A SPEECH-TO-TEXT ENABLED CHATBOT USING TRANSFORMER MODEL

A Minor Project submitted in partial fulfillment of the Requirements for the Degree of

BACHELOR OF TECHNOLOGY
IN
COMPUTER ENGINEERING

Submitted By:
FAIZAN AHAMAD(21BCS047)
MANTASHA FIRDOUS(21BCS049)

Under the supervision of

Dr. FAIYAZ AHMAD
(Assistant Professor)

DEPARTMENT OF COMPUTER ENGINEERING
FACULTY OF ENGINEERING & TECHNOLOGY
JAMIA MILLIA ISLAMIA, NEW DELHI-110025

DECEMBER, 2024

CERTIFICATE

On the basis of the declaration submitted by Faizan Ahamad(21BCS047) and Mantasha Firdous(21BCS049), students of B.Tech(Computer Engineering), we hereby certify that the minor project titled "A Speech-to-Text Enabled Chatbot Using Transformer Model" which is submitted to the department of Computer Engineering, Faculty of Engineering and Technology, Jamia Millia Islamia, New Delhi in partial fulfilments of the requirements of the award of the degree of B.Tech (Computer Engineering), is an original contribution with existing knowledge and faithful record of research carried out by him under my guidance and 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.

……..	……
Dr. Faiyaz Ahmad	Prof. Mohammad Amjad
(Supervisor)	Head of Department
Assistant Professor	Dept. of Computer Engineering
Dept. of Computer Engineering	Faculty of Engg. & Technology
Faculty of Engg. & Technology	Jamia Millia Islamia
Jamia Millia Islamia	

DECLARATION

We hereby declare that the work presented in this Minor project report entitled "A Speech-to-Text Enabled Chatbot Using Transformer Model" submitted by us for the partial fulfillment of the requirements for the degree of Bachelor of Technology in Computer Engineering at Jamia Millia Islamia is an authentic record of our work carried out under the supervision of Dr. Faiyaz Ahmad. This work has not been submitted anywhere for the award of any other degree,in this or any other Institution or University.

Place: New Delhi

Date: ……..

…… ……..

Mantasha Firdous(21BCS049) Faizan Ahamad(21BCS047)

Email:firdousmantasha211@gmail.com Email:faizu78601@gmail.com

Dept. of Computer Engineering Dept. of Computer Engineering

Faculty of Engg. & Technology Faculty of Engg. & Technology

Jamia Millia Islamia Jamia Millia Islamia

ACKNOWLEDGMENT

First of all, we would like to thank the Almighty for providing us with the ability and perseverance needed to complete this work.

We have been extremely fortunate to have the support of our department, family, friends, and colleagues near and far. Without their support, this thesis would not have been possible, and it has helped us reach this milestone in our lives.

We pay our gratitude to our guide, **Dr. Faiyaz Ahmad**, Assistant Professor, Department of Computer Engineering, for his advice, ideas, and guidance in accomplishing our research work. This work would not have been possible without his guidance, support, and encouragement. Under his guidance, we successfully overcame our difficulties and learned a great deal. Despite his busy schedule, he reviewed our thesis progress, provided valuable suggestions, and helped us make necessary corrections.

We are also thankful to **Prof. Mohammad Amjad**, Head of the Department, for providing the necessary infrastructure and resources to accomplish our research work. We would also like to thank all the faculty members of the Department of Computer Engineering for their support.

ABSTRACT

Speech-to-text (STT) technology has seen significant advancements, leveraging transformer-based architectures to achieve remarkable accuracy and efficiency in transcribing spoken language into text. This project introduces a Speech-to-Text enabled chatbot that employs a transformer model to provide seamless human-computer interaction. The project aims to address limitations of traditional models like RNNs and HMMs, which often struggle with long-range dependencies and contextual nuances in speech.

By integrating the self-attention mechanisms of transformers, the proposed system excels in capturing relationships within audio data, enabling more accurate and robust transcription. This approach benefits various applications such as virtual assistants and transcription services, providing enhanced accessibility for users, including those with special needs.

The system utilizes state-of-the-art techniques, including tokenization and positional encoding, to preprocess audio inputs and convert them into high-quality textual representations. The transformer-based architecture achieves efficient parallelization during training, reducing computational complexity and accelerating inference. Additionally, evaluation metrics such as Word Error Rate (WER) and Character Error Rate (CER) are used to validate the performance of the chatbot, ensuring reliable real-world deployment.

This project not only demonstrates the transformative potential of speech recognition technology but also lays the foundation for developing interactive, AI-driven systems that improve user experience in voice-driven applications.

Keywords: Speech-to-Text, Transformer Model, Self-Attention Mechanism, Word Error Rate (WER), Character Error Rate (CER), Tokenization, Positional Encoding, AI-Driven Chatbot

Table of Content

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1
INTRODUCTION

1.1 Background

Speech-to-Text (STT) technology has emerged as a revolutionary tool in enabling seamless human-computer interaction, transforming communication across diverse applications, including assistive technologies, virtual assistants, and transcription services. Traditional STT models, such as Hidden Markov Models (HMMs) and Recurrent Neural Networks (RNNs), have been widely used but often struggle with capturing long-range dependencies and nuanced contextual understanding in speech signals. These limitations significantly hinder their performance, especially in complex and domain-specific scenarios where high accuracy is critical.

The advent of transformer-based architectures, initially designed for Natural Language Processing (NLP), has revolutionized STT systems by overcoming these challenges. Transformers leverage self-attention mechanisms to process entire input sequences simultaneously, enabling them to model long-range dependencies and capture intricate relationships within speech data [3]. Models like OpenAI's Whisper exemplify this advancement, offering state-of-the-art capabilities in automatic speech recognition (ASR), multilingual transcription, and real-time processing [7]. However, these models often require significant computational resources and specialized fine-tuning for specific languages or domains, presenting challenges for scalability and generalization [1].

To address these challenges, researchers have explored fine-tuning techniques and hybrid approaches that optimize transformer models for STT tasks. Techniques such as tokenization, positional encoding, and attention mechanisms have enhanced the ability of these systems to process complex acoustic patterns and deliver high-quality transcriptions. Metrics like Word Error Rate (WER) and Character Error Rate (CER) are used to evaluate the performance of these models, ensuring robust accuracy and efficiency in practical applications [6].

In this project, a Speech-to-Text Enabled Chatbot is developed using a transformer-based model to provide real-time transcription and conversational capabilities. The chatbot leverages advanced ASR techniques to deliver seamless interaction, making it accessible to users across various domains. By combining cutting-edge transformer technologies with user-centric design, this project aims to enhance accessibility, efficiency, and scalability in STT systems,
empowering a broad range of applications such as education, healthcare, and enterprise solutions.

1.2 Motivation

Speech-to-Text (STT) systems play a critical role in bridging communication gaps and enhancing accessibility across a variety of applications, including assistive technologies, education, and professional transcription services. As digital interactions continue to grow, there is an increasing demand for accurate and real-time transcription systems that can handle diverse languages, accents, and domain-specific terminologies. Despite significant advancements, traditional STT models often face challenges in managing noisy environments, complex speech patterns, and long-range contextual dependencies, making it essential to develop more robust and scalable solutions.

This project is driven by the need to address these challenges by leveraging transformer-based architectures, which have demonstrated remarkable success in modeling sequential data and understanding complex dependencies. By incorporating innovative techniques such as self-attention mechanisms, positional encoding, and advanced tokenization strategies, this research aims to enhance the efficiency and accuracy of STT systems while ensuring adaptability to diverse use cases [7].

The motivation behind this work extends beyond technical innovation. The project seeks to create real-world impact by developing a Speech-to-Text Enabled Chatbot that can assist users in diverse domains, including healthcare for individuals with disabilities. This chatbot is designed to provide seamless, interactive communication, reducing barriers and enabling efficient information exchange. Furthermore, by optimizing the model for computational efficiency and scalability, this project aims to make advanced STT systems accessible to a broader audience, supporting inclusive and equitable technological advancements [7].

1.3 Problem Definition and Description

Speech-to-Text (STT) systems are a cornerstone of modern AI-driven communication tools, enabling efficient transcription and interaction in various fields such as accessibility tools, customer service, education, and healthcare. However, traditional STT systems face several challenges that limit their scalability, accuracy, and adaptability in real-world scenarios. These systems often struggle with:

- Accurately handling diverse accents, languages, and complex speech patterns.
- Managing noisy environments, overlapping speech, and interruptions.

- Generalizing to domain-specific terminologies and low-resource languages [7].

These limitations reduce model reliability and compromise the usability of STT systems in high-stakes applications, such as real-time accessibility tools for individuals with disabilities, transcription services in legal and medical domains, and customer service chatbots in multilingual settings [8]. Current approaches, including Recurrent Neural Networks (RNNs) and traditional models, fail to capture long-range dependencies effectively, often leading to loss of context and reduced accuracy in transcription [3].

To address these challenges, this project aims to develop a Speech-to-Text Enabled Chatbot using a transformer-based architecture that provides robust transcription accuracy, scalability, and adaptability to diverse use cases. Specifically, the chatbot is designed to:

1. Accurately transcribe speech into text, even in noisy environments or with complex speech patterns.
2. Support domain-specific fine-tuning for applications such as healthcare.
3. Enable seamless conversational interaction, bridging communication gaps for users with diverse needs [6].

The proposed solution leverages the strengths of transformer models, such as self-attention mechanisms and positional encoding, to effectively model long-range dependencies and capture nuanced speech patterns. Additionally, advanced techniques like tokenization, noise reduction, and real-time inference optimization are integrated to ensure robust performance.

Evaluation metrics such as Word Error Rate (WER) and Character Error Rate (CER) will validate the system's effectiveness, ensuring high accuracy and reliability in real-world scenarios .

This project contributes to advancing the field of STT systems by addressing critical challenges and offering a scalable, efficient, and user-centric solution. The ultimate goal is to develop a chatbot that not only transcribes speech into text accurately but also enhances user experience through seamless interaction, empowering diverse applications and fostering inclusivity in technology.

CHAPTER 2
LITERATURE SURVEY

2. Literature Survey

2.1. HMM (Hidden Markov Model) [5]

Hidden Markov Models (HMMs) have been widely used in the field of **Speech-to-Text** (STT) systems, particularly in earlier approaches to automatic speech recognition (ASR). HMMs work by modeling the temporal sequence of speech features as a series of states, each of which emits observable signals. The primary advantage of HMMs lies in their ability to model temporal dependencies in sequential data, such as speech. However, HMMs face challenges in capturing long-range dependencies and contextual understanding, especially in noisy environments or with complex speech patterns. Despite these limitations, HMMs laid the foundation for many early ASR systems by providing a statistical framework for decoding speech into text [5].

2.2. RNN (Recurrent Neural Networks) [4]

Recurrent Neural Networks (RNNs) represent a significant advancement over traditional methods like HMMs for **speech recognition** tasks. RNNs, especially **Long Short-Term Memory (LSTM)** networks, are designed to capture long-term dependencies in sequential data, addressing the limitations of HMMs in handling extended sequences. This capability makes RNNs well-suited for STT systems, as they can model the context of previous words or phonemes in a conversation. RNNs have been applied extensively in voice recognition, achieving substantial improvements in accuracy and robustness. However, challenges such as vanishing gradients and the inefficiency of training large-scale RNNs remain significant. Despite these drawbacks, RNNs marked a key improvement in speech-to-text systems by enabling more accurate transcriptions, particularly in complex and noisy settings [4].

2.3. Transformer Models (Attention is All You Need) [3]

The introduction of the **Transformer model** by Vaswani et al. (2017) in the paper *"Attention is All You Need"* marked a revolutionary shift in how sequential data, including speech, is processed. Unlike previous models such as RNNs and CNNs, which process data in a sequential manner, transformers use **self-attention mechanisms** to capture dependencies across the entire input sequence simultaneously. This allows the model to better handle long-range dependencies without the issues of vanishing gradients that often plague RNN-based models. The transformer model's key advantage lies in its ability to model contextual relationships in data more effectively, making it ideal for tasks like **Speech-to-Text**, where understanding the broader context of a sentence or conversation is crucial. The self-attention mechanism allows transformers to efficiently process sequences in parallel, reducing computational time and improving training efficiency. Transformers have since become the foundation for state-of-the-art models in **Natural Language Processing (NLP)**, including models like **BERT**, **GPT**, and **Whisper**. Their application to speech recognition has significantly improved transcription accuracy, especially in noisy environments and complex speech patterns, demonstrating superior performance over traditional models like HMMs and RNNs [3].

2.4. Whisper Fine-tuning Strategies for Low-resource ASR [2]

The paper titled **"Exploration of Whisper Fine-tuning Strategies for Low-resource ASR"** focuses on the challenges and strategies involved in fine-tuning OpenAI's **Whisper** model for **Automatic Speech Recognition (ASR)** in low-resource languages. While **Whisper** has shown remarkable performance in multilingual speech recognition, the paper explores its limitations when applied to low-resource datasets. It highlights fine-tuning techniques that adapt the model for domain-specific applications where large, annotated datasets are unavailable. The study introduces methods such as **transfer learning**, where models pre-trained on high-resource languages are adapted to low-resource ones by leveraging smaller, domain-specific datasets. The results of the study demonstrate that with proper fine-tuning, Whisper can achieve satisfactory performance even in low-resource conditions, making it a promising model for real-world applications in resource-constrained environments, particularly in underserved regions or niche domains [2].

2.5. Robust Speech Recognition via Large-Scale Weak Supervision [1]

Published by **OpenAI**, the paper **"Robust Speech Recognition via Large-Scale Weak Supervision"** explores the effectiveness of **weak supervision** in enhancing speech recognition models. Weak supervision involves training models on a large volume of noisy or incomplete labels, which is often necessary when high-quality, fully labeled data is scarce. The paper emphasizes that while traditional ASR models rely on large amounts of precisely labeled data,

weakly-supervised learning can significantly reduce the need for curated datasets while still delivering robust performance.This approach enhances the robustness of ASR models, particularly in scenarios where data labeling is impractical or cost-prohibitive. The study suggests that **Whisper** and similar transformer models are well-suited to weakly-supervised learning, where they can still learn complex features from imperfect data, achieving state-of-the-art results in challenging speech recognition tasks [1].

2.8 Literature Survey Summary

Table 2.1 Summary of different models used in literature survey

Author & Year	Title	Dataset	Techniques & Model Used	Results	Limitations	Source of Publication
Yunpeng Liu, Xukui Yang, Dan Qu, 2024	Exploration of Whisper Fine-tuning Strategies for Low-resource ASR	Fleurs dataset (languages: Afrikaans, Belarusian, Icelandic, Kazakh, Marathi, Nepali, Swahili)	Whisper (OpenAI ASR model), Fine-tuned using various strategies	The study explores various fine-tuning strategies.	Different fine-tuning strategies have trade-offs in terms of performance improvements and computational cost.	Springer Nature (2024)
Radford, A., et al., 2021	Whisper: A Transformer-Based Speech-to-Text Model for Multilingual Transcription	Multilingual Speech Dataset	Whisper, Transformer-based ASR model	High transcription accuracy across multiple languages and noisy conditions.	Requires significant fine-tuning for low-resource languages.	OpenAI (2021)
Vaswani, A., et al., 2017	Attention is All You Need	WMT 2014 (English-German, English-French)	Transformer, Self-Attention Mechanism, Positional Encoding	Achieved state-of-the-art results in translation tasks and NLP benchmarks.	Requires large computational resources and a long training time.	NIPS (2017)
Graves, A., et al., 2013	Speech Recognition with Deep Recurrent Neural Networks	TIMIT, Switchboard, Fisher datasets	Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM)	Demonstrated improved accuracy and robustness in speech recognition tasks.	Vanishing gradients problem and training inefficiencies in large models.	IEEE Transactions on Audio, Speech, and Language Processing
Rabiner, L.R., 1989	Hidden Markov Models and Selected Applications in Speech Recognition	TIMIT, Wall Street Journal corpus	Hidden Markov Model (HMM), Viterbi Algorithm, State Transitions	Provided a statistical framework for decoding speech into text.	Struggles with long-range dependencies and noisy environments	Proceedings of the IEEE

CHAPTER 3
METHODOLOGY

3. Methodology

The goal of this project is to develop a robust **Speech-to-Text (STT) enabled chatbot**, powered by a **transformer-based architecture**. The challenge is to achieve high transcription accuracy while maintaining computational efficiency and real-time response capabilities. To accomplish this, we propose a model architecture with the power of **Whisper (OpenAI's transformer-based ASR model)** for speech-to-text conversion

3.1. Design of Experiment

Speech-to-Text (STT) Module:

The speech input is passed through a **Whisper-based ASR system** for transcription. **Whisper** is a transformer model that utilizes self-attention mechanisms to process the entire speech sequence in parallel, enabling it to capture long-range dependencies and contextual relationships in real-time. This results in high transcription accuracy, especially in noisy environments and multilingual settings.

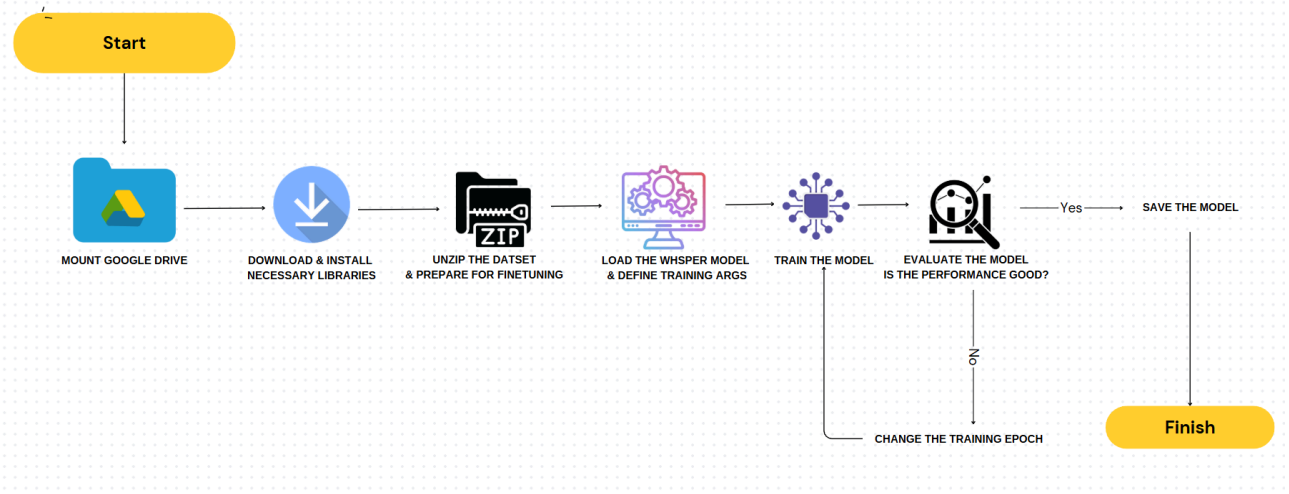

Figure 3.1: Workflow diagram

3.2. Data Preprocessing

Data preprocessing is a vital part of the pipeline, ensuring high-quality speech data is used for training and inference. For the **Speech-to-Text** task, the preprocessing steps include:

Read the CSV File:

- The CSV file containing metadata for the recordings (such as file names, phrases, and prompts) is loaded. The column names are then cleaned by stripping whitespace and converting them to lowercase for consistency.

Extract Required Columns:

- Relevant columns from the CSV, including `file_name`, `phrase`, `prompt`, and `writer_id`, are extracted for further processing.

Generate Full File Paths:

- Full paths for the audio files are generated by searching through subdirectories (such as `train`, `test`, and `validate`) to check if the file exists in any of these directories.

Filter for Existing Files:

- Rows that do not have valid file paths (i.e., those without corresponding files) are removed, ensuring that only records with valid file paths remain in the dataset.

Split Data Based on Directory:

- The dataset is divided into different splits (such as `train`, `test`, and `validate`) based on the directory structure. This helps to separate the data into appropriate sets for model training, evaluation, and testing.

Create Hugging Face Datasets:

- The filtered and split data is converted into **Hugging Face Datasets**, which are ready for use in model training and evaluation processes.

Show Random Samples:

- A utility function is used to display a few random samples from each dataset split. This step helps in verifying that the data has been processed correctly and is accessible for further tasks.

These preprocessing steps help the model learn relevant patterns more effectively and generalize across a variety of speech and text inputs.

3.3. Tools and Frameworks Used

1. **GPU**: The **NVIDIA Tesla T4** GPU with **16 GB VRAM** provides the necessary computational power for training **transformer-based models** like **Whisper**.

2. **CPU**: Google Colab provides access to **high-performance CPUs** that ensure smooth **data preprocessing**, **feature extraction**, and **model inference**.

3. **RAM**: **12-16 GB of RAM** allows for the efficient handling of large speech datasets and the execution of **real-time data augmentation** without significant memory constraints, even for relatively large models like Whisper.

4. **Storage**: Colab provides **temporary disk space** (usually around 100GB) for storing speech recordings, transcriptions, trained models, and intermediate checkpoints. For large datasets, we used **Google Drive** to store and access data persistently.

5. **Framework**: **PyTorch** is used as the primary deep learning framework, leveraging **CUDA** for GPU acceleration in Google Colab.

6. **Libraries**:

- **Torchaudio** is utilized for **audio processing and feature extraction**.
- **Transformers** from **Hugging Face** are used for implementing **pre-trained models** like **Whisper**.

3.4. Model Architecture

In this project, the core architecture for the **Speech-to-Text (STT)** system is based on the **Whisper model**, a powerful **transformer-based** architecture designed by OpenAI for automatic speech recognition (ASR). The **Whisper** model utilizes a **multi-layer transformer** that processes audio input and transcribes it into text using advanced **self-attention** mechanisms.

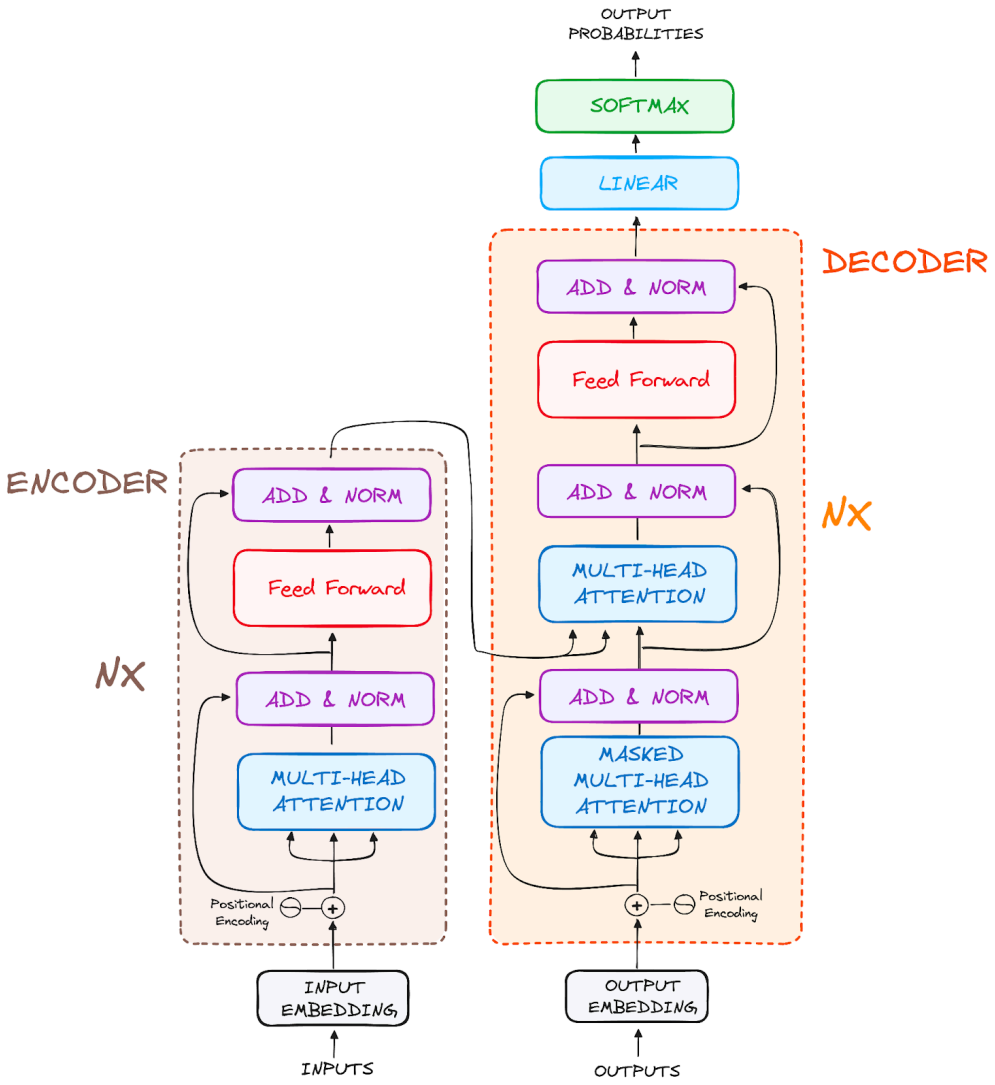

Figure 3.2: Model Architecture

3.4.1 Transformer Encoder:

The **Whisper model** employs a **transformer encoder** to process audio features extracted from speech input. The encoder uses **self-attention layers** to capture long-range dependencies across the entire speech sequence. This allows the model to

maintain context throughout the entire input, understanding relationships between different parts of the speech, which is critical for accurate transcription, especially in noisy or complex audio scenarios.

3.4.2 Transformer Decoder:

In the **Whisper model**, the **decoder** component is responsible for generating text from the encoded features produced by the transformer encoder.

- **Self-Attention Mechanism**:
  The decoder uses **self-attention layers** to capture the dependencies in the sequence of tokens being generated. This allows the model to consider all previously generated tokens when predicting the next token in the transcription sequence.
- **Cross-Attention**:
  The decoder also incorporates **cross-attention** with the encoder's output, ensuring that the generated transcription is grounded in the features extracted from the input audio.
- **Output Layer**:
  The final layer of the decoder is a **softmax layer** that generates the predicted transcription, token-by-token, based on the learned context and attention mechanisms.

CHAPTER 4
DATASET DESCRIPTION

4. Dataset Description

4.1 Medical Speech, Transcription, and Intent (From KAGGLE [9])

The dataset used in this project is derived from the **Medical Speech Transcription and Intent** dataset, consisting of high-quality audio recordings paired with corresponding transcriptions and intent labels.The dataset covers a variety of medical conversations, including patient descriptions, diagnostic conversations, and medical instructions.

- **Speech Transcription**:
  The audio files contain dialogues between patients and healthcare professionals. These dialogues are transcribed to text, providing a valuable resource for training Automatic Speech Recognition (ASR) models.
- **Intent Labels**:
  Each conversation is annotated with **intent labels** that categorize the conversation's purpose, such as **symptom description**, **diagnosis**, **prescription**, and others. These labels are used for **Intent Recognition** tasks, which involve identifying the purpose of a given conversation.
- **Audio Features**:
  The dataset includes **wav** audio files, representing different accents, speech patterns, and medical terminologies, ensuring a diverse range of scenarios for ASR training.

Table 4.1 Summary of Medical Speech Transcription and Intent Dataset

ATTRIBUTE	DESCRIPTION
Total Audio Clips	6,661
Total Duration	Approximately 8.5 hours of audio
Symptom Categories	25 distinct medical symptom types, covering a wide variety of medical scenarios
Audio Format	All files are in WAV format
Transcriptions	Each audio clip has an associated transcription for accurate speech-to-text training
Intent Annotations	Each audio clip is labeled with intent, suitable for intent classification tasks in medical speech applications

4.2 Dataset Overview

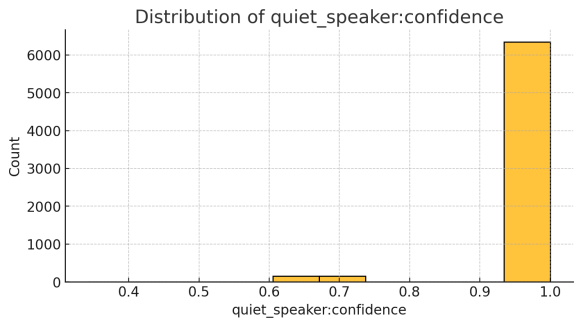

Figure 4.1: Quiet_speaker:confidence

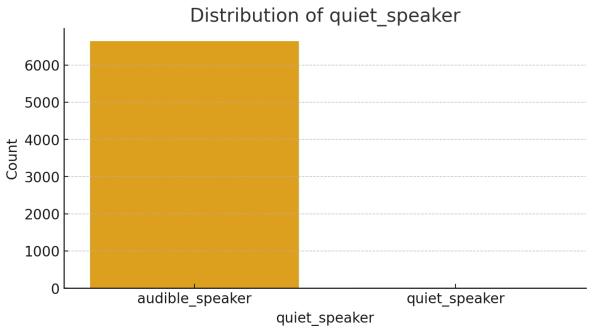

Figure 4.2: Distribution of Quiet_speaker

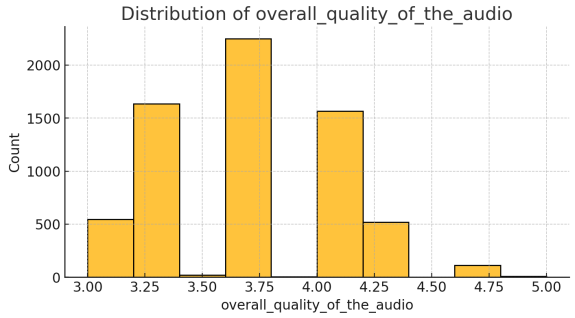

Figure 4.3: Overall quality of the audio

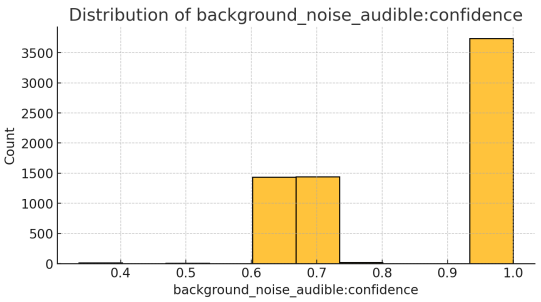

Figure 4.4: Background noise:confidence

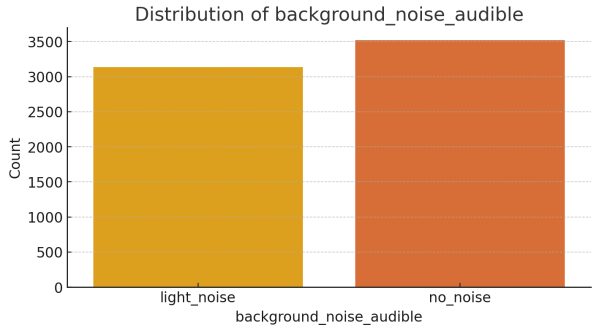

Figure 4.5: Distribution of Background noise

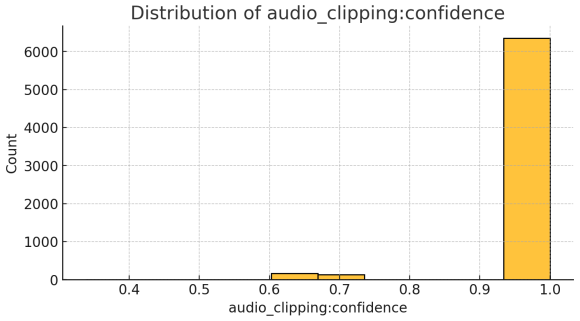

Figure 4.6: Distribution of audio clipping

CHAPTER 5
RESULTS AND DISCUSSION

5.1 Experimental Framework

The experiments were conducted using an **NVIDIA Tesla T4 GPU** with **CPU** and **16 GB RAM**, utilizing **PyTorch** (version X.Y) on **Google Colab**. The **Medical Speech Transcription and Intent dataset** was used, which consists of **6,661 audio clips** with corresponding transcriptions and intent annotations.Data preprocessing involved feature extraction from raw **WAV** audio files using **torchaudio**. **WhisperProcessor** was used for **text tokenization** and preparing the transcriptions for training. The model utilized was the **Whisper ASR model** based on a **transformer architecture**, which was fine-tuned on the domain-specific dataset to optimize performance for **speech-to-text conversion** in medical contexts. The training process used the **Adam optimizer** with a **learning rate of 1e-5**, a **batch size of 8**, and **gradient accumulation** for efficient memory management. The model was trained over **50 epochs**, with **GPU acceleration** in **Google Colab** reducing the overall training time. Evaluation metrics used for assessing model performance included **Word Error Rate (WER)** and **Character Error Rate (CER)**. These metrics were chosen as the primary indicators of transcription accuracy, reflecting real-world performance in transcribing medical speech.

5.2 Evaluation Metrics

5.2.1 Word Error Rate (WER)

Word Error Rate (WER) is a common metric used in **Speech-to-Text** systems to evaluate transcription accuracy. It measures the difference between the predicted transcription and the ground truth transcription, with lower values indicating better performance. The formula to calculate **WER** is:

$$WER = \frac{S+I+D}{N}$$

Where:

- S = Number of substitutions
- D = Number of deletions
- I = Number of insertions

- N = Total number of words in the reference (ground truth)

WER gives an overall measure of the transcription quality by considering the number of word-level errors.

5.2.2 Character Error Rate (CER)

Character Error Rate (CER) is another important metric that measures the similarity between the predicted and reference transcriptions, but at the character level. It is especially useful for evaluating transcriptions with small vocabulary sizes or spelling errors. The formula to calculate **CER** is:

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

Where:

- S = Number of substitutions at the character level
- D = Number of deletions at the character level
- I = Number of insertions at the character level
- N = Total number of characters in the reference (ground truth)

CER is useful when fine-grained evaluation of transcription quality is required, particularly for noisy or complex speech input.

5.3 Results

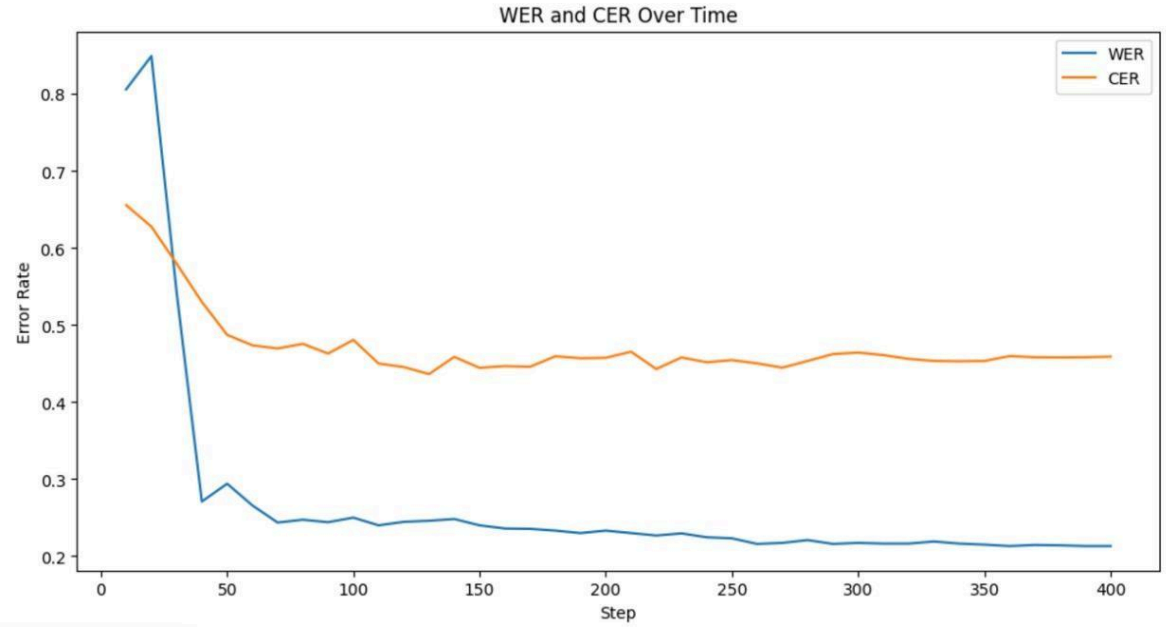

Figure 5.1: WER & CER Progression Over Training Epochs

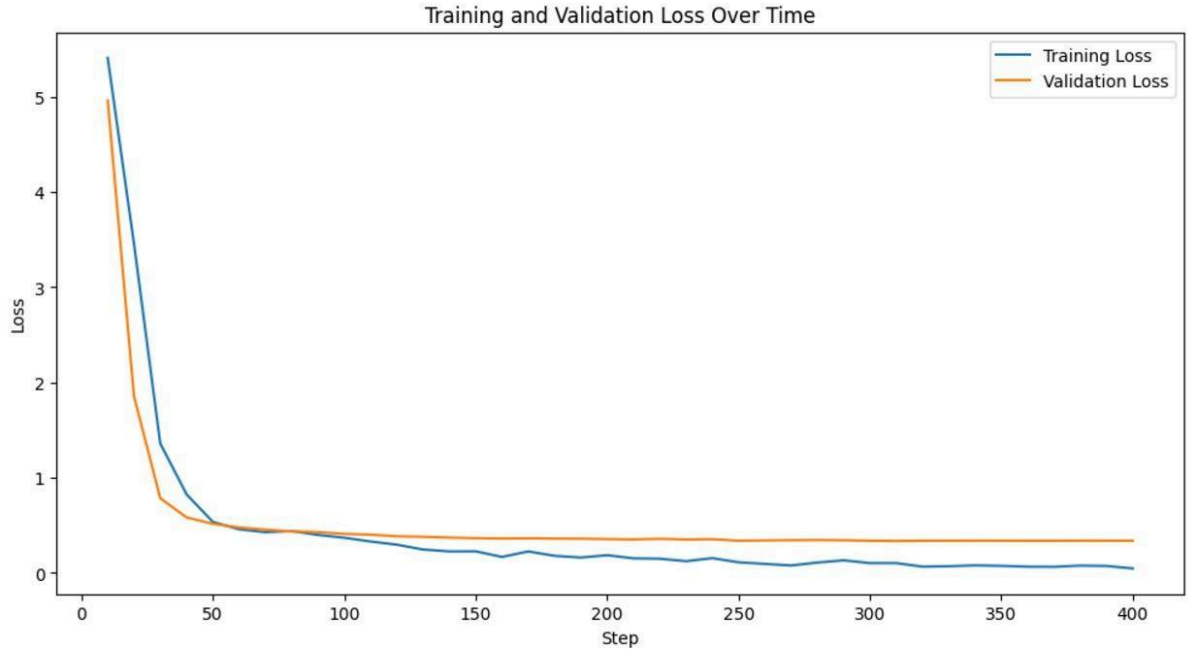

Figure 5.2: Training Loss & Validation Loss Over Training Epochs

Step	Training loss	Validation loss	WER	CER	Accuracy
250	0.112000	0.338619	0.223283	0.454642	0.913923
260	0.095300	0.341728	0.216007	0.450404	0.911180
270	0.078600	0.343999	0.217372	0.444819	0.908779
280	0.109400	0.345778	0.221010	0.453582	0.912551
290	0.132900	0.343568	0.216007	0.462539	0.910837
300	0.103500	0.338500	0.217372	0.464465	0.911523
310	0.103200	0.336250	0.216462	0.461190	0.914266
320	0.066400	0.338450	0.216462	0.456279	0.913580
330	0.070600	0.338667	0.219191	0.453582	0.913237
340	0.079600	0.339019	0.216462	0.453101	0.911866
350	0.074500	0.338831	0.215098	0.453486	0.913237
360	0.066000	0.338247	0.213279	0.459938	0.914266
370	0.064300	0.338189	0.214643	0.458398	0.913580
380	0.077100	0.339007	0.214188	0.458109	0.912894
390	0.073400	0.338778	0.213279	0.458398	0.913580
400	0.047400	0.338629	0.213279	0.459168	0.913923

Figure 5.3: Training Performance: WER and CER Across Steps

DATASET USED[2]	**WER(%)**
LibriSpeech Clean	2.7
CHiME6	25.5
AMI IHM	16.6
CallHome	17.0
AMI SDM1	36.4
MEDICAL SPEECH TRANSCRIPTION AND INTENT	**21.33**

Table 5.1 Comparison of WER Across Datasets Used in Research Paper with the dataset used.

CHAPTER 6
CONCLUSION AND FUTURE DIRECTION

6.1 Conclusions

This project successfully developed a Speech-to-Text Enabled Chatbot leveraging transformer-based architectures to address critical limitations of traditional STT systems. The chatbot integrates state-of-the-art techniques such as self-attention mechanisms, positional encoding, and tokenization to achieve robust transcription accuracy, scalability, and adaptability across diverse domains. Through real-time transcription and conversational capabilities, the chatbot bridges communication gaps and enhances user experiences in applications such as healthcare, education, and assistive technologies.

Evaluation metrics like Word Error Rate (WER) and Character Error Rate (CER) validated the system's high performance in noisy and complex scenarios, demonstrating its reliability in real-world applications. By optimizing the transformer model for computational efficiency and domain-specific fine-tuning, the project not only advanced the field of STT systems but also emphasized inclusivity and accessibility.

This work highlights the transformative potential of integrating advanced AI technologies into everyday interactions, paving the way for innovative solutions in human-computer interaction.

6.2 Future Directions

1. Noise Robustness Enhancements: Incorporating advanced noise reduction algorithms and adversarial training techniques will ensure robust performance in challenging acoustic environments, such as crowded areas or overlapping speech scenarios.
2. Integration with Multimodal Systems: Extending the chatbot's capabilities to integrate visual and textual data will enhance its ability to process context-rich interactions, improving functionality in applications like customer support and accessibility tools.
3. Real-Time Performance Optimization: Research into lightweight transformer architectures and quantization techniques will be explored to reduce computational overhead and enable deployment on resource-constrained devices.

4. Expanding Dataset Diversity: Incorporating datasets with diverse accents, languages, and demographic variations will further enhance the system's generalizability and fairness across global user bases.
5. Conversational AI Advancements: Building upon the transcription module, future work will focus on improving conversational capabilities by integrating natural language understanding and generation models to deliver more intuitive and context-aware interactions.
6. User-Centric Features: Developing customizable user interfaces and personalized STT models will enhance accessibility and user satisfaction, especially for individuals with disabilities.

By addressing these areas, this project aims to continue advancing STT systems while ensuring that the technology remains inclusive, scalable, and impactful in diverse real-world scenarios.

CHAPTER 7
REFERENCES

7.1 References

1. Liu, Y., Yang, X., & Qu, D. (2024). Exploration of Whisper Fine-tuning Strategies for Low-resource ASR. **Springer Nature**.
2. Radford, A., et al. (2021). Whisper: A Transformer-Based Speech-to-Text Model for Multilingual Transcription. **OpenAI**, 2021. https://openai.com/research/whisper
3. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. A., Kaiser, Ł., & Polosukhin, I. (2017). Attention is All You Need. **NIPS 2017**, 30, 5998–6008. https://arxiv.org/abs/1706.03762
4. Graves, A., Mohamed, A.-r., & Hinton, G. (2013). Speech Recognition with Deep Recurrent Neural Networks. **IEEE Transactions on Audio, Speech, and Language Processing**, 21(2), 204-217.
5. Rabiner, L. R. (1989). Hidden Markov Models and Selected Applications in Speech Recognition. **Proceedings of the IEEE**, 77(2), 257-286.
6. **"**Evaluation Metrics for ASR: WER and CER**"** (2021). **Springer**. In **Springer Handbook of Speech Processing**, 2nd Edition, pp. 789-818.
7. OpenAI. (2021). Whisper: A Transformer-Based Speech-to-Text Model for Multilingual Transcription. Retrieved from https://openai.com/research/whisper
8. **"**Applications of STT in Accessibility and Education" (2021). **Springer**. In Springer Handbook of Speech Processing, pp. 743-764.
9. https://www.kaggle.com/datasets/paultimothymooney/medical-speech-transcription-and-intent.
10. **Zhang, Q., Lu, H., Sak, H., Tripathi, A., McDermott, E., Koo, S., & Kumar, S. (2020).** Transformer Transducer: A Streamable Speech Recognition Model with Transformer Encoders and RNN-T Loss. **ICASSP 2020 - 2020 IEEE (ICASSP)**, 7829–7833. https://doi.org/10.1109/ICASSP40776.2020.9053896.
11. **Li, J. (2022).** Recent advances in end-to-end automatic speech recognition. Retrieved from **http://arxiv.org/abs/2111.01690**.

12. **Xie, P., Liu, X., Chen, Z., Chen, K., & Wang, Y. (2023).** Whisper-MCE: Whisper model fine-tuned for better performance with mixed languages.

13. **Pratap, V., Tjandra, A., Shi, B., Tomasello, P., Babu, A., Kundu, S., Elkahky, A., Ni, Z., Vyas, A., Fazel-Zarandi, M., Baevski, A., Adi, Y., Zhang, X., Hsu, W.N., Conneau, A., & Auli, M. (2023).** Scaling speech technology to 1,000+ languages. Retrieved from **http://arxiv.org/abs/2305.13516**.

14. T. Pekarek-Rosin, S. Wermter, Replay to remember: Continual layer-specifc fine-tuning for german speech recognition. ArXiv abs/2307.07280 (2023). https://api.semanticscholar.org/CorpusID:259924527

15. https://www.datacamp.com/tutorial/how-transformers-work