RSET
RAJAGIRI SCHOOL OF
ENGINEERING & TECHNOLOGY
(AUTONOMOUS)

Project Phase 2 Report On

Malayalam-to-English Real Time Speech Translation System

Submitted in partial fulfillment of the requirements for the award of the degree of

Bachelor of Technology

in

Computer Science and Engineering

By

Abijith Lohidakshan(U2003006)
Adhithyan R(U2003010)
Ajay A(U2003013)
Akhil Jose Francis(U2003017)

Under the guidance of
Mr.Sandy Joseph

Department of Computer Science and Engineering
Rajagiri School of Engineering & Technology (Autonomous)
(Parent University: APJ Abdul Kalam Technological University)
Rajagiri Valley, Kakkanad, Kochi, 682039
May 2024

CERTIFICATE

This is to certify that the project report entitled ***"Malayalam-to-English Real Time Speech Translation System"*** *is a bonafide record of the work done by* ***Abijith Lohidakshan(U2003006),Adhithyan R(U2003010),Ajay A(U2003013)*** *&* ***Akhil Jose Francis(U2003017)*** *submitted to the Rajagiri School of Engineering & Technology (RSET) (Autonomous) in partial fulfillment of the requirements for the award of the degree of Bachelor of Technology (B. Tech.) in Computer Science and Engineering during the academic year 2023-2024.*

Mr. Sandy Joseph
Project Guide
Asst.Professor
Dept. of CSE
RSET

Dr. Sminu Izudheen
Project Coordinator
Professor
Dept. of CSE
RSET

Dr. Preetha K. G.
Professor & Head of the Department
Dept. of CSE
RSET

ACKNOWLEDGMENT

We wish to express our sincere gratitude towards **Dr. P. S. Sreejith**, Principal of RSET, and **Dr. Preetha K. G.**, Head of the Department of **Computer Science and Engineering** for providing us with the opportunity to undertake this project, **Malayalam-to-English Real Time Speech Translation System**.

We are highly indebted to our project coordinator, **Dr. Sminu Izudheen**, Professor, Department of Computer Science and Engineering, for her valuable support.

It is indeed our pleasure and a moment of satisfaction to express our sincere gratitude to our project guide **Mr. Sandy Joseph** for his patience and all the priceless advice and wisdom he has shared with us.

Last but not the least, we would like to express our sincere gratitude towards all other teachers and friends for their continuous support and constructive ideas.

Abijith Lohidakshan
Adhithyan R
Ajay A
Akhil Jose Francis

Abstract

The goal of the proposed project is to develop a sophisticated real-time speech translation system specifically designed for Malayalam to English conversion. Sophisticated machine learning models based on Long Short-Term Memory (LSTM) that are deliberately deployed in three phases form its foundation. For the purpose of accurately transcribing the spoken Malayalam, the system first uses an LSTM model for speech-to-text conversion. The transcribed Malayalam text is then translated into English using a different LSTM model, which preserves linguistic subtleties and context. The process of translating a text from English to speech ends with another LSTM model converting the text back to speech.The system takes the malayalam speech input from the first user through its user interface. The input speech is then processed in the back-end that consist of the three modules. The converted English speech is given as the output in the second user's device.

This system is perfect for dynamic environments like live conversations and public speeches since it operates in real-time and has low latency. Its machine learning algorithms enable it to learn continuously from user interactions, improving the accuracy of its translations and adjusting to different speech patterns and accents.

The ability to translate instantly and accurately during real-time contacts is one of its greatest benefits; it provides smooth, organic communication channels free from annoying lags. This device facilitates easy communication between Malayalam and English speakers by translating speech inputs instantly, hence removing linguistic barriers.

Contents

List of Abbreviations

Abbreviation	**Expansion**
TTS	Text-to-Speech
LSTM	Long Short Term Memory
LPC	Linear Predictive Coding
MOS	Mean Opinion Score
BiLSTM	Bidirectional Long Short Term Memory
CBHG	Convolutional Bank Highway Gated Recurrent Unit
WER	Word Error Rate
MFCC	Mel-Frequency Cepstral Coefficient
RNN	Recurrent Neural Networks
NLP	Natural Language Processing
HMM	Hidden Markov Model
NMT	Neural Machine Translation
DNN	Deep Neural Networks
DBLSTM	Deep Bidirectional Long Short Term Memory
POS	Part Of Speech
RMSE	Root Mean Squared Error
LSD	Least Squares Distance
ASR	Automatic Speech Recognition
GAN	Generative Advarsarial Networks
MEMS	Micro Electromechanical Systems
API	Application Programming Interface
BLEU	Bilingual Evaluation Understudy

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

Effective language translation is crucial in our increasingly globalized society, and this is especially true given the variety of Indian languages. By using Long Short-Term Memory (LSTM) networks, this study seeks to answer the urgent demand for real-time voice conversion between Malayalam and English. The intricacies of informal speech are often beyond the capabilities of current translation methods, impeding effective communication across a range of industries. This research is important because it has the potential to improve educational access, promote cultural inclusion, boost economic growth, and demonstrate the state-of-the-art use of LSTM networks in solving practical language problems. The goal of this project is to overcome communication gaps and provide a reliable real-time speech conversion technology from Malayalam to English, paving the door for more effective and inclusive interactions in different contexts.

1.2 Problem Definition

The motivation behind this project lies in the imperative to overcome linguistic barriers between Malayalam and English speakers.Effective cross-cultural communication is hampered by existing translation systems inability to accurately translate spoken Malayalam. Motivated by a dedication to promoting cultural fusion and enhancing educational opportunities, the project seeks to enable more seamless contact amongst different fields. The creation of a precise and effective real-time speech conversion technology is crucial given the possible financial implications of commercial dealings and partnerships. Using LSTM networks is a novel strategy that demonstrates the project's commitment to pushing the frontiers of technology to solve linguistic difficulties in the real world. The end goal is to build an LSTM-based system that can translate spoken Malayalam into English dy-

namically while taking into consideration contextual variations, phonetic complexity, and regional accents. This would improve communication in a variety of industries.

1.3 Scope & Motivation

1.3.1 Scope

This project's scope includes creating a reliable real-time speech conversion system with LSTM networks specifically for translating Malayalam to English. The system is designed to accommodate a wide variety of linguistic circumstances, taking into consideration the many regional accents, phonetic nuances, and contextual differences that are present in Malayalam speech that is spoken informally. Furthermore, the scope encompasses the incorporation of the produced system into real-world applications, including technology, business, and education, to guarantee its flexibility and usefulness in authentic environments. With a focus on developing a flexible solution that goes beyond the present constraints of speech-to-speech translation, the project will investigate the possibility of scalability to support future developments in deep learning and natural language processing.

1.3.2 Motivation

- Cultural Bridging:

  Promote cultural affinities and overcome language barriers by assisting Malayalam and English speakers in communicating and understanding one another.

- Educational Accessibility:

  By giving people access to a technology that enables them to listen to and comprehend content in both Malayalam and English, you may improve language acquisition and advance accessibility and diversity in education.

- Practical Applications in Business and Tourism:

  By facilitating efficient communication between Malayalam and English speakers, you may help businesses and the tourism sector by improving user experience and customer service.

- Technological Advancement and Innovation:

  Demonstrate how machine learning and natural language processing can be used to create sophisticated language translation systems that will promote innovation in technology.

- Personalized Content Consumption:

  Enable consumers to access material in the language of their choice, offering a customized experience and enhancing user involvement across several platforms, such as media consumption and instant messaging.

1.4 Objectives

- Develop an intuitive and user-friendly React Native mobile app interface, allowing users to easily log in, connect with others, and initiate real-time speech translation with minimal effort.

- Create and train machine learning models for speech recognition, machine translation and speech synthesis implemented using advanced algorithms and techniques.

- Implement a reliable Django server backend that performs user authentication and integration of advanced speech recognition, machine translation, and speech synthesis models to ensure accurate and efficient translation in real-time.

- Allow users to communicate one-on-one with other users in real time, while also allowing self-translation on one's own device to promote interactions in person and language learning scenarios.

- Overcome language barriers and enhance cross-cultural connections by facilitating effective and seamless communication between Malayalam and English speakers.

1.5 Assumptions & Challenges

1.5.1 Assumptions

- **Clear audio with minimal noise :** The Malayalam audio input provided to the system must have minimum background noise as possible. Background noise

can interfere with the speech recognition system, making it difficult to accurately transcribe and translate spoken words.

- **Stable and reasonably fast internet connection :** The speech translation process is done as a cloud-based service. A stable internet connection ensures that the audio data is transmitted without significant delays. This enables smooth processing and translation of input Malayalam speech to text.

- **Reasonable speech rate and clear pronunciation :**Speech translation systems require a reasonable speech pace and clear pronunciation to function properly. Users must speak at a rate that allows for accurate real-time recognition. Clear pronunciation improves the system's capacity for understanding phonetic details decreasing the possibility of errors. Users are recommended to speak at a reasonable tempo and clearly to improve the overall performance and reliability of the speech translation system.

1.5.2 Challenges

The project has numerous obstacles, including difficulties in obtaining effective real-time speech conversion due to the innate structural disparities between the two languages syntax, grammar, and sentence structures. An further problem is the wide variety of accents and dialects seen among Malayalam speakers; in order to achieve comprehensive inclusivity, the system must be able to recognize and adjust to these variances. Furthermore, the system's ability to accurately translate and transcribe spoken words depends critically on maintaining high voice recognition accuracy, especially when dealing with background noise, different speaking tempos, and intonations.

1.6 Societal / Industrial Relevance

This project has a wide range of applications in both societal and industrial settings. It revolutionizes education in society by allowing pupils to access an abundance of global knowledge through real-time translations. Furthermore, it improves healthcare interactions by removing linguistic barriers between patients and healthcare practitioners. In the business world, the technology is critical in providing clear communication and developing global partnerships. It improves the user experience in customer service by offering

multilingual support. Furthermore, it is critical in emergency situations, providing efficient communication amongst people speaking different languages. Overall, the project encourages inclusivity, knowledge exchange, and efficient interaction across multiple sectors, benefiting both society and industry.

1.7 Organization of the Report

The report is organized into various sections, beginning with an introduction that discusses the project's scope, motivation, objectives, assumptions, and problems. The report's literature survey part includes an overview of existing voice recognition, translation, and synthesis systems, as well as related research. It addresses the many methodologies and techniques utilized in prior research, such as deep learning, neural networks, and natural language processing. The section also discusses the benefits and limits of various approaches, as well as prospective areas for further research. The project's hardware and software requirements are described, followed by the presentation of the system architecture, which includes the proposed architectural design, sequence diagram, module division, and module-wise diagram. The work breakdown and module schedule are then shown. Finally, the report concludes with a summary and the key findings and conclusions of the study with a references section, citing the sources used throughout the report

Chapter 2

Literature Survey

2.1 Existing System

1. **Automatic Speech Recognition (ASR):**
   - ASR is the technology that converts spoken language into written text. It plays a crucial role in speech translation systems by transcribing the spoken words in Malayalam into text format that can be further processed for translation.
2. **Machine Translation (MT):**
   - MT systems translate text from one language to another. For Malayalam to English translation, MT would be employed to convert the Malayalam text (resulting from ASR) into English. Traditional rule-based systems have been succeeded by statistical models and, more recently, neural machine translation (NMT).
3. **Neural Machine Translation (NMT):**
   - NMT is a type of machine translation that utilizes neural networks to improve translation accuracy. It has proven to be effective in capturing complex language patterns and dependencies, making it suitable for translating between languages like Malayalam and English.
4. **Speech-to-Text (STT) Services:**
   - STT services, provided by companies like Google and IBM, convert spoken words into written text. These services can be used as the initial step in processing spoken Malayalam, converting it into a format suitable for translation.
5. **Natural Language Processing (NLP):**

- NLP techniques are integrated into speech translation systems to enhance the understanding of context, idiomatic expressions, and cultural nuances. NLP helps in making the translations more contextually accurate and linguistically natural.

6. **Deep Learning Models:**

   - Deep learning models, such as recurrent neural networks (RNNs) and transformers, have been employed in speech translation systems to improve the learning and generalization capabilities of the models. Transformers, in particular, have shown great success in various natural language processing tasks.

7. **Speech Translation Apps and Platforms:**

   - Various applications and platforms offer speech translation services. Google Translate and Microsoft Translator are examples of platforms that provide speech translation capabilities. Users can speak or input text in Malayalam, and the platform translates it into English or vice versa.

2.2 Malayalam Speech to Text Conversion Using Deep Learning[1]

- The initial step of the method involves recording and storing words in .wav files to create a Malayalam dataset. This entails recording using Pyaudio, a Python package that provides built-in audio methods. Different speakers record multiple samples of each word, which are saved as distinct files. This phase guarantees a large and varied dataset for the speech recognition system's testing and training.

- Particular features are applied to signals in order to facilitate voice recognition. Mel frequency Cepstral Coefficients are the main method for feature extraction (MFCCs). When the feature extraction process is finished, MFCCs are obtained, and they analyze the frequencies with sensitivity comparable to human perception. This methodical approach is vital to capture the necessary features of the voice signals for precise recognition.

- To train the system, a long short-term memory network (LSTM) is used. Three equal portions of the database are picked, with two thirds going toward training

and the remaining third going for testing. This section enables thorough system performance validation and training. To make sure the system can generalize and recognize speech from a variety of sources, it is also tested using voice from speakers who are unknown to the user.

- The system utilizes Hidden Markov Model (HMM) and LSTM as classification techniques. These methods provide an elegant statistical framework for modeling speech patterns using a Markov process, allowing for effective speech classification and recognition. The combination of HMM and LSTM enhances the system's ability to accurately identify and transcribe spoken words.

- The software used for testing and training is PyCharm, an integrated development environment for Python programming. Additionally, TensorFlow, an open-source software library for machine learning, is utilized for training and inference of deep neural networks. These tools provide a robust and efficient environment for developing and testing the speech recognition system.

- The deep learning model's efficacy is evaluated through the utilization of accuracy and F1 score metrics. These metrics offer insightful information about how well the system works and how well it can translate Malayalam speech into text. The assessment procedure guarantees that the system satisfies the intended criteria for precision and dependability.

- The goal of the system is to develop it into a speaker-independent system that can manage a big vocabulary of continuous and linked words. This extension would greatly improve the system's usability and relevance in practical situations. The approach also seeks to highlight the potential societal influence of Malayalam by helping illiterate people write words in the language and encouraging everyday usage of the local tongue.

2.3 Real Time Translation of Malayalam Notice Boards to English Directions[2]

- This research examines how to translate Malayalam notice boards into English directions more accurately in real-time by utilising Neural Machine Translation (NMT) algorithms.

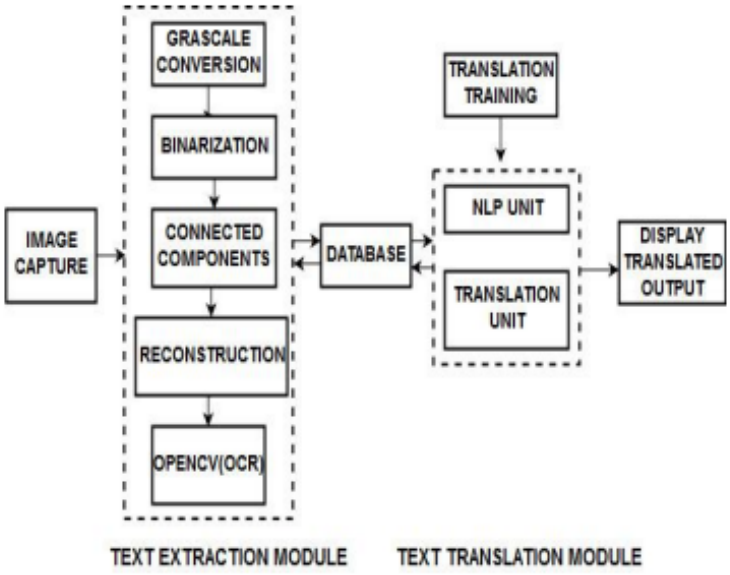


Figure 2.1: Block diagram of the system

- The self-attention mechanism and the bidirectional LSTM make up the two components of the suggested phrase embedding model.
- Using a large volume of data that has undergone some text preprocessing (text cleaning) can improve the translation's accuracy.

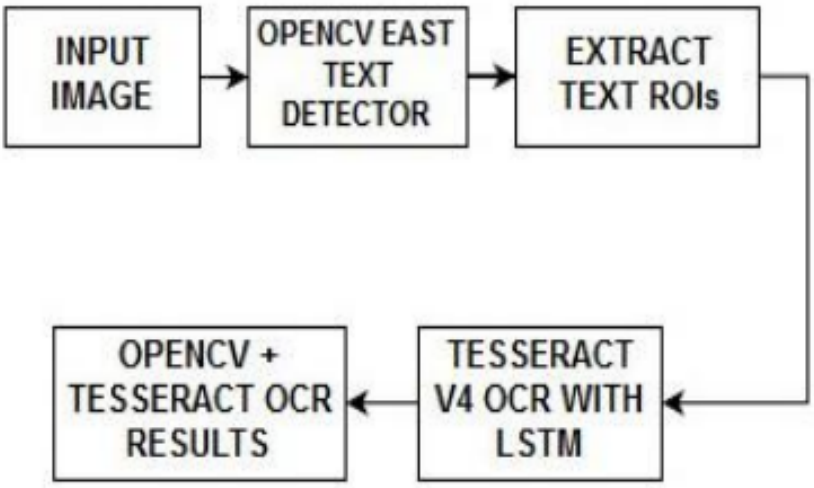


Figure 2.2: OpenCV OCR pipeline

- The accuracy of the translated sentence in English serves as a gauge for the quality of the translation, with a 75% accuracy rate attained.

- Because the dataset was short, the model that employed the character level embedding model performed marginally better than the word-based model, producing translated results that were more accurate, even when morphologically rich languages like Malayalam were used.

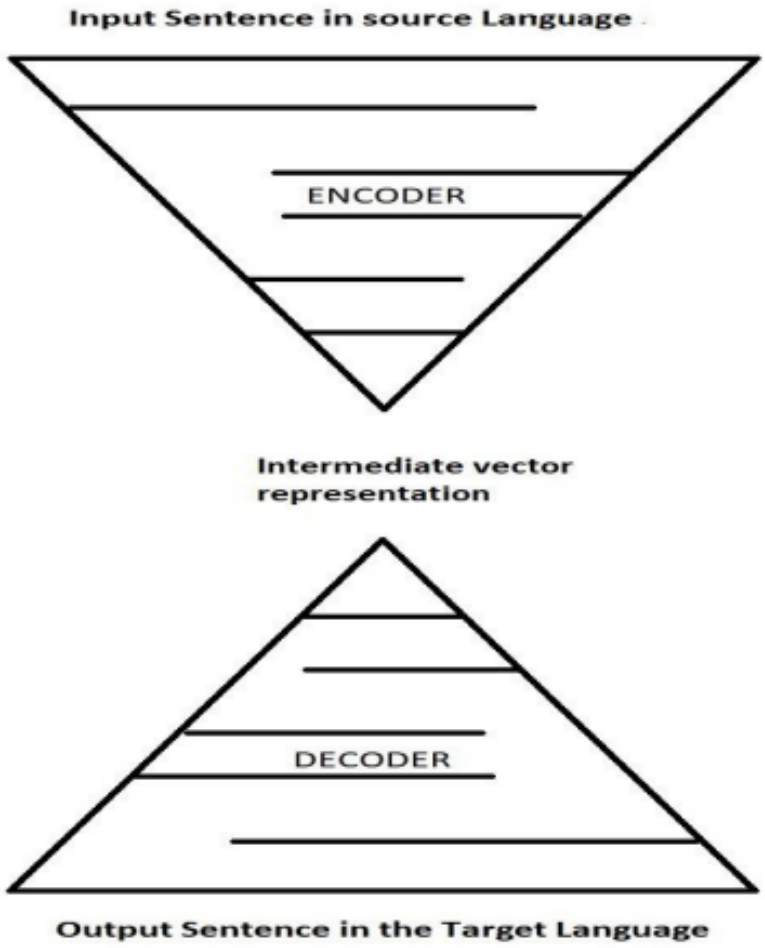


Figure 2.3: Translation Unit

- The authors propose that using NLP to process the dataset, training with cleaned data, and increasing the quantity of data the model is trained with can all help to increase system performance.

- The goal of the study is to improve translation accuracy by utilising natural language processing (NLP) and deep learning. It is intended to translate Malayalam written and printed on various media, in addition to notice boards.

- Future research, according to the authors, might include other modes, such as text-

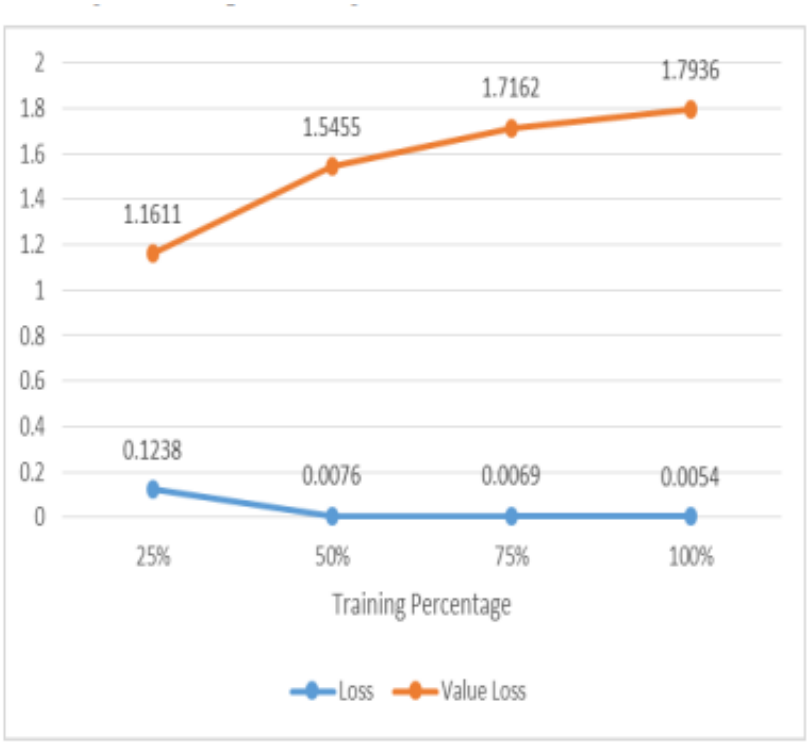


Figure 2.4: Graph showing the change in loss during each stage of training

voice, voice-text, and real-time voice-voice translation.

2.4 TTS Synthesis with Bidirectional LSTM based Recurrent Neural Networks[3]

- The paper presents a hybrid system of DNN and BLSTM-RNN for parametric TTS synthesis, compares its performance to HMM and DNN-based systems, and encourages future research into DBLSTM-RNN with a deeper structure and a larger corpus.
- The tests make use of a corpus of female American English native speakers.
- Speech signals are captured at 16 kHz and translated into static and dynamic LSPs.
- The phonetic and prosodic contexts include quin-phone, phone position, syllable and word in phrase and sentence, word and phrase length, syllable stress, TOBI, and POS of word.
- HMM-based TTS employs five-state, left-to-right HMM phone models, with each state represented by a single Gaussian output distribution with diagonal covari-

ance. The HMM is a statistical model that simulates the probability distribution of observations given a hidden state.

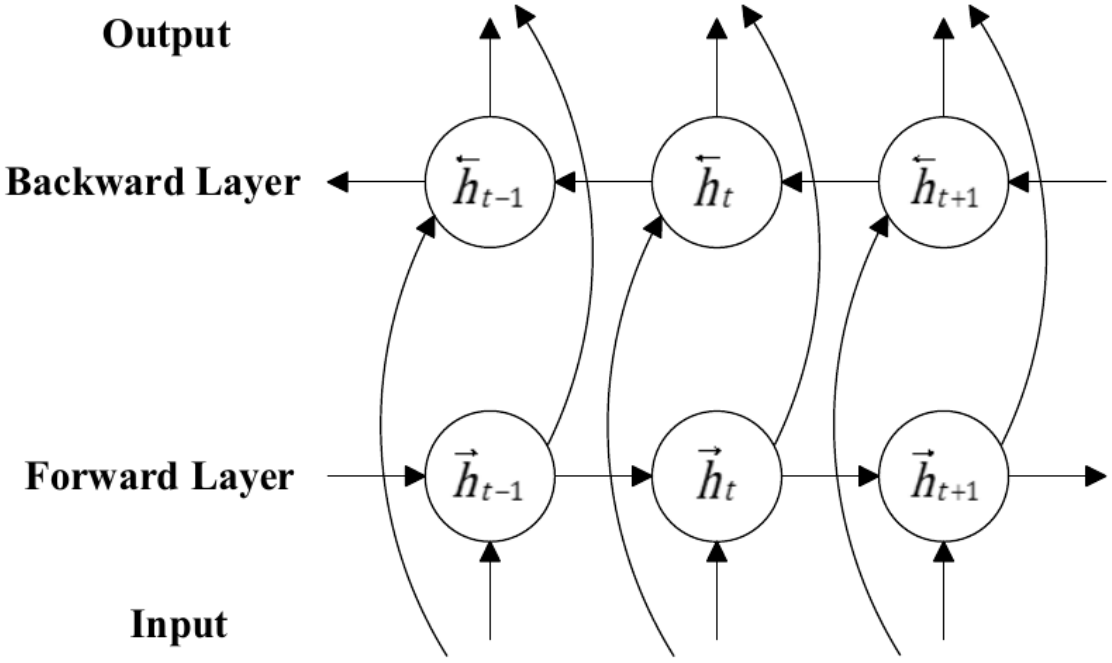


Figure 2.5: Bidirectional RNN Model

- DNN-based TTS employs a DNN with 6 hidden layers and 512 nodes per layer, as well as a DNN with 3 hidden layers and 1024 nodes per layer. DNN is a neural network composed of multiple layers of nodes, each of which is linked to all nodes in the previous layer.

- The same input and output features are used to train the BLSTM-RNN model as the DNN-based TTS. BLSTM-RNN is a type of recurrent neural network that captures the correlation or co-occurrence information between any two instants in a speech utterance using bidirectional LSTM cells.

- In the TTS synthesis, the vocoder waveform is used as the output feature. Acoustic parameters such as spectral envelope and dynamic features are used to construct the vocoder waveform.

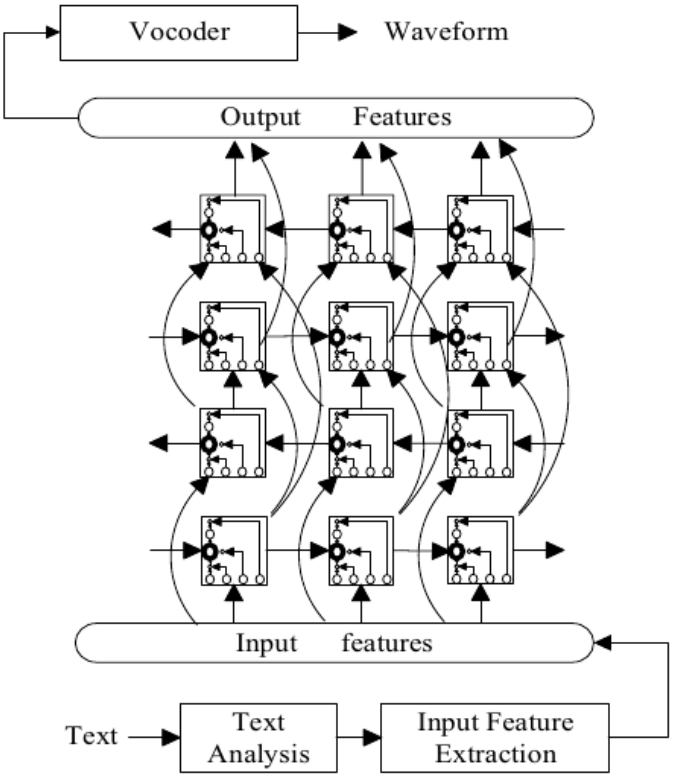


Figure 2.6: DBLSTM RNN Model

- F0 distortion in RMSE, V/U swapping errors, and normalized spectrum distance in LSD are used to objectively quantify synthesis quality.
- An AB preference test between speech sentence pairs synthesized by two different systems is used as the subjective assessment.
- The Hybrid system, which is a combination of DNN and BLSTM-RNN, outperforms the HMM and DNN systems as its preferred scores, 59% and 55%, are higher compared to HMM system with 22% and DNN system with 20%.
- The Hybrid system also has improved LSD of natural and generated spectra trajectories by over 0.1 dB.

Measures / Model	LSD (dB)	V/U Error rate	F0 RMSE (Hz)
HMM (2.89M)	3.74	5.8%	17.7
DNN_A (1. 55M)	3.73	5.8%	15.8
DNN_B (2.59 M)	3.73	5.9%	15.9
Hybrid_A (2.30M)	3.61	5.7%	16.4
Hybrid_B (3.61M)	3.54	5.6%	15.8

Figure 2.7: Results of LSD, V/U error rates and F0 RMSE for HMM, DNN and Hybrid systems

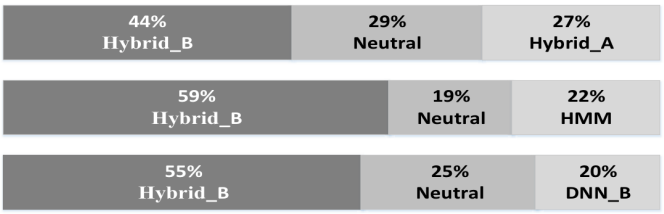


Figure 2.8: Preference scores of HMM, DNN and Hybrid systems

2.5 Development of Speaker-Independent ASR System for Kannada Language [4]

- The article presents a study on the development of an ASR system for the Kannada language, which has limited resources for continuous speech recognition.

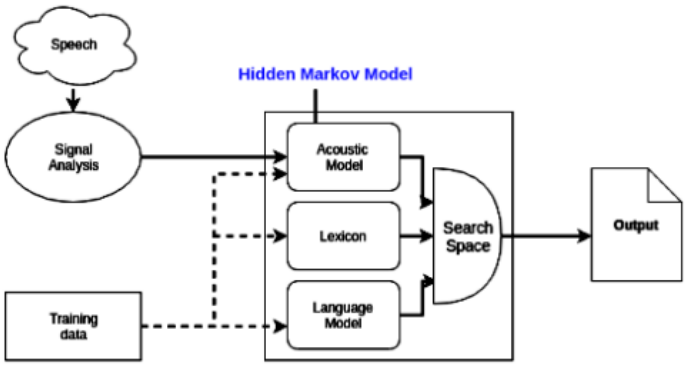


Figure 2.9: Working of HMM

- The authors suggest using modelling methods like triphone, monophone, SGMM, and hybrid modelling to raise the accuracy of Kannada speech data recognition.

- The study makes use of a database of 2800 speakers who were collected in real-world settings across the state of Karnataka. All speaker wave scripts were transcribed and validated.
- There are 49 phonemic symbols in the Kannada language, which are divided into three categories: vowels (Swaragalu), consonants (Vyanjanagalu), and semivowels (Yogavaahakagalu)..
- Based on the WER, the model's efficiency is calculated, and the outcomes are evaluated using popular datasets like TIMIT and Aurora-4.

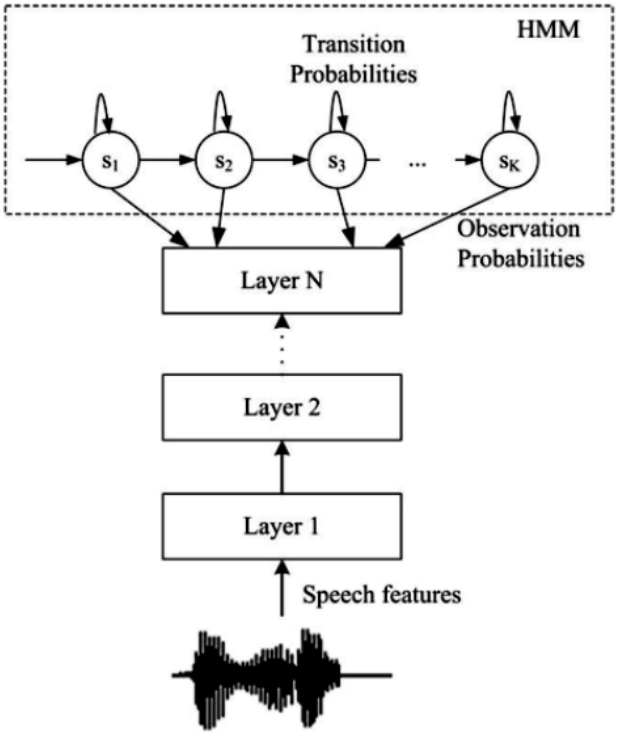


Figure 2.10: DNN-HMM Hybrid Device Mode

- The results of the experiment indicate that Kannada speech data has a higher recognition rate than the most advanced databases; the WERs for the monophone, triphone, DNN-HMM, and GMM-HMM acoustic models are 8.23 percent, 5.23 percent, 4.05 percent, and 4.64 percent, respectively.
- By tackling the problem of Kannada continuous speech recognition, the research advances the field of speech recognition and local language technology development.

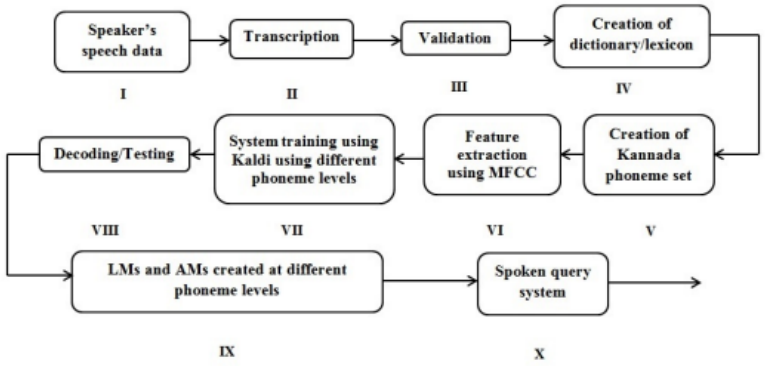


Figure 2.11: Schematic view of CSR for Kannada dialect

Phonemes	WER_1		WER_2		WER_3		WER_4		WER_5	
	CKSD	TIMIT	CKSD	TIMIT	CKSD	TIMIT	CKSD	TIMIT	CKSD	TIMIT
MONO	8.23	8.96	8.54	9.03	8.36	8.81	8.64	8.79	8.66	8.98
tri1_600_2400	7.52	7.86	7.48	7.86	7.26	7.59	7.81	8.06	7.65	8.15
tri1_600_4800	6.57	6.86	6.61	6.82	6.75	7.03	6.81	7.14	6.69	7.28
tri1_600_9600	6.24	6.54	6.37	6.85	6.48	6.94	6.22	6.76	6.35	6.86
tri2_600_2400	7.45	7.58	7.24	7.52	7.14	7.49	7.35	7.84	7.27	7.68
tri2_600_4800	6.52	6.98	6.27	6.94	6.29	6.84	6.35	6.85	6.33	6.73
tri2_600_9600	5.59	6.12	5.54	6.04	5.38	5.96	5.84	6.16	6.01	6.53
tri3_600_2400	5.79	6.25	5.61	6.09	5.54	6.02	5.58	6.12	5.88	6.24
tri3_600_4800	5.45	5.95	5.62	6.01	5.38	5.93	5.48	5.97	5.41	5.86
tri3_600_9600	5.12	5.62	5.34	5.96	5.27	5.81	5.23	5.88	5.32	5.59
SGMM	4.86	4.97	5.12	5.59	4.84	5.81	4.89	5.29	4.92	5.31

Figure 2.12: The represntation of WER at the different phoneme levels for the continous Kannada speech database

Phonemes	WER_1		WER_2		WER_3		WER_4		WER_5	
	CKSD	TIMIT	CKSD	TIMIT	CKSD	TIMIT	CKSD	TIMIT	CKSD	TIMIT
SGMM+MMI_it1	5.23	6.45	5.06	5.89	4.98	6.02	5.21	6.66	4.95	8.98
SGMM+MMI_it2	5.38	6.59	5.64	6.97	6.29	7.21	5.83	6.82	6.02	7.13
SGMM+MMI_it3	5.88	6.94	5.64	7.23	6.02	7.89	5.98	6.68	6.23	7.63
SGMM+MMI_it4	6.02	7.82	5.88	6.97	6.23	7.85	5.81	6.85	6.01	7.69
DNN+HMM	4.56	6.02	4.67	5.92	5.01	6.23	4.05	5.87	5.21	6.90
DNN+SGMM_it1	4.87	6.23	4.65	5.02	4.94	5.45	5.10	5.67	4.86	5.54
DNN+SGMM_it2	4.59	5.24	4.62	5.14	4.85	5.41	5.03	5.67	5.24	5.89
DNN+SGMM_it3	5.31	6.12	4.92	5.61	5.09	5.84	4.86	5.29	4.64	5.77
DNN+SGMM_it4	4.99	5.64	5.06	6.11	4.85	5.90	5.22	5.93	5.59	6.28

Figure 2.13: The depiction of WER for hybrid modelling techniques at different phoneme levels for continous Kannada speech database

Phonemes	WER_1		WER_2		WER_3		WER_4		WER_5	
	CKSD	Aurora-4	CKSD	Aurora-4	CKSD	Aurora-4	CKSD	Aurora-4	CKSD	Aurora-4
MONO	8.23	9.25	8.54	9.47	8.36	9.59	8.64	8.86	8.66	8.59
tri1_600_2400	7.52	7.86	7.48	8.45	7.26	8.21	7.81	8.58	7.65	8.68
tri1_600_4800	6.57	6.97	6.61	7.24	6.75	7.47	6.81	7.67	6.69	7.69
tri1_600_9600	6.24	6.54	6.37	7.35	6.48	7.81	6.22	7.29	6.35	7.03
tri2_600_2400	7.45	7.58	7.24	8.65	7.14	8.56	7.35	8.46	7.27	8.45
tri2_600_4800	6.52	6.98	6.27	7.25	6.29	7.64	6.35	7.61	6.33	7.97
tri2_600_9600	5.59	6.55	5.54	6.59	5.38	6.87	5.84	6.73	6.01	6.85
tri3_600_2400	5.79	6.25	5.61	6.58	5.54	6.69	5.58	6.67	5.88	6.95
tri3_600_4800	5.45	6.55	5.62	6.68	5.38	6.62	5.48	6.77	5.41	6.86
tri3_600_9600	5.12	6.62	5.34	6.84	5.27	6.85	5.23	6.23	5.32	6.97
SGMM	4.86	5.26	5.12	5.65	4.84	5.86	4.89	5.15	4.92	5.58

Figure 2.14: The represntation of WER at the different phoneme levels for the continous Kannada speech and Aurora-4 database

Phonemes	WER_1		WER_2		WER_3		WER_4		WER_5	
	CKSD	Aurora-4	CKSD	Aurora-4	CKSD	Aurora-4	CKSD	Aurora-4	CKSD	Aurora-4
SGMM+MMI_it1	5.23	7.02	5.06	6.89	4.98	6.91	5.21	7.23	4.95	7.21
SGMM+MMI_it2	5.38	6.65	5.64	7.35	6.29	6.86	5.83	6.54	6.02	6.58
SGMM+MMI_it3	5.88	7.28	5.64	7.86	6.02	7.61	5.98	7.54	6.23	6.98
SGMM+MMI_it4	6.02	8.01	5.88	7.61	6.23	8.28	5.81	7.81	6.01	8.03
DNN+HMM	4.56	5.68	4.67	5.27	5.01	5.81	4.05	6.21	5.21	5.97
DNN+SGMM_it1	4.87	5.94	4.65	5.68	4.94	5.94	5.10	6.01	4.86	6.24
DNN+SGMM_it2	4.59	5.54	4.62	5.64	4.85	5.29	5.03	6.31	5.24	6.10
DNN+SGMM_it3	5.31	6.54	4.92	5.58	5.09	5.64	4.86	5.68	4.64	5.93
DNN+SGMM_it4	4.99	6.14	5.06	5.59	4.85	6.21	5.22	5.92	5.59	6.34

Figure 2.15: The WER representation for hybrid modelling techniques for continous Kannada speech database and Aurora-4 database

2.6 Kannada to English Machine Translation Using LSTM [5]

- The methodology of this study builds a Seq2Seq model with an encoder-decoder mechanism. The RNN unit is an LSTM. The core objective of the project is to convert Kannada language into English text using NMT.

- A tokenizer that splits sentences into words, and assigns an integer code to each different word in the process, that begins with the input language translation into the destination language. Integer values are then transformed into vector values, which stand in for the input to the LSTM cell.

- In order to obtain an encoder vector, the LSTMs are arranged in series in the encoder unit, which is where the process starts. The value of internal states from the encoder's most recent LSTM, which contains details about earlier input elements, is the encoder vector. This vector serves as the first decoder LSTM's starting state value, aiding in the decoder's ability to make precise predictions. Both the training and inference processes use the same encoder technique.

- One word at a time will be produced throughout the inference procedure. As a result, only a single step is processed each time the LSTM decoder is called in a loop. The anticipated output at one time step serves as the input for the subsequent time step since the decoder states are saved and set as beginning states at each time step. The SoftMax activation function receives the anticipated output from the decoder and uses the vector values to choose which one to activate.

- The dataset utilized in this study is parallel-formatted Kannada-English Seq-Seq model based. The dataset is pre-processed in order to increase the accuracy of the

models' construction. With 41,000 data pairs and 40 epochs in the dataset, the study yielded an accuracy of 86.32% and an overall loss of 0.849.

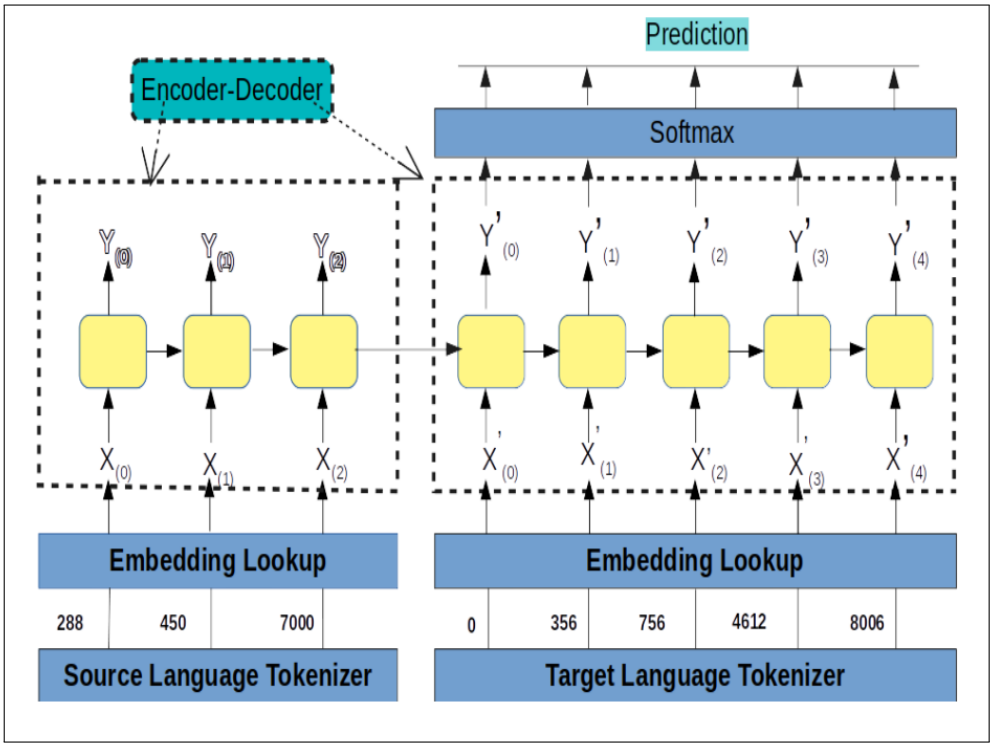


Figure 2.16: Proposed Architecture

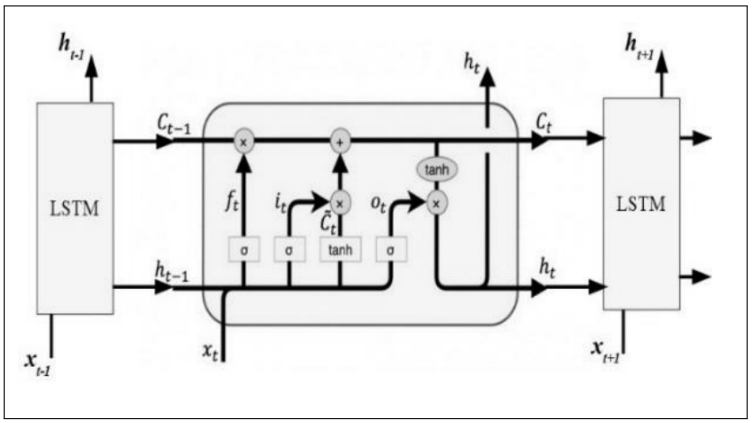


Figure 2.17: LSTM Model

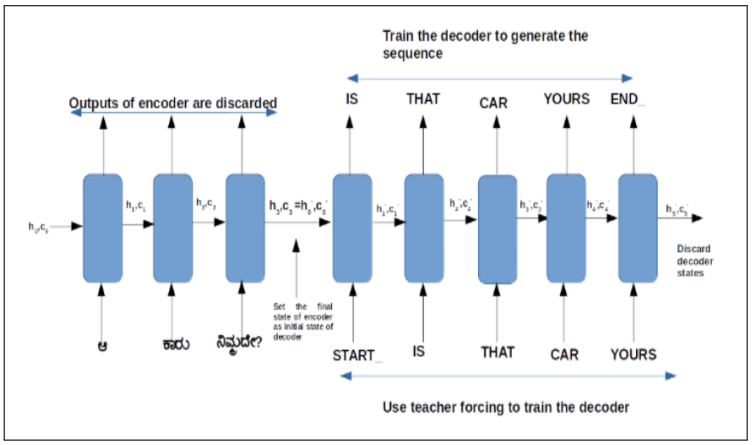


Figure 2.18: Encoder-decoder LSTM in training process

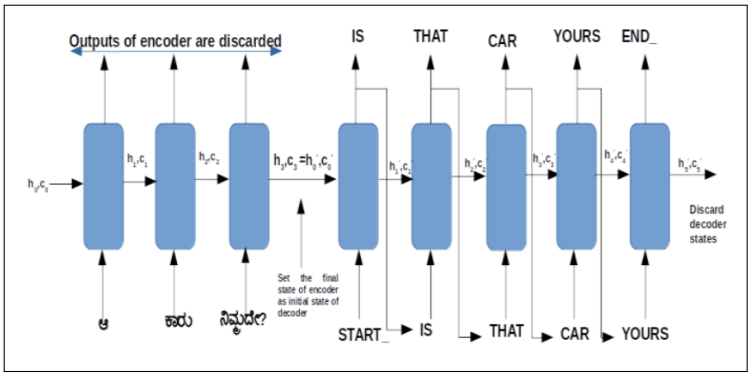


Figure 2.19: Encoder-decoder LSTM in inference process

BLEU (Weights)	Train (Scores normalized to 1)	Test (Scores normalized to 1)
BLEU-1 (1.0, 0, 0, 0)	0.642263	0.472143
BLEU-2 (0.5, 0.5, 0, 0)	0.559133	0.360877
BLEU-3 (0.3, 0.3, 0.3, 0)	0.497395	0.302902
BLEU-4 (0.25,0.25,0.25,0.25)	0.317888	0.173815

Figure 2.20: BLEU Scores Obtained in Inference

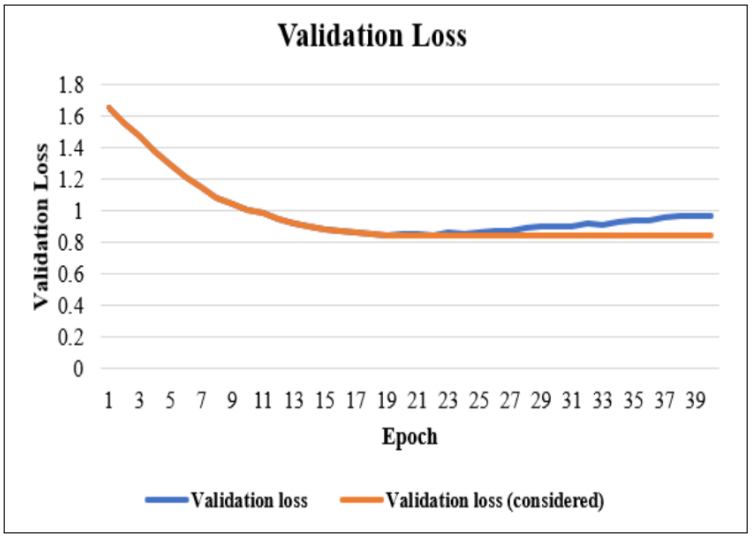


Figure 2.21: Validation Loss v/s Epochs Graph

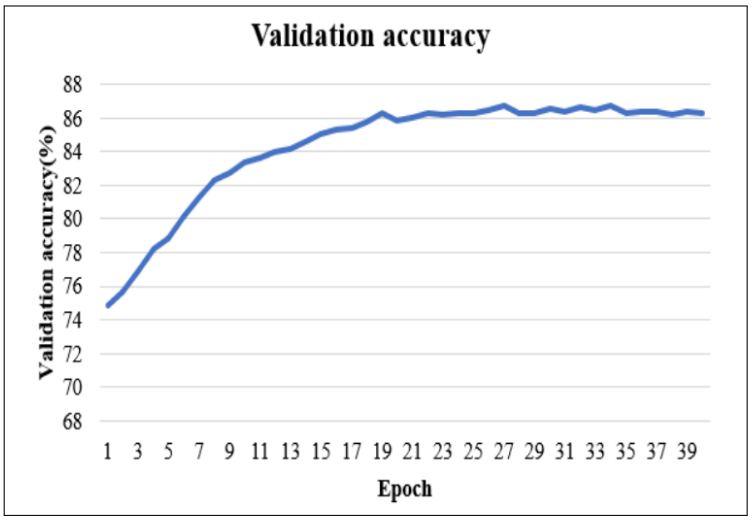


Figure 2.22: Validation Accuracy v/s Epochs Graph

2.7 Tigrigna Linguistic Text to Speech Synthesizer utilizing Concatenative Based Method and LSTM Model[6]

- The study collected a corpus of Tigrigna text data from various sources, including books, newspapers, and online resources.

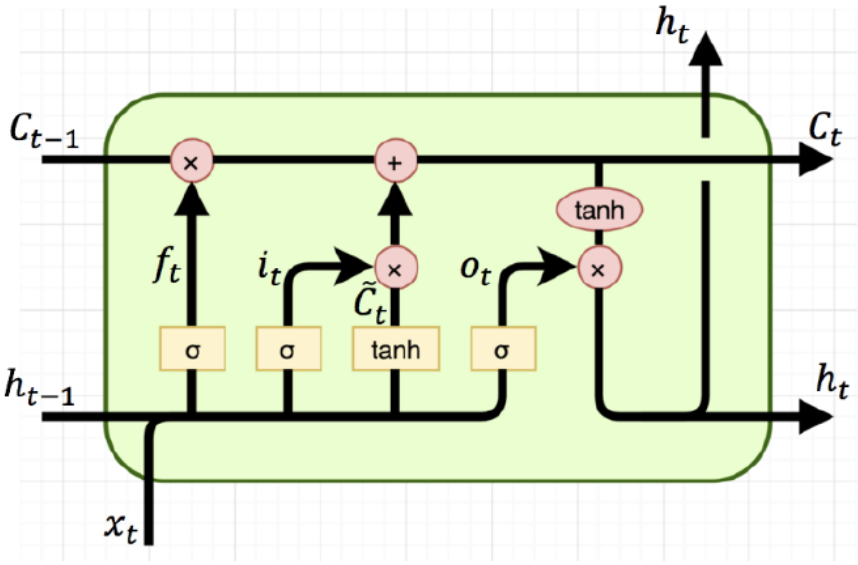


Figure 2.23: LSTM model

- The collected data was preprocessed to remove noise, normalize the text, and segment it into smaller units, such as words and syllables.
- Pitch and duration data, MFCCs, and LPC were among the features that the study identified from the preprocessed data.
- Concatenating pre-recorded speech fragments to generate words and sentences is the method of concatenative speech synthesis used in this study.
- In the study, the LSTM model was used to improve the quality and naturalness of the synthesized speech.
- In order to improve the prosodic consistency and naturalness of the synthesized speech, the study used a dialog fusion architecture.

- The study used a number of metrics, such as accuracy, precision, recall, F-score, and MOS, to assess the proposed system's performance.

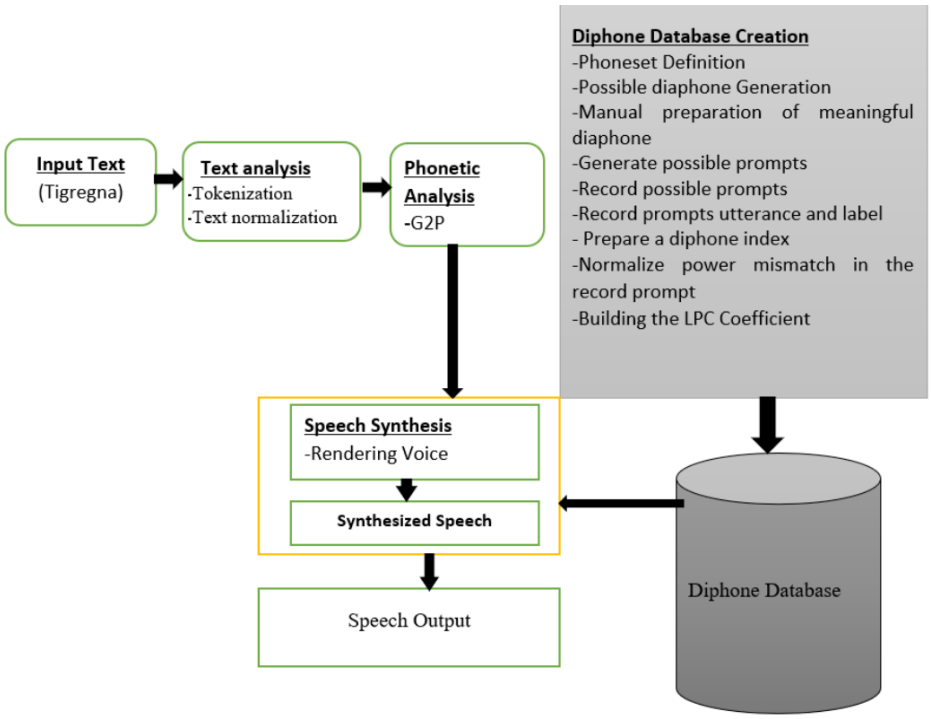


Figure 2.24: Proposed work architecture

- The study used a number of technologies, such as MATLAB, LPC, and Python, to put the suggested methodology into practice and assess the system's performance..
- The study detailed the system's overall performance in terms of word-level correctness, sentence-level naturalness and intelligibility, as well as its shortcomings and potential research areas.

Sentence	Intelligibility	Naturalness
1	3.25	3.15
2	3.15	3.25
3	3.30	3.50
4	3.40	3.35
5	3.40	4.25
6	3.00	3.20
7	3.15	3.30
8	3.35	3.05
9	3.25	3.40
10	3.40	3.30

Figure 2.25: Average MOS scores

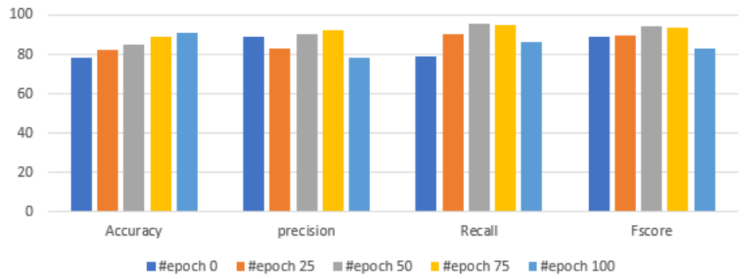


Figure 2.26: Performance of LSTM in different Epoch number

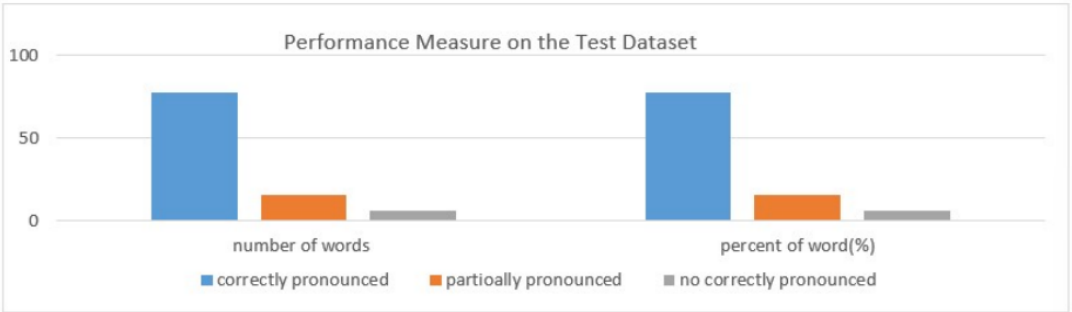


Figure 2.27: Performance Measure on the Test Dataset

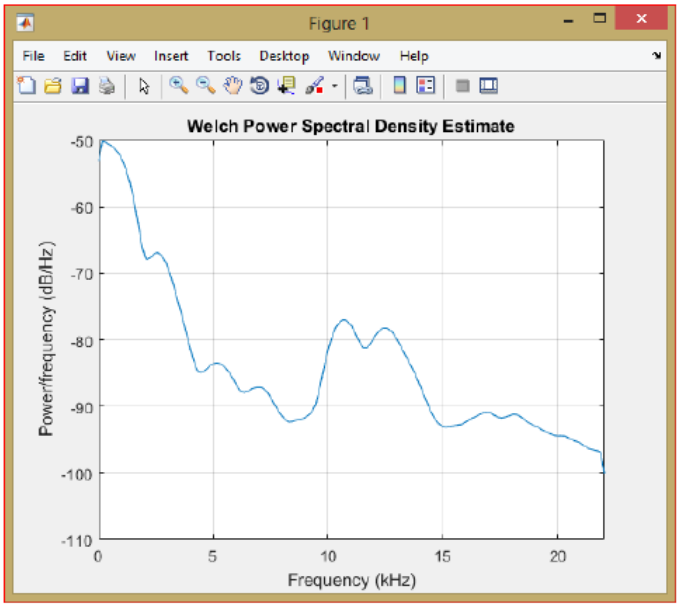
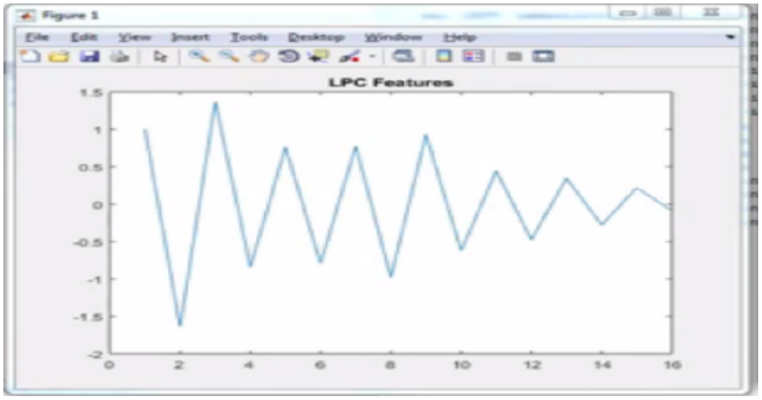


Figure 2.28: Formants of the word "arba"



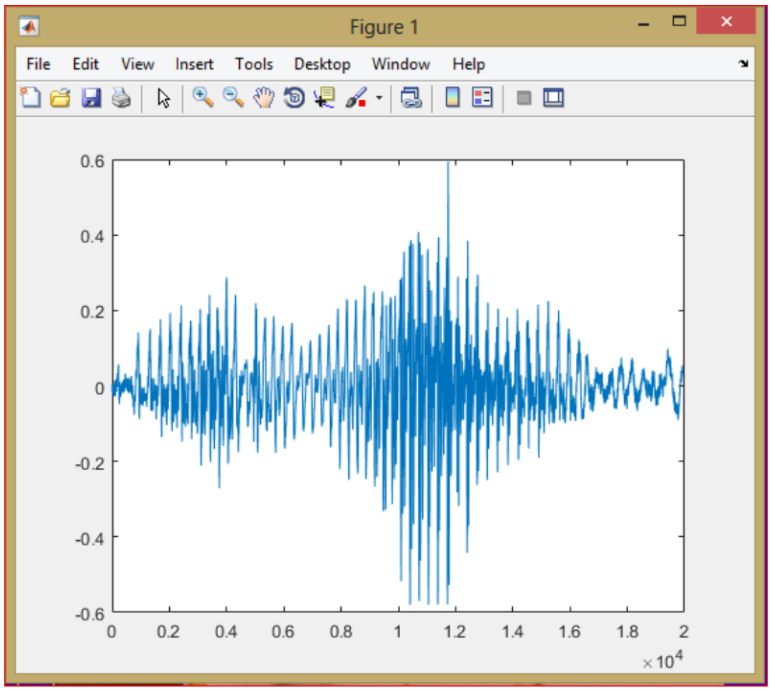


Figure 2.29: The word "arba" 's original signal

2.8 Text to speech system using Generative Adversarial networks [7]

- The conversion of text to speech involves the utilization of generative adversarial networks (GANs), MelGAN, for high-quality audio synthesis from text input
- The input text, consisting of two- or three-letter Hindi words, is preprocessed to prepare it for input into the neural network model. This may involve linguistic analysis and formatting to ensure compatibility with the model.
- The preprocessed text is then fed into the trained MelGAN model, which are designed to generate raw audio data from the input text. These models have been specifically tailored for audio synthesis and is able to produce high quality audio samples.
- The MelGAN model then synthesizes the audio waveform based on the input text, generating natural-sounding speech corresponding to the provided Hindi words.
- A feedforward generator is trained using a modified GAN architecture to produce raw speech audio from text input. Utilizing an ensemble of discriminators that operate on random windows of different sizes to analyze the audio in terms of its realism and correspondence to the utterance. Employing Mel-spectrogram inversions as input to the GAN set-up of discriminator and generator, creating a fully convolutional, non-autoregressive model with significantly fewer parameters.
- The system's effectiveness is evaluated through subjective human evaluation using MOS, a scale used to measure the quality of speech synthesis and quantitative metrics such as Frechet deep speech distance and Kernel deep speech distance.

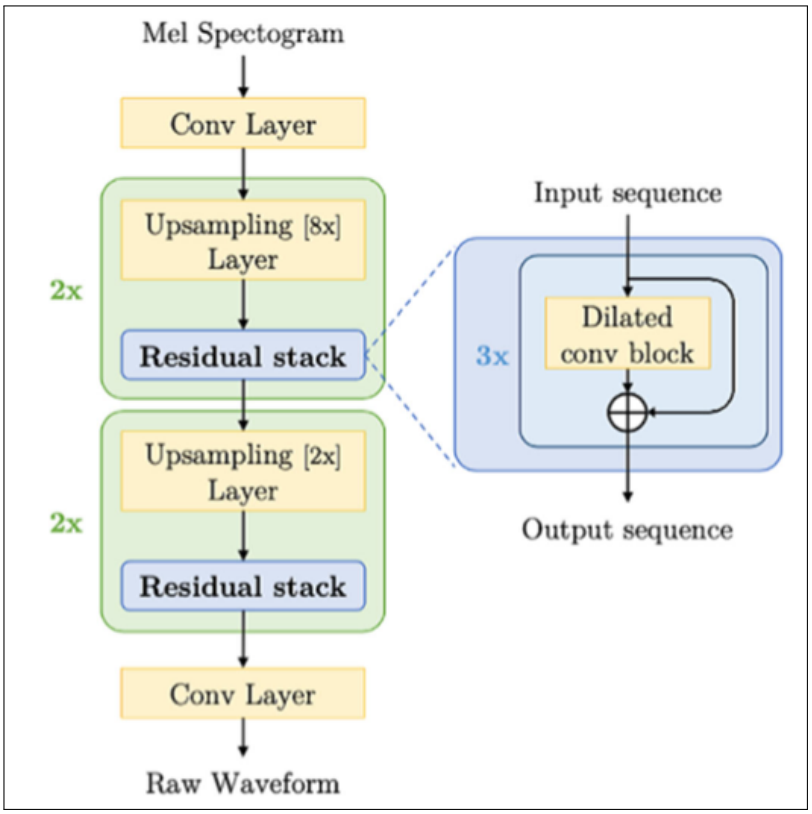


Figure 2.30: MelGAN generator architecture

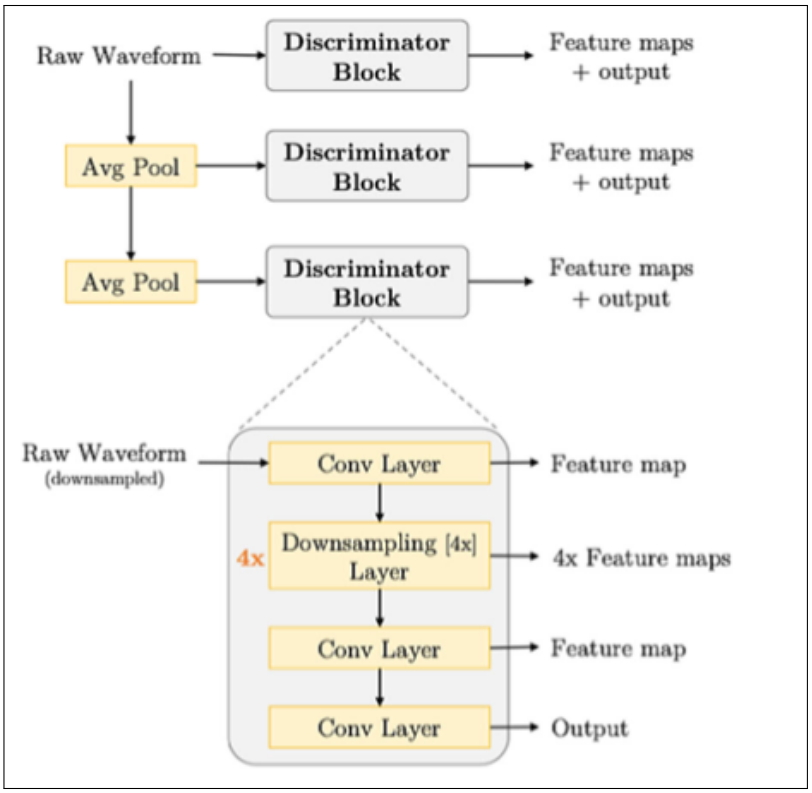


Figure 2.31: MelGAN discriminator architecture

Model	MOS
Wave glow	3.52
WaveGAN	3.72
MelGAN	4.11
Original	4.46

Figure 2.32: Comparison of MelGAN with MOS

Chapter 3

Hardware and Software Requirements

3.1 Hardware Requirements:

- **Microphone:** A good-quality microphone is essential for capturing clear and accurate speech input. It must have good sensitivity and frequency response so that it can capture the human speech without distortion. Since the system is employed in a mobile application, it uses the mobile microphone which are of type MEMS. MEMS microphones generally possess a frequency range of 20 Hz-20 KHz and a sensitivity between –46 dBV and –35 dBV.

- **Processor:** A powerful central processing unit is required to handle tasks such as speech recognition, text translation and synthesis. A modern multi core processor with x86-64 architecture is recommended.CPUs with higher clock speeds ranging from 1 GHz to 3.8GHz are beneficial for real-time applications as they can process instructions more quickly.

- **Memory (RAM):** Sufficient RAM is necessary for processing language models and handling the translation task effectively.LSTM models require substantial memory for efficient training and inference. Having higher RAM capacity can further improve performance, especially for a system with multiple models that deals with large datasets. Hence, a system with at least 16 GB RAM or higher is required.

- **Graphics Processing Unit (GPU):** A powerful Graphics Processing Unit (GPU) is essential for real-time speech translation systems, especially those using LSTM, due to the intense computational requirements of deep learning models.NVIDIA processors with CUDA support are used for machine learning tasks that uses deep learning models.

- **Storage:** A minimum of 8GB is required store huge audio datasets. Faster storage is preferable, such as Solid State Drive(SSD) especially when dealing with large datasets and frequent read/write operations during model training and inference.

3.2 Software Requirements:

- **Operating Software:** Since the project developed is a mobile application, an operating system is required for its smooth operation. The application is developed in such a way that it must run in both android and iOS.

- **Machine Learning Tools:**

  - TensorFlow and Keras

    TensorFlow is a free and open-source software library for machine learning.It can be used across a range of tasks but has a particular focus on training and inference of deep neural networks. Keras is the high-level API of the TensorFlow platform. It provides an approachable, highly-productive interface for solving machine learning (ML) problems, with a focus on modern deep learning. TensorFlow and Keras are together used for extensive training and implementation of the LSTM models.

- **Back End Framework:**

  - **Django**:

    Django is a free and open-source, Python-based web framework that runs on a web server. Django provides a robust set of tools and features for interacting with databases and a built-in admin interface for managing application data. It us used to integrate and facilitate the transfer of data between various components in the system.

- **Front End Developement tool:**

  - **React Native**

    React Native is an open-source UI software framework used to develop applications that run on multiple platforms. It consist of libraries that provide

functionalities which support both iOS and Android. Here, it is used to develop the application for taking the malayalam speech input from the user and provides the translated English speech as the output.

Chapter 4

System Architecture

4.1 System Overview

Malayalam to English Real-time speech conversion system involves a three-step process.This three-model approach integrates the strengths of LSTM and Bidirectional LSTM architectures, providing a comprehensive solution for real-time speech conversion from Malayalam to English.

- **Malayalam Speech to Text Conversion (LSTM-based Model):**
  - This model converts spoken Malayalam into text accurately.It uses a recurrent neural network (RNN) known as a Long Short-Term Memory (LSTM) network, set up in an encoder-decoder architecture within a sequence-to-sequence (Seq2Seq) model. Relevant data is captured by the encoder as it processes the input sequence of acoustic features taken from the Malayalam voice. The matching textual output is produced by the decoder in turn. Accurately reproducing the subtle nuances of spoken Malayalam depends on LSTM's capacity to record long-term dependencies. This model is trained using a wide collection of Malayalam speech samples, which guarantees that the model can accommodate a range of accents, intonations, and speaking speeds. The model architecture involves inputting Mel-frequency cepstral coefficients (MFCC) extracted from audio files and producing tokenized Malayalam text as output. Utilizing the Indic BERT tokenizer, the text is tokenized and padded to a maximum length of 40 tokens, while the audio features are padded to match a maximum length of 40 coefficients. The model was trained for 30 epochs with a batch size of 32 and optimized using the Adam optimizer with a sparse categorical cross entropy loss function. During training, 20% of the dataset was reserved for

validation. This LSTM model contributes to the advancement of speech technology for Malayalam, facilitating the conversion of spoken Malayalam into textual form.

- **Malayalam Text to English Text Conversion (Bidirectional LSTM-based Model):**
  - For the translation stage, the second model in the pipeline uses a Bidirectional LSTM (BiLSTM) network, which makes it easier to accurately translate Malayalam text into English text.This model makes use of the bidirectional nature of the LSTM to record contextual dependencies in both forward and backward directions. It is also organized within a Seq2Seq framework. The Malayalam text is processed by the encoder, and the matching English text is produced by the decoder. The BiLSTM model is highly proficient in comprehending the structural and contextual subtleties inherent in Malayalam text, enabling precise language translation. It was trained on parallel datasets that comprised aligned phrases in both Malayalam and English. The model architecture involved an encoder-decoder framework, with the encoder consisting of an embedding layer followed by a Bidirectional LSTM layer to capture the contextual information of the Malayalam text. Similarly, the decoder comprised an embedding layer, an LSTM layer, and a Dense layer to generate the English translations based on the encoded information. Tokenization was performed using the BERT-base-uncased tokenizer for English and the Indic BERT tokenizer for Malayalam, followed by padding to ensure uniform sequence lengths. The dataset was split into training and validation sets with an 80-20 split, and the model was trained using the Adam optimizer and sparse categorical cross entropy loss function over 30 epochs with a batch size of 64.

- **English Text to English Speech Conversion (Bidirectional LSTM-based TTS Model):**
  - A Bidirectional LSTM-based Text-to-Speech (TTS) model is used in the last stage. Natural-sounding English voice is synthesized by this model using the translated English text. The LSTM's bidirectional nature enhances the nat-

uralness of the synthesized speech by enabling the model to accurately capture the written text's contextual flow and intonation patterns. The SpeechT5 framework consists of a shared encoder-decoder network and six modal-specific (speech/text) pre/post-nets. After preprocessing the input speech/text through the pre-nets, the shared encoder-decoder network models the sequence-to-sequence transformation, and then the post-nets generate the output in the speech/text modalit based on the output of the decoder. The model architecture starts by loading the LJ Speech dataset, which contains audio and corresponding transcripts. Then preprocess the dataset by converting audio samples into a uniform sampling rate, tokenizing the text transcripts, and preparing speaker embeddings using a pre-trained Speaker Recognition model. After preprocessing, define a data collator to handle padding and batching of input features and labels. Specify training arguments such as batch size, learning rate, and evaluation strategy. Then, initialize the Seq2SeqTrainer with these arguments, along with the SpeechT5 model, training, and evaluation datasets. During training, the model is fine-tuned on the training dataset while periodically evaluating performance on the test dataset. Use gradient accumulation, gradient checkpointing, and mixed-precision training to efficiently utilize computational resources and speed up training. Training progress is logged, and the best-performing model is saved.

4.2 Architectural Design

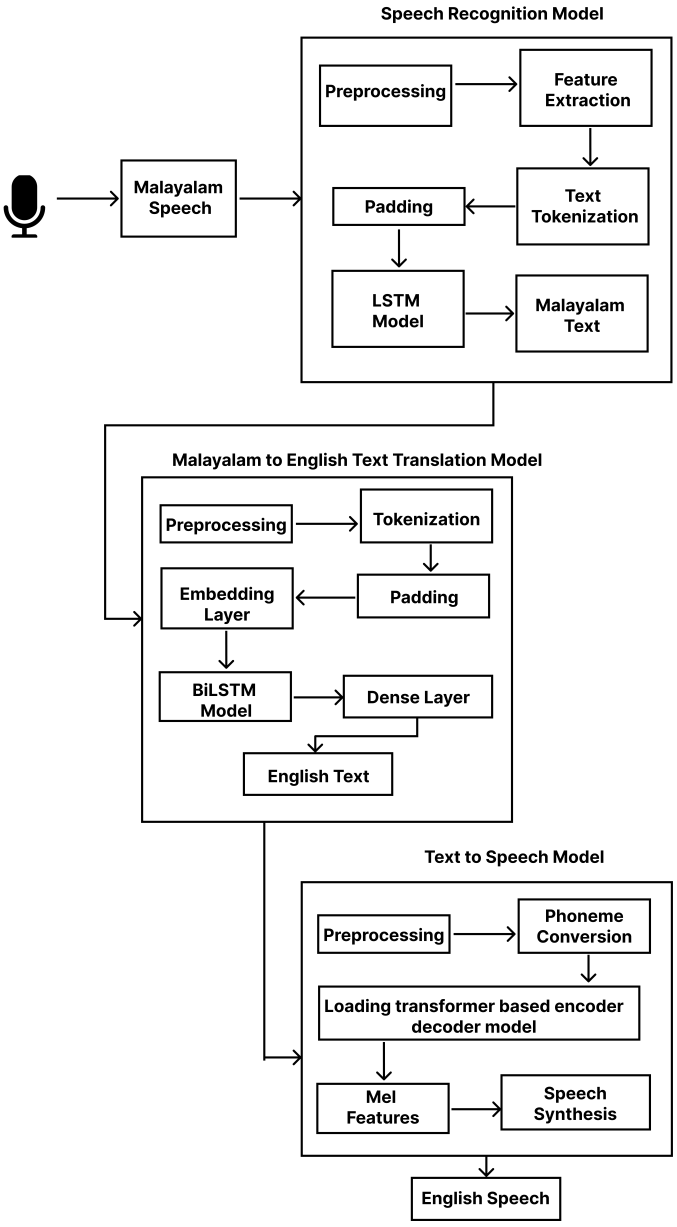


Figure 4.1: Architecture Diagram

4.3 Sequence Diagram

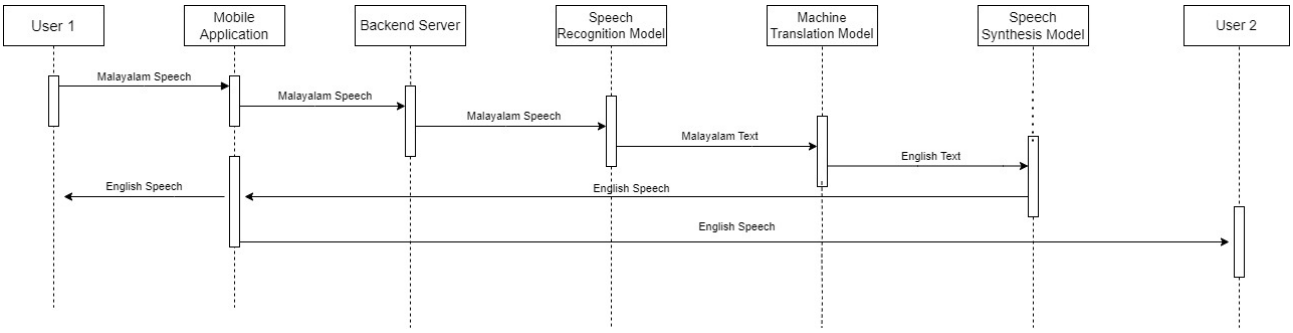


Figure 4.2: Sequence diagram

4.4 Module Division

4.4.1 Speech Recognition Module:

The Speech Recognition Module converts spoken words in Malayalam given as the input to corresponding Malayalam text. It follows a multistep process:

1. The module begins by receiving an audio signal containing spoken Malayalam input audio.
2. Through signal processing techniques, the recieved audio signal is analyzed to extract relevant features in the form of Mel Frequency Ceptral Coefficients(MFCC).
3. The obtained audio features are padded to a length of 40 coefficients and tokenized Malayalam text is produced as the output.
4. The text is tokenized and padded to a length of maximum 40 tokens using IndicBERT tokenizer.

The end result is an accurate transcription of the spoken words in Malayalam, providing a foundation for further language understanding.

4.4.2 Machine Translation Module:

The Machine Translation Module focuses on translating Malayalam text to English text. The process involves the usage of encoder-decoder mechanism that involves several key steps:

1. The tokenized malayalam text is fed into the encoder.

2. The encoder conmprises an embedding layer followed by a Bidirectional LSTM layer to capture the contextual information of the Malayalam text.

3. The decoder uses an embedding layer, an LSTM layer, and a Dense layer to generate the English translations based on the encoded information.

4. Tokenization was performed using the BERT-base-uncased tokenizer for English followed by padding to ensure uniform sequence lengths.

The outcome is an English version of the input Malayalam text that maintains semantic fidelity.

4.4.3 Speech Synthesis Module:

The Speech Synthesis Module is responsible for converting English text to spoken words through SpeechT5 framework that has a shared encoder-decoder network and six modal-specific (speech/text) pre/post-nets.. This involves:

1. The module takes sentences in English as input followed by preprocessing through the prenets.

2. The encoder-decoder network models the sequence-to-sequence transformation, and then the post-nets generate the output in the speech/text modalit based on the output of the decoder.

The module's goal is to create an output that closely resembles human speech, providing a seamless and intelligible listening experience.

4.4.4 User Interface: Mobile App

The User Interface is a mobile app designed to offer a user-friendly experience for individuals who want to seamlessly translate spoken Malayalam to English speech. Key features of the app include:

- Speech Input: Users can easily input spoken Malayalam through the app's intuitive interface.

- Translation Output: The app provides an instant and accurate translation of the spoken Malayalam into English speech.
- User-Friendly Design: The interface is designed for simplicity and accessibility, ensuring a positive user experience.

The Mobile App serves as the gateway for users to interact with the entire speech translation system, making language translation more accessible and efficient.

4.4.5 Dataset:

The dataset used in the development of the speech translation system plays a crucial role in training and evaluating the various modules. Key aspects of the dataset include:

- Multilingual Corpus: A diverse dataset containing spoken Malayalam paired with its corresponding English translations.
- Transcription Accuracy: Ensuring high-quality transcriptions of spoken Malayalam to facilitate accurate training of the speech recognition module.
- Translation Pairs: An extensive collection of Malayalam-English translation pairs to train and evaluate the machine translation module effectively.
- Text-to-Speech Examples: Samples of English sentences for training and testing the speech synthesis module.

The dataset's quality and diversity directly impact the performance of the entire speech translation system.

4.4.6 Django Back-end Server:

The back-end server acts as the foundation for translating the Malayalam input speech to corresponding English speech . Key features and components include:

- API Endpoints: Providing well-defined API endpoints for communication with the mobile app for the sending and receiving of input and output speech respectively.

- Model Deployment: The models for speech recognition, machine translation and speech synthesis is implemented in the back-end server. It receives the input speech from the application and processes it by passing the input through the models and returns the output English speech back to the mobile application

- Scalability: The server is designed such that it is able to handle the increasing workload without compromising the performance and functionality..

- Data Processing: Performing real-time data processing, including signal processing for speech recognition and translation tasks.

- Communication: The back-end server acts as an intermediary between the two mobile applications for sending translated speech in both directions.

The backend server forms the backbone of the speech translation system, orchestrating the flow of data and computations to deliver accurate and timely results.

4.5 Module wise diagram

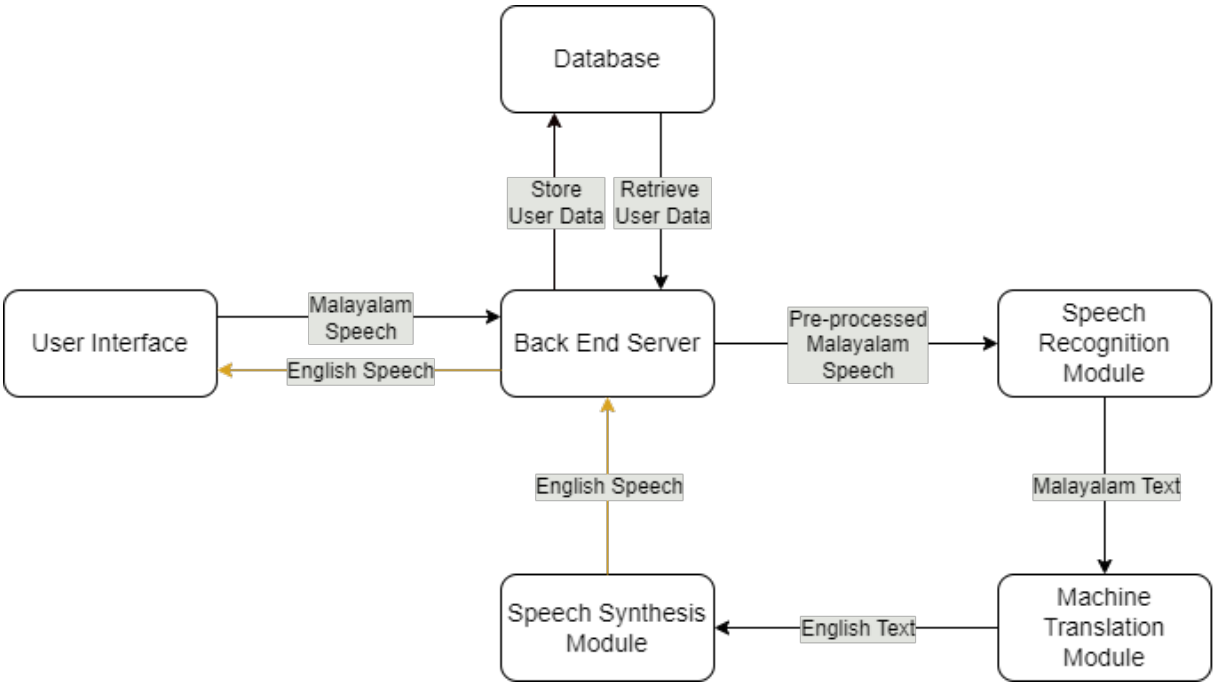


Figure 4.3: Module wise diagram

4.6 Work Breakdown & Responsibilities

4.6.1 User Interface

Front-End

Developed by Adhithyan R and Akhil Jose Francis using React Native. The application features a screen with microphone icon which can used to capture audio and send to Django server as POST request. The response received as string is displayed on the screen.

4.6.2 Speech Recognition Module

Back-End

Managed by Abijith Lohidakshan and Ajay A. The audio received as POST request from Front-End is processed and passed to LSTM Model which converts it to speech in string format and sends it as response.

Dataset Collection

Performed by Adhithyan R and Akhil Jose Francis. Malayalam Speech-to-Text dataset was collected from Kaggle.

LSTM Model

Developed by Abijith Lohidakshan and Ajay A. Utilises Tensorflow library for creation and training of the model. The model consists of encoder, decoder and embedding layer. The encoder captures and encodes the relevant information from the input to create a context vector. The embedding layer creates dense vectors for the word indices in the vocabulary. The decoder uses context vector from encoder and produces target sequence.

4.6.3 Ljspeech dataset preprocessing for text to speech synthesis model

Done by Adhithyan R and Akhil Jose Francis.

4.6.4 Malayalam speech recognition

Done by Adhithyan R and Akhil Jose Francis. Developed a LSTM neural network model using the IMaSC dataset. The model architecture involves inputting Mel-frequency cepstral coefficients (MFCC) extracted from audio files and producing tokenized Malayalam text as output.The model was trained for 30 epochs with a batch size of 32 and optimized using the Adam optimizer with a sparse categorical crossentropy loss function.

4.6.5 Malayalam to english text translation

Done by Abijith Lohidakshan and Ajay A. Used the Olam Dataset comprising pairs of Malayalam sentences and their corresponding English translations. Uses Bidirectional LSTM layer. The decoder comprised an embedding layer, an LSTM layer, and a Dense layer. The dataset was split into training and validation sets with an 80-20 split, and the model was trained using the Adam optimizer and sparse categorical crossentropy loss function over 30 epochs with a batch size of 64.

4.6.6 English text to speech

Done by Abijith Lohidakshan and Ajay A. They trained a transformer based encoder decoder model for text to speech generation. They used ljspeech dataset for finetune and training the transformer encoder decoder model. The transformer model here used for training is speecht5 model.

4.6.7 Training and testing for text to speech synthesis model

Done by Adhithyan R and Akhil Jose Francis. we load the LJ Speech dataset. preprocess the dataset by converting audio samples into a uniform sampling rate, tokenizing the text transcripts, and preparing speaker embeddings using a pre-trained Speaker Recognition model. During training, the model is fine-tuned on the training dataset while periodically evaluating performance on the test dataset. We use gradient accumulation, gradient checkpointing, and mixed-precision training to efficiently utilize computational resources and speed up training. Training progress is logged, and the best-performing model is saved.

4.6.8 Django project creation, module integration

Done by Abijith Lohidakshan and Ajay A. With the development of a custom application suited to various capabilities, the Django project's launch signaled the commencement of painstaking groundwork. Every aspect of this project, from specifying the structure of the application to setting up important URL endpoints, showed meticulous forethought. These endpoints were entry points to particular functions, such as user authentication and user interaction facilitation, which were inextricably related to these endpoints' activities. The project sought to provide a smooth user experience by methodically integrating necessary functionalities, like signing up and logging in. What emerged was a solid Django application ready to achieve its goals precisely and effectively.

4.6.9 Integration

Done by Abijith Lohidakshan and Ajay A. Integration of the 3 modules was done to finalize the project.

4.7 Work Schedule - Gantt Chart

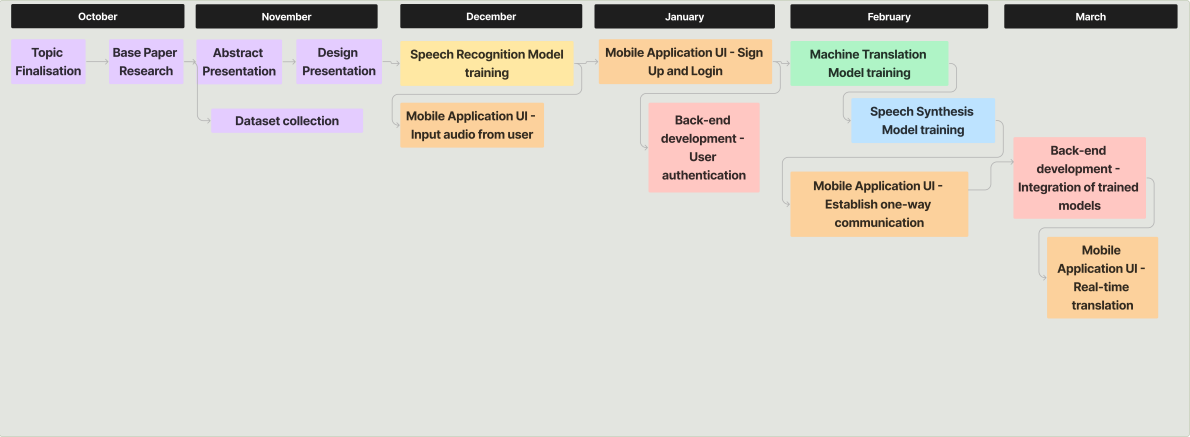


Figure 4.4: Gantt Chart

SLNO.	MODULE	SCHEDULE
1	LITERATURE SURVEY	OCTOBER 2023
2	DATASET COLLECTION	NOVEMBER 2023
3	SPEECH RECOGNITION MODEL DEVELOPMENT	DECEMBER 2023
4	GUI AND BACKEND DEVELOPMENT	JANUARY 2024
5	MACHINE TRANSLATION MODEL DEVELOPMENT	FEBRUARY 2024
6	SPEECH SYNTHESIS MODEL DEVELOPMENT	MARCH 2024

Table 4.1: Work Schedule

Chapter 5

Conclusion

The introduction of a real-time speech translation technology from Malayalam to English is a critical step in promoting international understanding and communication. By acting as a catalyst, this cutting-edge system breaks down the conventional barriers preventing cross-cultural relationships. Through the smooth translation of spoken Malayalam into English, it increases inclusivity and makes it possible for a wider range of people to interact with content from around the world.

Despite these obstacles, there is much promise for this kind of technology in terms of bridging linguistic divides and promoting efficient international communication. Also, this system supports efficient message passing.It can bring people together across language barriers, facilitate smooth communication, and allow knowledge and ideas to be shared without language barriers. The system's importance in a worldwide society where efficient communication is crucial is shown by its capacity to bridge the gap between Malayalam and English, two languages with different linguistic structures and cultural contexts.

In summary, the creation of a real-time speech translation system from Malayalam to English poses significant challenges in terms of precision, ambiguity, and context preservation. However, the system's overall potential to overcome language barriers and promote efficient communication on a worldwide level cannot be overstated. With the help of this technical innovation, we may move closer to a society where people understand each other better and where language barriers no longer prevent people from exchanging ideas and cultures.

Chapter 6

Result

Important progress was made in the areas of speech and text processing for the Malayalam and English languages as a result of the project. Using the IMaSC dataset, a reliable Long Short-Term Memory (LSTM) model was created for text translation of spoken Malayalam. This model improved Malayalam speech-to-text conversion skills by demonstrating excellent accuracy in tokenized Malayalam text production from audio inputs. Furthermore, an advanced text translation model demonstrating remarkable accuracy with high semantic similarity measurements was developed to translate Malayalam sentences into English with ease. Regarding the English language, a transformer-based model was effectively trained to provide high-quality synthesized speech output from textual inputs using text-to-speech conversion. Techniques like the sparse categorical cross entropy loss function and Adam optimizer were used to optimize the training process, guaranteeing effective convergence and top performance on validation and test datasets. Additionally, the project effectively showcased the potential of voice synthesis, garnering favorable comments from users and emphasizing the model's capacity to produce speech that sounds natural based on text inputs. All things considered, these findings represent noteworthy progressions in Malayalam and English voice technology, benefiting the larger domain of natural language processing.

References

[1] S. R. A. S. A. Arun HP, Jithin Kunjumon, "Malayalam speech to text conversion using deep learning," *IOSR Journal of Engineering*, 2021.

[2] C. V. B. B. R. P. S. Akshay K, Aravind Das A. M, "Real time translation of malayalam notice boards to english directions," *International Journal of Computer Applications*, 2019.

[3] F. X. F. K. S. Yuchen Fan, Yao Qian, "Tts synthesis with bidirectional lstm based recurrent neural networks," *International Speech Communication Association*, 2014.

[4] e. L. B. C. M. C. Anushka Chaudhari, Vina M. Lomte, "Development of speaker-independent automatic speech recognition system for marathi language," *Indian Journal of Science and Technology*, vol. 8, 2023.

[5] K. N. Pushpalatha, K. S. Ravikumar, M. S. Kasyap, M. H. S. Murthy, and J. Paul, "Kannada to english machine translation using lstm," *Internation Information & Engineering Technology Association*, 2020.

[6] M. Araya and M. Alehegn, "Text to speech synthesizer for tigrigna linguistic using concatenative based approach with lstm model," *Indian Journal of Science and Technology*, 2022.

[7] G. Atkar and P. Jayaraju, "Speech synthesis using generative adversarial network for improving readability of hindi words to recuperate from dyslexia," *Neural Computing and Applications*, 2021.

Appendix A:
FINAL PRESENTATION

MALAYALAM-TO-ENGLISH SPEECH TRANSLATION SYSTEM

ABIJITH LOHIDAKSHAN
ADHITHYAN R
AJAY A
AKHIL JOSE FRANCIS

GUIDE: SANDY JOSEPH

PROBLEM DEFINITION

- In our increasingly interconnected world, effective communication across language barriers is more critical than ever.
- The project aims to create a speech translation system that can convert spoken Malayalam into English.
- This project addresses the need for a transition system from Malayalam to English.

PROJECT
OBJECTIVES

- The primary objective of the project is to design, develop, and implement a one-way communication speech translation system that converts spoken Malayalam into spoken English

.

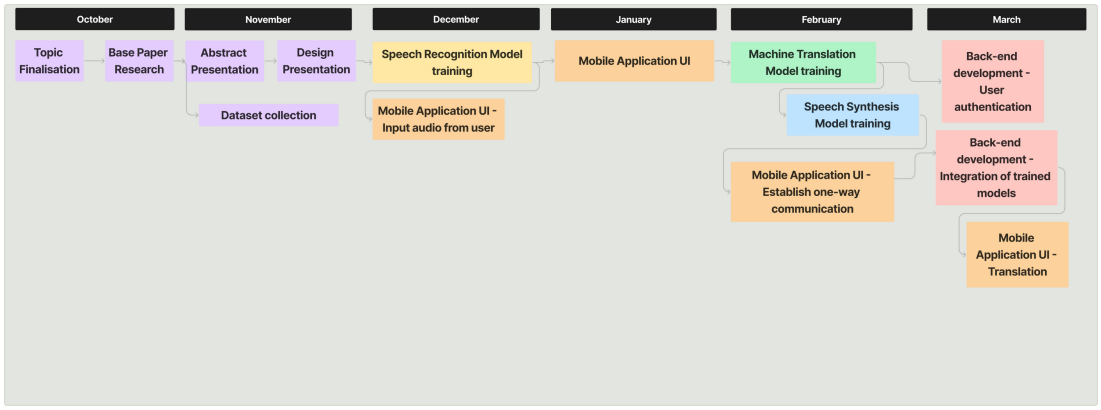
- This system will be integrated into a user-friendly application, allowing users to make calls and provide spoken Malayalam as input, with the system delivering English speech as the output.

NOVELTY OF IDEA & SCOPE OF IMPLEMENTATION

- Seamless integration of speech-to-text, text translation, and text-to-speech functionalities, enabling the complete conversion of Malayalam to English. By combining these components, the model offers a comprehensive solution for language translation, bridging the gap between Malayalam and English speakers. Its ability to accurately translate and synthesize speech enhances accessibility and communication across linguistic boundaries.

- Used in airports, train stations, and public transportation hubs, tourism offices and visitor centers, museums and cultural institutions, hospitals and healthcare facilities etc.

PROJECT
GANTT CHART



WORK DONE DURING
30% EVALUATION

Implemented Speech to text model

- The speech-to-text module converts spoken Malayalam into text through feature extraction, encoding, embedding. This process ensures an accurate transcription of spoken words by understanding language context and decoding the audio signal effectively.

Mobile App

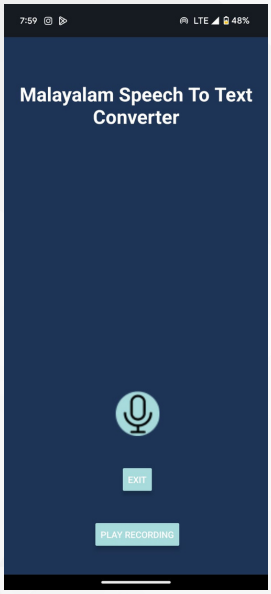
- The User Interface is a mobile app designed to offer a user-friendly experience for individuals to input Malayalam speech.

WORK DONE DURING
60% EVALUATION

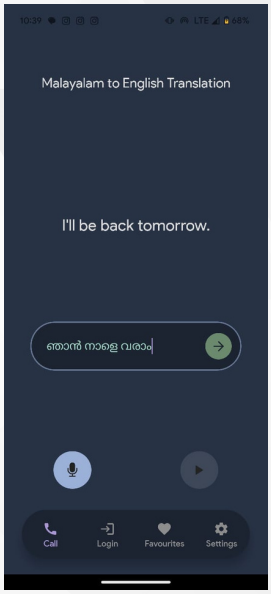
- Implemented Machine Translation Module.
  - Involves the development of an Bidirectional LSTM model that translates text in English language to its corresponding Malayalam text.
  - The model processes the preprocessed input text in forward and backward directions for capturing the contextual information.

- Developed mobile application
  - A mobile application was created which provides a playback function for Malayalam speech input and provides English text as the output for Malayalam text input.
  - The app sends the malayalam text input to the Django server that uses the Bi LSTM model to predict the English text which is returned back to application

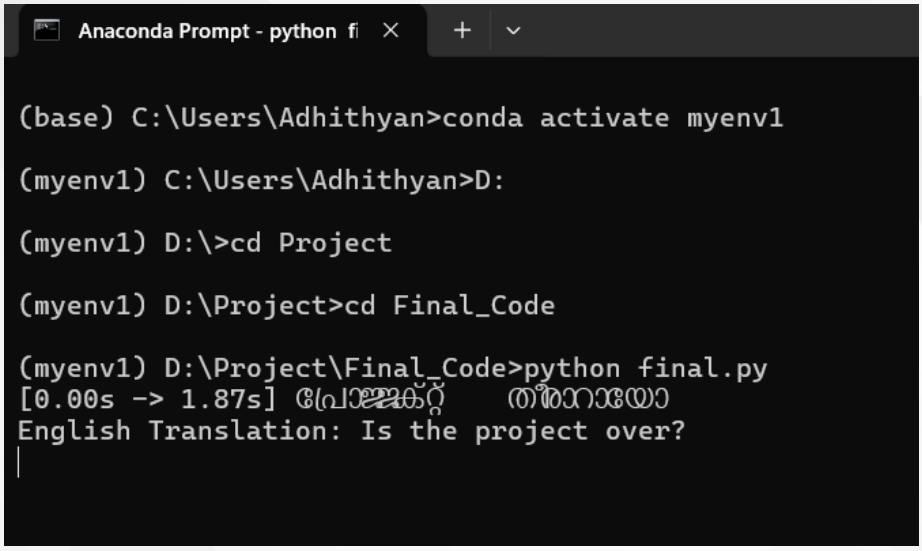
INTERIM RESULTS



30%



60%



100%

WORK DONE DURING
100% EVALUATION

- Implementation of Malayalam speech to English speech translation system.
- The system consist of three main modules:
  - Speech recognition module
  - Machine translation module
  - Speech synthesis module
- The system is presented through a mobile application that accepts Malayalam speech as the input from one user's device.The English speech is produced as the output in other user's device.
- It uses a Django server module that does the translation process through the above three models implemented using LSTM.

FUTURE SCOPE

- The future scope of a Malayalam to English speech translation system deals with the further improvements and expansion that can be done on the system.
- The system's coverage can be extended by making the system able to support multiple regional languages as input and converting it into corresponding English speech.
- Another scope of expansion can be done by integrating the system with AI chatbots so that the user can give the input in the preferred regional language.
- There are numerous other areas in which system can be improved such as cross-platform integration, domain specific translation, privacy and security and adaptation to different slangs and dialects.

TASK DISTRIBUTION

- **Akhil Jose Francis**
  - Front-end mobile application
  - Dataset collection for Malayalam speech to text and English text to speech
  - Preprocessing the Malayalam speech to text dataset
  - Preprocessing the English text to speech dataset
  - Feature Extraction for Malayalam speech to text dataset
- **Adhithyan R**
  - Front-end mobile application
  - Creating model architecture for Malayalam speech to text
  - Training and evaluation
  - Dataset collection for Malayalam text to English text

TASK DISTRIBUTION

- **Ajay A**
  - Preprocessing Malayalam text to English text dataset
  - Architecture creation
  - Training and evaluation
  - Back-end development using Django
  - Final integration

- **Abijith Lohidakshan**
  - Feature extraction for English text to speech
  - Architecture creation for English text to speech
  - Training and evaluation
  - Back-end development using Django
  - Final integration

CONCLUSION

- In conclusion, a Malayalam to English speech translation system serves as a major step in promoting global communication and understanding.
- The system also bridges the communication barriers which makes the information accessible to a wide variety of audience, fostering inclusivity.
- The development of such a system involves challenges such as accuracy, ambiguity, and context preservation.
- Overall, this system holds significant potential for breaking down language barriers and facilitating effective communication in a globalized world.

REFERENCES

- Malayalam Speech to Text Conversion Using Deep Learning.Arun HP, Jithin Kunjumon, Sambhunath , Ancy S Ansalem;IOSR Journal of Engineering.
- Real Time Translation of Malayalam Notice Boards to English Directions. Akshay K, Aravind Das A.M, Carral Vincent, Betty Babu, Rasmi P.S.International Journal of Computer Applications (0975 – 8887),Volume 178 – No. 26.
- TTS Synthesis with Bidirectional LSTM based Recurrent Neural Networks. Yuchen Fan , Yao Qian, Fenglong Xie, Frank K. Soong.
- Machine Translation between Malayalam & English.Dr.Sreelekha S. Linguistics Journal volume 14

REFERENCES

- Development of Speaker-Independent Automatic Speech Recognition System for Kannada Language; Praveen Kumar, H S Jayanna; INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY
- Text to Speech Synthesizer for Tigrigna Linguistic using Concatenative Based approach with LSTM model, Mezgebe Araya , Minyechil Alehegn, INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY
- Kannada to English Machine Translation Using Deep Neural Network; Pushpalatha Kadavigere Nagaraj, Kshamitha Shobha Ravikumar, Mydugolam Sreenivas Kasyap, Medhini,Hullumakki Srinivas Murthy, Jithin Paul; International Information and Engineering Technology Association
- Speech synthesis using generative adversarial network for improving readability of Hindi words to recuperate from dyslexia; Geeta Atkar, Priyadarshini Jayaraju; Neural Computing and Applications

STATUS OF PAPER PUBLICATION

- The contents of the article including the abstract,introduction,related works,proposed architecture,implementation, conclusion & result has been done.
- The draft has been submitted to our guide.

t

THANK YOU

Appendix B: Vision, Mission, Programme Outcomes and Course Outcomes

Vision, Mission, Programme Outcomes and Course Outcomes

Institute Vision

To evolve into a premier technological institution, moulding eminent professionals with creative minds, innovative ideas and sound practical skill, and to shape a future where technology works for the enrichment of mankind.

Institute Mission

To impart state-of-the-art knowledge to individuals in various technological disciplines and to inculcate in them a high degree of social consciousness and human values, thereby enabling them to face the challenges of life with courage and conviction.

Department Vision

To become a centre of excellence in Computer Science and Engineering, moulding professionals catering to the research and professional needs of national and international organizations.

Department Mission

To inspire and nurture students, with up-to-date knowledge in Computer Science and Engineering, ethics, team spirit, leadership abilities, innovation and creativity to come out with solutions meeting societal needs.

Programme Outcomes (PO)

Engineering Graduates will be able to:

1. Engineering Knowledge: Apply the knowledge of mathematics, science, engineering fundamentals, and an engineering specialization to the solution of complex engineering problems.

2. Problem analysis: Identify, formulate, review research literature, and analyze complex engineering problems reaching substantiated conclusions using first principles of mathematics, natural sciences, and engineering sciences.

3. Design/development of solutions: Design solutions for complex engineering problems and design system components or processes that meet the specified needs with appropriate consideration for the public health and safety, and the cultural, societal, and environmental considerations.

4. Conduct investigations of complex problems: Use research-based knowledge including design of experiments, analysis and interpretation of data, and synthesis of the information to provide valid conclusions.

5. Modern Tool Usage: Create, select, and apply appropriate techniques, resources, and modern engineering and IT tools including prediction and modeling to complex engineering activities with an understanding of the limitations.

6. The engineer and society: Apply reasoning informed by the contextual knowledge to assess societal, health, safety, legal, and cultural issues and the consequent responsibilities relevant to the professional engineering practice.

7. Environment and sustainability: Understand the impact of the professional engineering solutions in societal and environmental contexts, and demonstrate the knowledge of, and need for sustainable development.

8. Ethics: Apply ethical principles and commit to professional ethics and responsibilities and norms of the engineering practice.

9. Individual and Team work: Function effectively as an individual, and as a member or leader in teams, and in multidisciplinary settings.

10. Communication: Communicate effectively with the engineering community and with society at large. Be able to comprehend and write effective reports documentation. Make effective presentations, and give and receive clear instructions.

11. Project management and finance: Demonstrate knowledge and understanding of engineering and management principles and apply these to one's own work, as a member and leader in a team. Manage projects in multidisciplinary environments.

12. Life-long learning: Recognize the need for, and have the preparation and ability to engage in independent and lifelong learning in the broadest context of technological change.

Programme Specific Outcomes (PSO)

A graduate of the Computer Science and Engineering Program will demonstrate:

PSO1: Computer Science Specific Skills
The ability to identify, analyze and design solutions for complex engineering problems in multidisciplinary areas by understanding the core principles and concepts of computer science and thereby engage in national grand challenges.

PSO2: Programming and Software Development Skills
The ability to acquire programming efficiency by designing algorithms and applying standard practices in software project development to deliver quality software products meeting the demands of the industry.

PSO3: Professional Skills
The ability to apply the fundamentals of computer science in competitive research and to develop innovative products to meet the societal needs thereby evolving as an eminent researcher and entrepreneur.

Course Outcomes (CO)

Course Outcome 1: Model and solve real world problems by applying knowledge across domains (Cognitive knowledge level: Apply).

Course Outcome 2: Develop products, processes or technologies for sustainable and socially relevant applications (Cognitive knowledge level: Apply).

Course Outcome 3: Function effectively as an individual and as a leader in diverse teams and to comprehend and execute designated tasks (Cognitive knowledge level: Apply).

Course Outcome 4: Plan and execute tasks utilizing available resources within timelines, following ethical and professional norms(Cognitive knowledge level: apply)

Course Outcome 5: Identify technology/research gaps and propose innovative/creative solutions (Cognitive knowledge level: Analyze).

Course Outcome 6: Organize and communicate technical and scientific findings effectively in written and oral forms (Cognitive knowledge level: Apply).

Appendix C: CO-PO-PSO Mapping

CO-PO AND CO-PSO MAPPING

	PO 1	PO 2	PO 3	PO 4	PO 5	PO 6	PO 7	PO 8	PO 9	PO 10	PO 11	PO 12	PSO 1	PSO 2	PSO 3
CO 1	2	2	2	1	2	2	2	1	1	1	1	2	3		
CO 2	2	2	2		1	3	3	1	1		1	1		2	
CO 3									3	2	2	1			3
CO 4					2			3	2	2	3	2			3
CO 5	2	3	3	1	2							1	3		
CO 6					2			2	2	3	1	1			3

3/2/1: high/medium/low

JUSTIFICATIONS FOR CO-PO MAPPING

MAPPING	LOW/MEDIUM/ HIGH	JUSTIFICATION
100003/ CS722U.1-PO1	M	Knowledge in the area of technology for project development using various tools results in better modeling.
100003/ CS722U.1-PO2	M	Knowledge acquired in the selected area of project development can be used to identify, formulate, review research literature, and analyze complex engineering problems reaching substantiated conclusions.

100003/ CS722U.1-PO3	M	Can use the acquired knowledge in designing solutions to complex problems.
100003/ CS722U.1-PO4	M	Can use the acquired knowledge in designing solutions to complex problems.
100003/ CS722U.1-PO5	H	Students are able to interpret, improve and redefine technical aspects for design of experiments, analysis and interpretation of data, and synthesis of the information to provide valid conclusions.
100003/ CS722U.1-PO6	M	Students are able to interpret, improve and redefine technical aspects by applying contextual knowledge to assess societal, health and consequential responsibilities relevant to professional engineering practices.
100003/ CS722U.1-PO7	M	Project development based on societal and environmental context solution identification is the need for sustainable development.
100003/ CS722U.1-PO8	L	Project development should be based on professional ethics and responsibilities.
100003/ CS722U.1-PO9	L	Project development using a systematic approach based on well defined principles will result in teamwork.
100003/ CS722U.1-PO10	M	Project brings technological changes in society.

100003/ CS722U.1-PO11	H	Acquiring knowledge for project development gathers skills in design, analysis, development and implementation of algorithms.
100003/ CS722U.1-PO12	H	Knowledge for project development contributes engineering skills in computing & information gatherings.
100003/ CS722U.2-PO1	H	Knowledge acquired for project development will also include systematic planning, developing, testing and implementation in computer science solutions in various domains.
100003/ CS722U.2-PO2	H	Project design and development using a systematic approach brings knowledge in mathematics and engineering fundamentals.
100003/ CS722U.2-PO3	H	Identifying, formulating and analyzing the project results in a systematic approach.
100003/ CS722U.2-PO5	H	Systematic approach is the tip for solving complex problems in various domains.
100003/ CS722U.2-PO6	H	Systematic approach in the technical and design aspects provide valid conclusions.
100003/ CS722U.2-PO7	H	Systematic approach in the technical and design aspects demonstrate the knowledge of sustainable development.

100003/ CS722U.2-PO8	M	Identification and justification of technical aspects of project development demonstrates the need for sustainable development.
100003/ CS722U.2-PO9	H	Apply professional ethics and responsibilities in engineering practice of development.
100003/ CS722U.2-PO11	H	Systematic approach also includes effective reporting and documentation which gives clear instructions.
100003/ CS722U.2-PO12	M	Project development using a systematic approach based on well defined principles will result in better teamwork.
100003/ CS722U.3-PO9	H	Project development as a team brings the ability to engage in independent and lifelong learning.
100003/ CS722U.3-PO10	H	Identification, formulation and justification in technical aspects will be based on acquiring skills in design and development of algorithms.
100003/ CS722U.3-PO11	H	Identification, formulation and justification in technical aspects provides the betterment of life in various domains.
100003/ CS722U.3-PO12	H	Students are able to interpret, improve and redefine technical aspects with mathematics, science and engineering fundamentals for the solutions of complex problems.

100003/ CS722U.4-PO5	H	Students are able to interpret, improve and redefine technical aspects with identification formulation and analysis of complex problems.
100003/ CS722U.4-PO8	H	Students are able to interpret, improve and redefine technical aspects to meet the specified needs with appropriate consideration for the public health and safety, and the cultural, societal, and environmental considerations.
100003/ CS722U.4-PO9	H	Students are able to interpret, improve and redefine technical aspects for design of experiments, analysis and interpretation of data, and synthesis of the information to provide valid conclusions.
100003/ CS722U.4-PO10	H	Create, select, and apply appropriate techniques, resources, and modern engineering and IT tools for better products.
100003/ CS722U.4-PO11	M	Students are able to interpret, improve and redefine technical aspects by applying contextual knowledge to assess societal, health and consequential responsibilities relevant to professional engineering practices.
100003/ CS722U.4-PO12	H	Students are able to interpret, improve and redefine technical aspects for demonstrating the knowledge of, and need for sustainable development.
100003/ CS722U.5-PO1	H	Students are able to interpret, improve and redefine technical aspects, apply ethical principles and commit to

		professional ethics and responsibilities and norms of the engineering practice.
100003/ CS722U.5-PO2	M	Students are able to interpret, improve and redefine technical aspects, communicate effectively on complex engineering activities with the engineering community and with society at large, such as, being able to comprehend and write effective reports and design documentation, make effective presentations, and give and receive clear instructions.
100003/ CS722U.5-PO3	H	Students are able to interpret, improve and redefine technical aspects to demonstrate knowledge and understanding of the engineering and management principle in multidisciplinary environments.
100003/ CS722U.5-PO4	H	Students are able to interpret, improve and redefine technical aspects, recognize the need for, and have the preparation and ability to engage in independent and life-long learning in the broadest context of technological change.
100003/ CS722U.5-PO5	M	Students are able to interpret, improve and redefine technical aspects in acquiring skills to design, analyze and develop algorithms and implement those using high-level programming languages.
100003/ CS722U.5-PO12	M	Students are able to interpret, improve and redefine technical aspects and contribute their engineering skills in computing and information engineering domains like

		network design and administration, database design and knowledge engineering.
100003/ CS722U.6-PO5	M	Students are able to interpret, improve and redefine technical aspects and develop strong skills in systematic planning, developing, testing, implementing and providing IT solutions for different domains which helps in the betterment of life.
100003/ CS722U.6-PO8	H	Students will be able to associate with a team as an effective team player for the development of technical projects by applying the knowledge of mathematics, science, engineering fundamentals, and an engineering specialization to the solution of complex engineering problems.
100003/ CS722U.6-PO9	H	Students will be able to associate with a team as an effective team player to Identify, formulate, review research literature, and analyze complex engineering problems
100003/ CS722U.6-PO10	M	Students will be able to associate with a team as an effective team player for designing solutions to complex engineering problems and design system components.
100003/ CS722U.6-PO11	M	Students will be able to associate with a team as an effective team player, use research-based knowledge and research methods including design of experiments, analysis and interpretation of data.

100003/ CS722U.6-PO12	H	Students will be able to associate with a team as an effective team player, applying ethical principles and commit to professional ethics and responsibilities and norms of the engineering practice.
100003/ CS722U.1-PSO1	H	Students are able to develop Computer Science Specific Skills by modeling and solving problems.
100003/ CS722U.2-PSO2	M	Developing products, processes or technologies for sustainable and socially relevant applications can promote Programming and Software Development Skills.
100003/ CS722U.3-PSO3	H	Working in a team can result in the effective development of Professional Skills.
100003/ CS722U.4-PSO3	H	Planning and scheduling can result in the effective development of Professional Skills.
100003/ CS722U.5-PSO1	H	Students are able to develop Computer Science Specific Skills by creating innovative solutions to problems.
100003/ CS722U.6-PSO3	H	Organizing and communicating technical and scientific findings can help in the effective development of Professional Skills.