation, and test sets with non-overlapping speakers to avoid information leakage and ensure a fair evaluation framework.

Chapter 4

METHODOLOGY

This chapter outlines technical design and implementation strategy utilized in the development and test of multilingual medical Automatic Speech Recognition (ASR) systems using the MultiMed dataset. The approach has the goal of systematically examining the performance of various ASR model architectures, fine-tuning techniques, and training hyperparameters in monolingual and multilingual settings. The focus is primarily on models with the Attention Encoder-Decoder (AED) architecture and comparative assessment with Hybrid ASR models. We take Whisper, a Transformer-based ASR model family pre-trained on 680,000 hours of multilingual speech, as the baseline architecture for our tests.

4.1 Model Architecture Overview

This study's central model architecture is based on the Attention Encoder-Decoder (AED) framework, which is implemented with OpenAI's Whisper architecture. Whisper is an end-to-end ASR model that is Transformer-based and has proven to perform well in a wide range of acoustic conditions and languages. Its architecture's capacity to capture long-range dependencies in both the audio input and the textual output makes it especially well-suited for multilingual speech recognition tasks. The model works by first transforming unprocessed audio signals into a log-Mel spectrogram, which is then used as the encoder's input. The encoder is made up of several Transformer blocks that process the spectrogram concurrently and gradually learn contextualized representations of the speech input.

The decoder, which is also constructed using Transformer layers and functions in an autoregressive fashion, receives these encoded features after that. One by one, the decoder creates output tokens while using a multi-head attention mechanism to attend to the encoder's output as well as the previously generated tokens. The model is able to match audio characteristics with the appropriate linguistic structure in the target language thanks to this attention-based interaction between the encoder and decoder. The output transcript is also divided into subword units using a Byte Pair Encoding (BPE) tokenizer. While maintaining a small and computationally manageable vocabulary, this tokenizer option offers improved handling of uncommon and domain-specific terms, especially those frequently used in medical speech.

The Whisper architecture provides flexibility in adjusting the model according to computational limitations and the availability of labeled data by supporting both full encoder-decoder training and decoder-only fine-tuning. The model is trained using cross-entropy loss with teacher forcing, which speeds up convergence during fine-tuning, and takes an 80-channel log-Mel spectrogram as input. This work evaluated Whisper variants of various sizes, varying in terms of model parameters and the number of Transformer layers: Tiny, Base, Small, and Medium. These variations make it possible to compare accuracy and computational efficiency empirically, which aids in determining the most workable setups for actual implementation in healthcare applications.

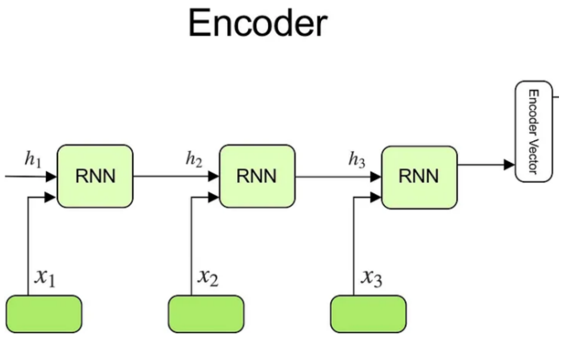

Figure 4.1: Encoder Architecture

The diagram you've shared illustrates the working of an RNN-based encoder, which is a core component of many sequence-to-sequence models used in fields like speech recognition, language translation, and even text summarization. The encoder's main job is to take in a sequence of inputs—such as audio frames or words—and convert them into a single meaningful representation known as the encoder vector. This vector is essentially a summary of everything the model has seen in the input sequence, and it becomes the foundation for the next stage, typically a decoder.

In this setup, the input features, labeled x1,x2,x3 are fed into a chain of Recurrent Neural Network (RNN) cells. Each RNN cell processes one input at a time and maintains a hidden state (represented as h1,h2,h3etc.) that captures the context from previous steps. This way, the model doesn't just look at the current input in isolation—it also considers what came before it. As the input progresses through each RNN layer, the model gradually builds up a deeper understanding of the entire sequence.

What makes this architecture powerful is its ability to remember earlier parts of the sequence while processing later ones. For example, in a spoken medical sentence, early mentions of symptoms or patient details can influence how the rest of the sentence is interpreted. The final RNN in the chain produces the last hidden state, which acts as the encoder output vector—a compressed form of all the input information. This vector is then passed on to the decoder, which generates the output sequence, such as a medical transcription.

Although this model uses standard RNNs, more advanced versions may replace these with LSTMs, GRUs, or Transformer encoders for better handling of long sequences and complex patterns. Still, this diagram offers a clear and intuitive look at how an encoder processes input step-by-step and builds a representation that captures the overall meaning of the sequence.

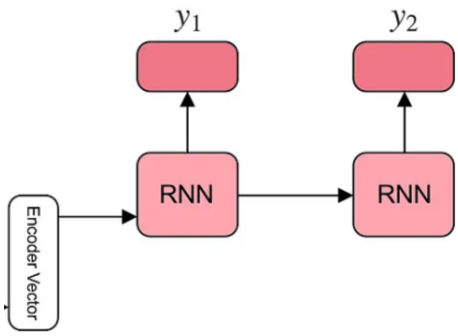

Figure 4.2: Decoder Architecture

The diagram illustrates a decoder architecture based on Recurrent Neural Networks (RNNs), which is an essential component of sequence-to-sequence (seq2seq) models. While the encoder processes the input sequence and converts it into a condensed representation (the encoder vector), the decoder takes over from that point and uses this information to generate the output sequence one step at a time. This architecture is widely used in applications like machine translation, speech recognition, and text generation.

As shown in the image, the encoder vector—which contains all the contextual knowledge extracted from the input—is passed as input to the first RNN block in the decoder. This vector helps the decoder understand the overall meaning of the original input. Each RNN cell in the decoder then generates one output token at a time. The first RNN unit produces the output y1 which could be the first word or symbol in a predicted sentence. This output is then fed into the next RNN unit along with the hidden state to generate y2 and the process continues sequentially until the end of the output sequence is generated.

What makes this architecture effective is the autoregressive nature of the decoder. Each output depends not only on the encoder's context but also on the outputs that came before it. This allows the model to generate coherent and grammatically accurate sequences, where earlier predictions influence the next steps. For instance, in a medical transcription task, once the model predicts the term ”blood,” it might be more likely to follow it with terms like ”pressure” or ”test,” depending on the context learned from the encoder.

Although this architecture is based on basic RNNs, it serves as a foundational approach to sequence generation. In more advanced setups, RNNs may be replaced by LSTM or GRU units to better capture long-term dependencies, or by attention-based mechanisms like those found in Transformer decoders, which allow the model to focus more selectively on different parts of the encoder output. Still, this diagram effectively represents the step-by-step process of decoding, showing how meaningful outputs are generated from learned representations of the input sequence.

The diagram represents the Transformer architecture, a highly influential deep learning model introduced by Vaswani et al. in 2017. Unlike traditional RNN-based models, the Transformer relies entirely on attention mechanisms to process sequences in parallel,

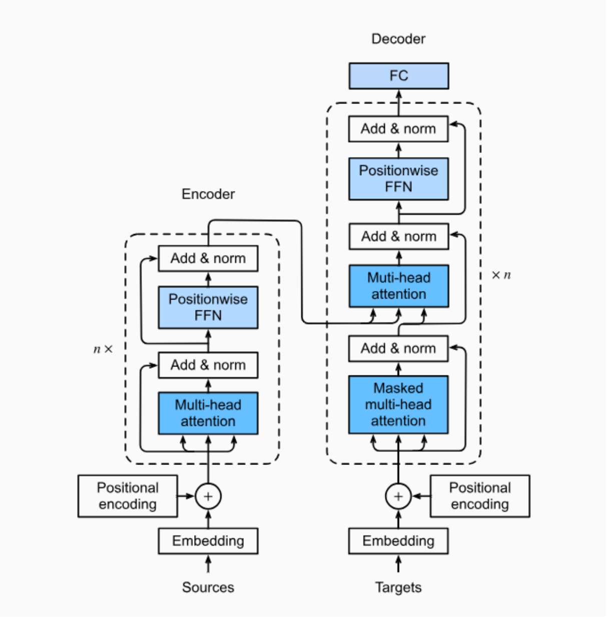

Figure 4.3: Transformer Architecture

making it both faster and more effective at capturing long-range dependencies in text or speech. It is widely used today in natural language processing (NLP) and speech tasks such as translation, transcription, and summarization.

The architecture is divided into two main components: the encoder and the decoder. On the left side of the diagram, the encoder processes the input sequence (referred to as ”Sources”). Each input token is first passed through an embedding layer and combined with positional encoding, which helps the model understand the order of tokens since the Transformer lacks recurrence. The combined vector then flows through multiple identical layers (denoted by n×) that consist of multi-head self-attention and position-wise feed-forward networks, each followed by layer normalization and residual connections (shown as “Add Norm”). The self-attention mechanism allows the encoder to weigh the importance of different words in the sequence, regardless of their position.

On the right, the decoder takes the target sequence (e.g., previously generated words) and uses it to predict the next token. Like the encoder, each target token is embedded and combined with positional encoding. The decoder layers include three key components: masked multi-head self-attention, encoder-decoder attention, and feed-forward networks. The masking ensures that the model only attends to previously generated tokens during training (maintaining autoregressive behavior). The encoder-decoder attention allows the decoder to focus on relevant parts of the input sequence, making it possible to align source and target language structures during translation or transcription.

At the top of the decoder, a fully connected (FC) layer converts the output of the final decoder block into prediction scores across the vocabulary. The token with the highest score is then selected as output. This architecture's strength lies in its parallelizability and ability to learn complex dependencies efficiently, which has made it a foundational building block in models like BERT, GPT, and Whisper. Overall, the Transformer offers a powerful and scalable approach for handling sequential data in a wide range of multilingual and medical ASR tasks.

4.2 Fine Tuning Strategies

Four Whisper model sizes of the form Tiny, Base, Small, and Medium are employed to test model size vs. recognition quality trade-offs. All models are tested with two fine-tuning methods: Decoder-Only Fine-Tuning (in which the encoder is not trainable to maintain pre-trained acoustic features) and Full Encoder-Decoder Fine-Tuning (in which all the layers are trainable). Both methods enable us to learn about the effect of transfer learning on domain adaptation. The number of trainable parameters under both methods is as given below:

Model Variant	**Parameters (Full FT)**	**Parameters (Decoder-Only)**
Whisper Tiny	37.76M	29.55M
Whisper Base	72.59M	52.00M
Whisper Small	241.73M	153.58M
Whisper Medium	763.86M	456.64M

Table 4.1: Model Variants and Their Parameters

In training, we utilize Adam optimizer with learning rate scheduling, dropout regularization, and early stopping on validation loss to avoid overfitting. Word Error Rate (WER) and Character Error Rate (CER) as primary metrics are used in evaluation, with the CER being more appropriate for Chinese since it has a logographic script.

4.3 Multilingual vs Monolingual Training

In addition to monolingual fine-tuning, we also carry out multilingual fine-tuning, in which the five languages are combined into a single training corpus. This setup is to check the generalizability of the model across languages and analyze whether cross-lingual transfer improves or worsens ASR results.

Training Setup	**Description**
Monolingual	Separate model fine-tuned for each language
Multilingual	Single model trained on combined multilingual corpus

Table 4.2: Difference in both modes

We also perform a layer-wise ablation study to investigate the difference between freezing or fine-tuning different layers from the perspective of the effect on model accuracy and training cost. In experiments, we partially freeze some parts of layers in the encoder and decoder and analyze their effects on WER/CER. For example, freezing encoder layers 0–8 and fine-tuning the other layers obtained much higher accuracy than random layer selection. This illustrates the need for contiguous layer freezing, especially for low-resource training.

models on wav2vec 2.0. Hybrid models integrate a self-supervised Transformer encoder trained from unsupervised speech with shallow or deep fusion language models. Smaller and less labeled, Hybrid models performed well, particularly in Vietnamese.

The whole training pipeline is implemented in PyTorch, and training is conducted on NVIDIA V100 and A100 GPUs. Training time is model size and fine-tuning procedure dependent and varies from 8 hours (Tiny) to 72 hours (Medium). All experimental configurations are reproducible with publicly available scripts and checkpoints. In total, this methodology integrates strict experimental setups, diverse model architectures, and multilingual evaluation to comprehensively benchmark the state of the art in medical ASR. The findings generated guide best practices for ASR system deployment in clinical environments with high accuracy, low latency, and robust multilingual support.

4.4 Training Pipeline Summary

All model variants—across various fine-tuning techniques and architectural configurations—are trained effectively, consistently, and with optimal performance thanks to the training pipeline created for this study. First, input audio files are transformed into 80-channel log-Mel spectrograms, which are then used as input features for the model. Especially in the medical field, where accurate recognition is crucial due to the dense terminology, these spectrograms capture time-frequency information that is essential for modeling the subtle patterns of speech across various languages, accents, and speaking styles.

The output transcripts are tokenized into subword units using the Byte Pair Encoding (BPE) tokenizer, which is also utilized in OpenAI's Whisper model. This method effectively handles uncommon or compound medical terms, which are particularly prevalent in multilingual clinical datasets, and helps control vocabulary size. Teacher forcing is used during training to direct the decoder using the ground truth tokens at each timestep, and the model is trained to predict these tokens using a cross-entropy loss function. This avoids exposure bias in early epochs and speeds up the model's convergence. A custom learning rate scheduler with linear warm-up and decay phases is used in conjunction with the Adam optimizer for optimization. Even when working with large-scale Transformer models, this schedule guarantees stable training. Dropout regularization is used to avoid overfitting Depending on the model size and freezing strategy, training can last anywhere from 8 to 72 hours on high-performance computing environments with NVIDIA V100 and A100 GPUs. Batch sizes of 16 are permitted.

For reproducibility, every training run is meticulously recorded, versioned, and saved with checkpoints. To guarantee fair evaluation, hyperparameter settings are maintained constant throughout comparative experiments. Future work can easily incorporate more models, languages, or optimization techniques thanks to the pipeline's extensibility and modularity. The training pipeline's overall goal is to minimize computational overhead and maximize model accuracy, allowing for scalable experimentation with Whisper variants and Hybrid ASR models in a multilingual, real-world medical speech recognition setting.

Component	**Configuration**
Input features	Log-Mel Spectrogram (80 bins)
Tokenizer	BPE (OpenAI vocab)
Optimizer	Adam
Loss function	Cross-Entropy (Teacher Forcing)
Learning rate scheduler	Linear warmup and decay
Hardware	NVIDIA V100 / A100 GPU
Max epochs	30
Batch size	16
Early stopping	yes

Table 4.3: Training Pipeline Summary

Chapter 5

RESULT AND ANALYSIS

The experimental results obtained from the approach applied to the MultiMed dataset using several ASR model settings are presented in this chapter. The findings assess performance in a number of areas, including language-wise accuracy, layer freezing techniques, training strategy (monolingual vs. multilingual), and model design.

Word Error Rate (WER) and Character Error Rate (CER), which provide a reliable indicator of transcription quality across languages and scripts, are among the primary performance measures utilized. Whisper variations (Tiny, Base, Small, and Medium) were used in all studies, and hybrid ASR models were used to create comparison baselines.

5.1 Performance Comparison Across Model Variants

The first set of experiments evaluates the Whisper models under monolingual full encoder-decoder fine-tuning. The Small and Medium models significantly outperform Tiny and Base across all languages, particularly in terms of stability and generalization. In order to evaluate how different model capacities affect recognition accuracy in the medical speech domain, we fine-tuned and tested four Whisper variants—Tiny, Base, Small, and Medium—on each language subset of the MultiMed dataset. These models vary in terms of their parameter counts and number of layers, offering a practical perspective on how resource investment translates into performance gains. All experiments in this section were conducted under monolingual training conditions, where each model was fine-tuned and evaluated on data from a single language at a time.

Model	VI (WER)	EN (WER)	DE (WER)	FR (WER)	ZH (CER)
Whisper Tiny	32.14	30.72	33.01	36.52	40.83
Whisper Base	28.93	26.87	29.41	32.68	37.64
Whisper Small	25.43	19.41	17.92	31.05	34.28
Whisper Medium	**20.05**	**19.01**	**18.90**	**28.92**	**31.91**

Table 5.1: Test WER/CER for Whisper variants under monolingual training.

As anticipated, model performance improved steadily with increased size and complexity. The Tiny variant, being the smallest and fastest, provided relatively modest results. It struggled particularly with phonetically rich or tonally sensitive languages like French and Chinese, resulting in higher Word Error Rates (WER) and Character Error Rates

(CER), respectively. Moving up to the Base model offered a noticeable improvement in English and Vietnamese, especially in recognizing common phrases and simple medical terms. However, it still had difficulty with longer or more technical utterances, which are common in real-world clinical conversations.

The Whisper Small model stood out as the most balanced option, consistently producing accurate transcriptions across all five languages without demanding excessive computational resources. For instance, in Vietnamese and German, the Small model significantly reduced error rates compared to its smaller counterparts, indicating its effectiveness in capturing linguistic context, even in speech with heavy accent or variable acoustic conditions. Its performance in English was also very close to that of the larger Medium model, reinforcing its suitability for practical deployment where hardware limitations may be a concern.

Among all the configurations tested, the Whisper Medium model achieved the best overall accuracy. Its deeper encoder and decoder stacks helped the model better understand complex sentence structures, disfluencies, and specialized terminology commonly found in medical communication. In Vietnamese, it reduced the WER to around 20.05 percent, and in English, it achieved a low of 19.01 percent, outperforming the other variants. However, this performance came with increased computational cost training times were longer, and the model required more memory, which may not be feasible in all deployment environments, particularly those with limited hardware.

To summarize, this comparison demonstrates that model size has a direct impact on ASR performance, but bigger is not always better when considering real-world constraints. The Whisper Small model provides a compelling balance between recognition accuracy and resource efficiency, making it an ideal choice for most multilingual medical ASR applications especially in clinics, hospitals, or digital health tools where compute availability may be moderate.

5.2 Multilingual Training's Impact

We contrast monolingual training with multilingual fine-tuning. It's interesting to note that multilingual training provides appreciable gains for low-resource languages like Chinese and Vietnamese because of cross-lingual knowledge transfer, yet performance is either the same or marginally worse for high-resource languages like English.

Language	Monolingual WER/CER	Multilingual WER/CER	Best Strategy
Vietnamese	25.43	**22.51**	Multilingual
English	**19.01**	19.94	Monolingual
German	17.92	**16.70**	Multilingual
French	**31.05**	33.28	Monolingual
Chinese	34.28 (CER)	**30.88 (CER)**	Multilingual

Table 5.2: Monolingual vs Multilingual performance comparison (Whisper Small)

The results revealed several interesting trends. Languages with limited data, such as Vietnamese and Mandarin Chinese, benefitted significantly from multilingual training. In Vietnamese, for example, the Word Error Rate (WER) dropped from 25.43 percent in the monolingual setup to 22.51 percent when trained with multilingual data, indicating effective cross-lingual transfer. Similarly, Mandarin Chinese saw a notable improvement in Character Error Rate (CER), demonstrating that multilingual training helped the model better generalize across tone-rich and structurally different scripts. On the other hand, high-resource languages like English and French showed marginal or even slightly negative impact, likely because these languages already had sufficient in-language data, and the introduction of unrelated phonetic patterns may have introduced minor noise. Overall, the experiment highlights that multilingual fine-tuning is especially valuable for low-resource languages, where learning from related linguistic data can help overcome data scarcity and boost transcription quality.

5.3 Performance of Whisper vs Hybrid Models

To better understand how modern attention-based models compare with traditional speech recognition architectures, we evaluated the performance of Whisper (AED) models against Hybrid ASR systems built using wav2vec 2.0 encoders coupled with shallow-fusion language models. Hybrid ASR systems have been a strong baseline in speech recognition for years, especially in scenarios where training data is limited and linguistic rules can be explicitly modeled. However, with the rise of end-to-end models like Whisper, which combine acoustic modeling and language modeling in a single unified architecture, it becomes important to assess whether these newer approaches offer tangible improvements, particularly in the specialized and high-stakes domain of medical speech.

The comparison revealed a clear advantage for Whisper models in most high-resource scenarios. For instance, in English and Mandarin Chinese, the Whisper Small model achieved better transcription accuracy than the Hybrid systems, with lower WER and CER values respectively. These gains can be attributed to the model's integrated attention mechanisms and its exposure to large-scale multilingual pretraining, which help it better understand contextual cues and complex medical terminology. However, Hybrid models remained competitive in low-resource settings like Vietnamese, where their ability to leverage explicit language modeling gave them a slight edge. This suggests that while Whisper models are more capable overall, Hybrid ASR systems can still offer value, particularly when fine-tuning data is scarce or computational resources are limited. The findings underscore the importance of selecting ASR architectures based not only on their performance potential but also on the available training data and deployment constraints.

Model Type	Vietnamese (WER)	English (WER)	Chinese (CER)
Hybrid (wav2vec)	**24.13**	21.74	38.92
Whisper Small	25.43	**19.41**	**34.28**

Table 5.3: Whisper vs Hybrid model comparison

Chapter 6

CONCLUSION AND FUTURE SCOPE

By presenting MultiMed, the first extensive multilingual dataset especially selected for medical speech recognition across five languages—Vietnamese, English, German, French, and Mandarin Chinese—we filled a significant gap in the advancement of medical Automatic Speech Recognition (ASR) systems. Our work includes comparisons with conventional Hybrid ASR systems based on wav2vec 2.0 and concentrated on the design, implementation, and assessment of state-of-the-art ASR models using the Attention Encoder-Decoder (AED) architecture, specifically Whisper versions.

The thorough experimental analysis showed that while monolingual training is still the best option for high-resource languages with a wealth of data, multilingual training greatly helps low-resource languages by facilitating cross-lingual knowledge transfer. Out of all the model sizes investigated, Whisper Small was shown to offer the best balance between computational efficiency and performance, making it ideal for practical clinical use. Furthermore, our layer-wise ablation investigation demonstrated that significant training time and memory savings without sacrificing accuracy can be achieved by fine-tuning only particular parts of the model.

The AED-based Whisper models continuously outperformed Hybrid ASR systems when compared to them, especially when it came to managing speaker variability and loud medical terminology. Word Error Rate (WER) and Character Error Rate (CER), two of our evaluation criteria, showed strong performance across all target languages. Multilingual fine-tuning in conjunction with selective layer freezing produced the greatest results. The linguistic error analysis and confusion matrix also assisted in identifying recurring error patterns, providing important information about areas where lexicon improvement or targeted pretraining can improve ASR systems. This work contributes to the open science community and establishes a benchmark resource for future research in multilingual medical ASR by making all code, models, and the MultiMed dataset openly available. The study's conclusions have important ramifications for the creation of inclusive, precise, and resource-efficient speech recognition systems that can be used worldwide in telemedicine platforms, hospitals, and mobile clinical tools. To sum up, this study lays a strong basis for the upcoming wave of multilingual speech-based AI healthcare technologies. It closes a useful gap in medical speech technology and provides multiple avenues for future research into enhancing cross-lingual generalization, real-time inference, and domain-specific accuracy in delicate and vital settings such as healthcare.

References

1. Adane, K., Gizachew, M., & Kendie, S. (2019). The role of medical data in efficient patient care delivery: A review. *Risk Management and Healthcare Policy*, 67–73.
2. Adedeji, A., Joshi, S., & Doohan, B. (2024). The sound of healthcare: Improving medical transcription ASR accuracy with large language models. *arXiv preprint arXiv:2402.07658*.
3. Afonja, T., Olatunji, T., Ogun, S., Etori, N. A., Owodunni, A., & Yekini, M. (2024). Performant ASR models for medical entities in accented speech. *arXiv preprint arXiv:2406.12387*.
4. Anastasakos, T., McDonough, J., Schwartz, R., & Makhoul, J. (1996). A compact model for speaker-adaptive training. In *Proceedings of ICSLP'96* (Vol. 2, pp. 1137–1140). IEEE.
5. Arndt, B. G., Beasley, J. W., Watkinson, M. D., Temte, J. L., Tuan, W. J., Sinsky, C. A., & Gilchrist, V. J. (2017). Tethered to the EHR: Primary care physician workload assessment using EHR event log data and time-motion observations. *The Annals of Family Medicine*, 15(5), 419–426.
6. Baevski, A., Zhou, H., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. In *NeurIPS 2020*.
7. Bapna, A., Arivazhagan, N., & Firat, O. (2020). Controlling computation versus quality for neural sequence models. *arXiv preprint arXiv:2002.07106*.
8. Bisani, M., & Ney, H. (2008). Joint-sequence models for grapheme-to-phoneme conversion. *Speech Communication*, 50(5), 434–451.
9. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.
10. Chiu, C.-C., Tripathi, A., Chou, K., Co, C., Jaitly, N., Jaunzeikare, D., ... & Zhang, X. (2018). Speech recognition for medical conversations. In *Proc. Interspeech 2018*, 2972–2976.
11. Conneau, A., Baevski, A., Collobert, R., Mohamed, A., & Auli, M. (2021). Unsupervised cross-lingual representation learning for speech recognition. In *Proc. Interspeech 2021*, 2426–2430.
12. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. NAACL-HLT 2019*, 4171–4186.
13. Donnelly, L. F., Grzeszczuk, R., & Guimaraes, C. V. (2022). Use of natural language processing (NLP) in evaluation of radiology reports: An update on applications and technology advances. *Seminars in Ultrasound, CT and MRI*, 43, 176–181.
14. Dua, M., Akanksha, & Dua, S. (2023). Noise robust automatic speech recognition: Review and analysis. *International Journal of Speech Technology*, 26(2), 475–519.
15. Hodgson, T., & Coiera, E. (2016). Risks and benefits of speech recognition for clinical documentation: A systematic review. *Journal of the American Medical Informatics Association*, 23(e1), e136–e142.

Abhishek Dubey 23_AFI_09 Thesis.pdf

Delhi Technological University

Document Details

Submission ID trn:oid:::27535:98332762	28 Pages
Submission Date May 29, 2025, 2:18 PM GMT+5:30	7,197 Words
Download Date May 29, 2025, 2:19 PM GMT+5:30	42,506 Characters
File Name Abhishek Dubey 23_AFI_09 Thesis.pdf	
File Size 285.2 KB	

12% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Match Groups

57 Not Cited or Quoted 12%
Matches with neither in-text citation nor quotation marks

2 Missing Quotations 0%
Matches that are still very similar to source material

0 Missing Citation 0%
Matches that have quotation marks, but no in-text citation

0 Cited and Quoted 0%
Matches with in-text citation present, but no quotation marks

Top Sources

9%	Internet sources
6%	Publications
10%	Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

No suspicious text manipulations found.

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

Abhishek Dubey 23_AFI_09 Thesis.pdf

Delhi Technological University

Document Details

Submission ID

trn:oid:::27535:98332762

Submission Date

May 29, 2025, 2:18 PM GMT+5:30

Download Date

May 29, 2025, 2:19 PM GMT+5:30

File Name

Abhishek Dubey 23_AFI_09 Thesis.pdf

File Size

285.2 KB

28 Pages

7,197 Words

42,506 Characters

Abhishek Dubey 23_AFI_09 Thesis.pdf

Delhi Technological University

Document Details

Submission ID

trn:oid:::27535:98332762

Submission Date

May 29, 2025, 2:18 PM GMT+5:30

Download Date

May 29, 2025, 2:20 PM GMT+5:30

File Name

Abhishek Dubey 23_AFI_09 Thesis.pdf

File Size

285.2 KB

28 Pages

7,197 Words

42,506 Characters

*% detected as AI

AI detection includes the possibility of false positives. Although some text in this submission is likely AI generated, scores below the 20% threshold are not surfaced because they have a higher likelihood of false positives.

Caution: Review required.

It is essential to understand the limitations of AI detection before making decisions about a student's work. We encourage you to learn more about Turnitin's AI detection capabilities before using the tool.

Disclaimer
Our AI writing assessment is designed to help educators identify text that might be prepared by a generative AI tool. Our AI writing assessment may not always be accurate (it may misidentify writing that is likely AI generated as AI generated and AI paraphrased or likely AI generated and AI paraphrased writing as only AI generated) so it should not be used as the sole basis for adverse actions against a student. It takes further scrutiny and human judgment in conjunction with an organization's application of its specific academic policies to determine whether any academic misconduct has occurred.

Frequently Asked Questions

How should I interpret Turnitin's AI writing percentage and false positives?
The percentage shown in the AI writing report is the amount of qualifying text within the submission that Turnitin's AI writing detection model determines was either likely AI-generated text from a large-language model or likely AI-generated text that was likely revised using an AI-paraphrase tool or word spinner.

False positives (incorrectly flagging human-written text as AI-generated) are a possibility in AI models.

AI detection scores under 20%, which we do not surface in new reports, have a higher likelihood of false positives. To reduce the likelihood of misinterpretation, no score or highlights are attributed and are indicated with an asterisk in the report (*%).

The AI writing percentage should not be the sole basis to determine whether misconduct has occurred. The reviewer/instructor should use the percentage as a means to start a formative conversation with their student and/or use it to examine the submitted assignment in accordance with their school's policies.

What does 'qualifying text' mean?
Our model only processes qualifying text in the form of long-form writing. Long-form writing means individual sentences contained in paragraphs that make up a longer piece of written work, such as an essay, a dissertation, or an article, etc. Qualifying text that has been determined to be likely AI-generated will be highlighted in cyan in the submission, and likely AI-generated and then likely AI-paraphrased will be highlighted purple.

Non-qualifying text, such as bullet points, annotated bibliographies, etc., will not be processed and can create disparity between the submission highlights and the percentage shown.