

TRIBHUVAN UNIVERSITY

INSTITUTE OF ENGINEERING

PASCHIMANCHAL CAMPUS, POKHARA

LAMACHAUR, POKHARA-16

[Subject code: EX707]

A MAJOR PROJECT PROPOSAL ON

“Speech recognition: Nepali Speech to text using Deep learning”

Submitted by:

CHANDRA SHEKHAR GIRI: PAS077BEI017

DIPESH KARKI: PAS077BEI018

NAVRAJ AWASTHI: PAS077BEI026

YOGENDRA BASKOTA: PAS077BEI045

Submitted to:

DEPARTMENT OF ELECTRONICS AND COMPUTER ENGINEERING

JESTHA, 2081

Lamachaur, Pokhara

**ABSTRACT**

Speech recognition technology has revolutionized human-computer interaction, facilitating seamless communication and accessibility across various domains. However, for languages like Nepali, which boasts a rich cultural heritage and a large speaker population, robust speech recognition solutions remain scarce. This project presents the development of a Nepali Speech-to-Text Recognition System using Deep Learning techniques on the Raspberry Pi platform. The system aims to bridge the gap in speech technology for the Nepali language, enabling accurate transcription of spoken Nepali words and phrases into text in real-time. Leveraging the versatility of Raspberry Pi hardware and the power of Deep Learning frameworks such as TensorFlow, the system integrates seamlessly with microphone and LCD modules to provide user-friendly speech recognition capabilities. Through a combination of data collection, model development, system integration, and evaluation, the project endeavors to deliver an accessible and efficient solution for Nepali speakers, fostering technological inclusivity and empowerment.

**Table of contents**

CHAPTER 1: INTRODUCTION...........................................................................6

1.1 Background...................................................................................................6

1.2 Problem Statement........................................................................................6

1.3 Objectives.....................................................................................................7

CHAPTER 2: LITERATURE REVIEW..................................................................7

CHAPTER 3: METHODOLGY..............................................................................8

3.1 Hardware Implementation...........................................................................8

3.2 System Architecture...................................................................................11

CHAPTER 4: SYSTEM DESIGN..................................................................

4.1 Requirement specification

4.1.1 Functional Requirements

4.1.1 Nonfunctional Requirements

4.2 Feasibility Analysis

4.2.1 Technical feasibility

4.2.2 Financial feasibility

4.2.3Market feasibility

4.2.4Operational feasibility

CHAPTER 5: Tools and techniques

5.1Hardware Requirement

5.1.1Raspberry Pi

5.1.2Microphone

5.1.3LCD module

5.2Software Requirement

5.2.1 Raspberry pi OS

5.2.2 Python

5.2.3 TensorFlow

CHAPTER 4: EPILOGUE....................................................................................12

4.1 Expected output.......................................................................................................12

4.2 Work Schedule.................................................................................................12

4.3 Budget estimation..................................................................................................12

CHAPTER 6 : Conclusion

REFERENCES........................................................................................................13

LIST OF FIGURES

Fig 1: Arduino Uno.........................................................................................................8

Fig 2: MPU 6050............................................................................................................9

Fig 3: Accelerometers.....................................................................................................9

Fig 4: GPS NEO-6M.......................................................................................................10

Fig 5: GSM Sim800L.....................................................................................................10

Fig 6: Schematic Diagram of the AAS............................................................................11

**Introduction**

Background

Over the past few decades, cell phones have become an indispensable source of communication for modern society and there has been a cosmic increase in the quantity of educational sectors everywhere throughout the world. Due to this enormous increase, education has become the doorstep of every individual. A large portion of these educational institutions are utilizing the same old procedures that are time consuming and more tedious for typing and especially for the fast typing. We can make calls and text messages from a source to a destination easily. It is known that verbal communication is the most appropriate modern of passing on and conceiving the correct information, avoiding misquotations.

To fulfill the gap over a long distance, verbal communication can take place easily on phone calls. A path-breaking innovation has recently come to play in the SMS technology using the speech recognition technology, where voice messages are being converted to text messages. Quite a few applications used to assist the disabled make use of speech to text translation. In modern civilized societies for communication between human speeches is one of the common methods. Different ideas formed in the mind of the speaker are communicated by speech in the form of words, phrases, and sentences by applying some proper grammatical rules. The speech is primary mode of communication among human being and also the most natural and efficient form of exchanging information among human in speech.

Most of the information in the digital world is available to a few who can read or understand a scrupulous language. Language technologies can provide solutions in the form of ordinary interfaces so the digital content can reach the masses and facilitate the exchange of information across different people speaking different languages. These technologies play a vital role in multilingual societies such as Nepal which has about 124 native languages. Speech to Text conversion takes input from the microphone in the form of speech & then it is converted into text form which is displayed on the screen. Speech processing is the study of speech signals, and the various methods which are used to process them. In this process various applications such as speech coding, speech synthesis, speech recognition and speaker recognition technologies; speech processing is employed.

Among the above, speech recognition is the most important one. The main purpose of speech recognition is to convert the acoustic signal obtained from a microphone or a telephone to generate a set of words. Words are generated by converting the analog signal into the digital signal and then plotting into the graph between the frequency and the time. The digital signal is converted into the frequency spectrum using the Fast Fourier transform, which creates a brighter pattern for the higher frequency terms and darker for the low frequency components. In order to extract and determine the linguistics conveyed by a speech wave we have to employ computers & electronic circuits. This process is performed for several applications such as security devices, household appliances, cellular phones, ATM machines and computers.

Analog audio must be converted into digital signals to be processed. This requires analog-to-digital conversion. For a computer to decipher the signal, it must have a digital database, or vocabulary, of words or syllables, and a speedy means of comparing this data with signals. The speech patterns are stored on the hard drive and loaded into memory when the program is run. A comparator checks these stored patterns against the output of the A/D converter.

Problem Statement

The lack of robust speech recognition systems for the Nepali language poses a barrier to technological accessibility and innovation for Nepali-speaking communities. Existing solutions are often limited in their accuracy, adaptability, and accessibility, failing to cater to the diverse linguistic nuances and socio-cultural contexts inherent in Nepali speech. This presents a pressing need for the development of a tailored speech recognition system capable of accurately transcribing Nepali speech into text, thereby unlocking a myriad of applications and opportunities for Nepali speakers.

Objectives

* Develop a robust Nepali Speech-to-Text Recognition System using Deep Learning techniques and integrate the system with Raspberry Pi to enhance accessibility and usability.
* Bridge the gap in speech technology for the Nepali language, empowering Nepali-speaking communities with innovative and inclusive technology solutions.

**Literature Review**

Supriya Khadka, Ranju G.C., Prabin Paudel, Rahul Shah, Basanta Joshi, Nepali Text-to-Speech Synthesis using Tacotron2 for Melspectrogram Generation[1], Proposes generating high-quality Nepali speech using the Tacotron2 model and HiFiGAN/WaveGlow vocoders. This method, which preprocesses text and fine-tunes the Tacotron2 model, achieved a record Mean Opinion Score of 4.03 for naturalness in Nepali Text-to-Speech tasks.

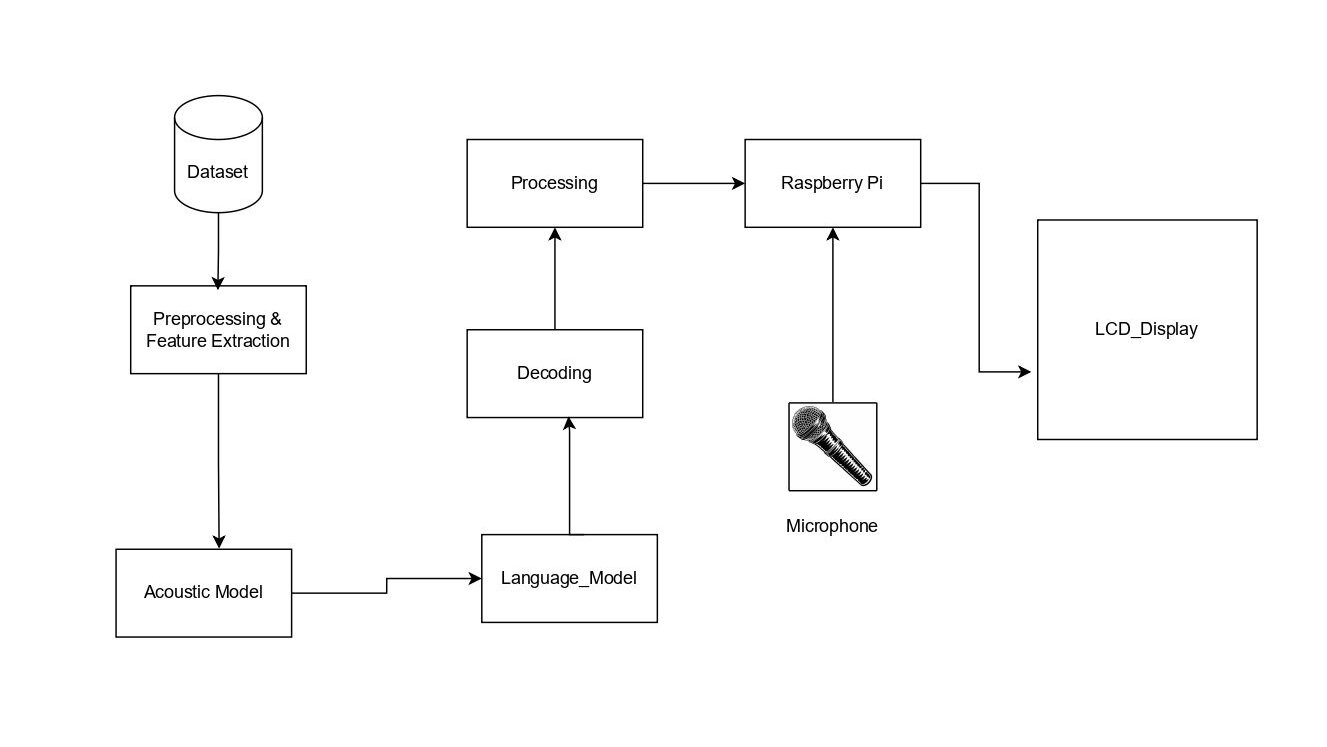
Basanta joshi, Bharat Bhatta, Sanjeeb Prasad panday, Ram Krishna Maharjan, A Novel Deep Learning Based Nepali Speech Recognition [2] (2022), Proposes a Nepali speech recognition model using CNN-GRU with data from Librispeech. MFCC extracts features, CNN-GRU develops the acoustic model, and CTC decodes. Performance is evaluated using Word Error Rate.

Manish Dhakal, Arman Chhetri, Aman Kumar Gupta, Prabin Lamichhane, Suraj Pandey, Subarna Shakya, Automatic speech recognition for the Nepali language using CNN, bidirectional LSTM and ResNet [3 ] (2022), Introduces an end-to-end deep learning ASR model for transcribing Nepali speech to text, trained on the OpenSLR dataset. After preprocessing to remove silent gaps, MFCCs are used as audio features. The best performance, with a character error rate of 17.06%, is achieved using a Bidirectional LSTM paired with ResNet and a one-dimensional CNN, employing CTC for loss calculation and beam search decoding.

Paribesh Regmi, Arjun Dahal, Basanta Joshi, Nepali Speech Recognition using RNN-CTC Model [4], (2019), introduces a Nepali Speech Recognition model using RNNs and CTC. The RNN processes sequential audio data, and CTC maximizes the probability of desired labels from RNN output. The model outputs Nepali text and uses a character set of 67 Nepali characters for transcription.

**Methodology**

**Block Diagram:**

****

Working Principle:

Translating Nepali speech into Nepali text involves several stages, integrating digital signal processing, natural language processing, and machine learning. The process starts with capturing the audio input via a microphone. This audio undergoes pre-processing, including noise reduction, volume normalization, and segmentation to ensure clarity and manageability. Feature extraction follows, converting the audio signals into formats like Mel-Frequency Cepstral Coefficients (MFCC) or spectrograms, which represent the sound's short-term power spectrum.

The core of the translation lies in the acoustic and language models. The acoustic model, often utilizing Hidden Markov Models (HMM) or Deep Neural Networks (DNN), translates audio features into phonetic units. The language model then ensures these units form coherent text by predicting word sequences, aiding in word disambiguation and maintaining grammatical and contextual integrity

Decoding combines the acoustic and language models to produce the most probable word sequence from the audio input, using algorithms like Viterbi or beam search. Post-processing refines this text, correcting spelling errors, adding punctuation, and ensuring proper formatting. Finally, the text is outputted, displayed on a screen, saved to a file, or integrated into another application. This entire mechanism leverages technologies such as DSP for audio processing, machine learning for model training, and NLP for language handling, supported by tools like TensorFlow, PyTorch, and Kaldi, to achieve accurate and efficient speech-to-text translation in Nepali.

**System Design**

**Requirement Specification**

1. Functional Requirements:

1.1 Speech Recognition:

The system must accurately transcribe spoken Nepali words and phrases into text.

It should support a wide range of Nepali accents, dialects, and speaking styles.

1.2 Real-time Processing:

The system must process audio input in real-time, providing immediate transcription feedback.

It should minimize latency to ensure smooth user interaction.

1.3 Hardware Integration:

The system must interface seamlessly with Raspberry Pi hardware components, including the microphone and LCD module.

It should utilize GPIO pins for hardware control and data transfer.

1.4 User Interface:

The system should have a user-friendly interface for initiating speech recognition, displaying transcribed text, and providing feedback.

It should incorporate error handling mechanisms to guide users in case of input errors or system failures.

2. Non-functional Requirements:

2.1 Accuracy:

The system should achieve high accuracy in transcribing Nepali speech, with minimal errors.

It should continuously improve accuracy through model optimization and training.

2.2 Performance:

The system should be optimized for efficient resource utilization, minimizing CPU and memory usage.

It should handle multiple concurrent requests without significant degradation in performance.

2.3 Portability:

The system should be portable and lightweight, capable of running on different Raspberry Pi models and configurations.

It should be easy to deploy and maintain in diverse environments.

2.4 Scalability:

The system should be scalable, capable of handling increasing volumes of speech data and user interactions.

It should support modular design principles to facilitate future enhancements and updates.

2.5 Security:

The system should prioritize user privacy and data security, adhering to best practices for data encryption and access control.

It should implement safeguards against unauthorized access and malicious attacks.

**Feasibility Analysis**

Feasibility analysis is a critical step in assessing the viability and potential success of a project. For the Nepali Speech-to-Text Recognition System, feasibility can be evaluated across various dimensions:

1. Technical Feasibility:

Data Availability: Availability of high-quality Nepali speech datasets for model training is crucial. If sufficient datasets exist or can be collected, technical feasibility is enhanced.

Hardware and Software Requirements: Raspberry Pi hardware and required software components must meet system requirements. Compatibility issues and resource constraints could affect feasibility.

Algorithm Complexity: Deep Learning algorithms for speech recognition should be implementable within the computational limitations of Raspberry Pi. Complex algorithms may require optimization or alternative approaches.

2. Economic Feasibility:

Cost Analysis: Costs associated with hardware components, software licenses, and development resources should be evaluated against the project budget. If costs exceed budget constraints, feasibility may be compromised.

Return on Investment (ROI): Potential benefits, such as increased accessibility for Nepali speakers and technological advancement, should outweigh project costs. A positive ROI indicates economic feasibility.

3. Operational Feasibility:

User Acceptance: The system's usability and effectiveness must meet user expectations. User feedback and testing can assess the system's operational feasibility.

Integration with Existing Systems: Integration with other systems or applications may be necessary. Compatibility and ease of integration should be evaluated to determine operational feasibility.

4. Legal and Regulatory Feasibility:

Compliance: The project must comply with legal and regulatory requirements related to data privacy, intellectual property rights, and speech recognition technology. Non-compliance could pose legal risks and affect feasibility.

Ethical Considerations: Ethical implications, such as ensuring user privacy and preventing bias in speech recognition algorithms, should be addressed to maintain feasibility.

Based on the feasibility analysis, if technical challenges can be overcome within budget constraints, economic benefits justify costs, operational requirements are achievable, and legal and regulatory compliance is ensured, the project is deemed feasible. Conversely, if significant barriers exist in any of these areas, further evaluation and mitigation strategies are necessary to enhance feasibility and project success.

**Tools and Techniques**

**Hardware**

1. Raspberry Pi

The Raspberry Pi microcomputer serves as the central hardware component, providing computational power and interfacing capabilities. It integrates seamlessly with other hardware modules, facilitating data processing and control tasks.



Fig: Raspberry Pi

1. Microphone Module:

The microphone module captures audio input from the environment, converting sound waves into digital signals. It consists of a microphone sensor, amplifier circuitry, and connectors for interfacing with the Raspberry Pi, enabling real-time audio capture.



Fig: Microphone module

1. LCD module

The LCD module serves as the output device, displaying transcribed text generated by the speech recognition system. It utilizes liquid crystal technology to produce clear and readable text output, enhancing user interaction and feedback.



Fig: LCD module

**Software:**

1. Raspberry Pi OS

Raspberry Pi OS, a lightweight Linux distribution, serves as the operating system, providing essential drivers and utilities for system management and software execution. It ensures compatibility with Raspberry Pi hardware and supports Python scripting.

1. Python:

Python, a versatile programming language, facilitates the development of software components. Python scripts control hardware interfaces, implement speech recognition algorithms, and manage system behavior, enabling seamless integration and functionality.

1. TensorFlow:

TensorFlow, a Deep Learning framework, powers the speech recognition model. It enables the creation and training of neural network architectures capable of accurately transcribing Nepali speech into text. Additionally, libraries such as pandas, numpy, and matplotlib complement Python's capabilities, facilitating data manipulation, numerical computation, and visualization for system optimization and analysis.

**Epilogue**

In the final stages of planning for the Nepali Speech-to-Text Recognition System, considerations for expected outputs, work scheduling, and budget estimations play pivotal roles in project completion and success.

1. **Expected Output:**

Upon completion, the Nepali Speech-to-Text Recognition System is expected to deliver:

Accurate transcription of spoken Nepali words and phrases into text in real-time.

Seamless integration with Raspberry Pi hardware components, including microphone and LCD module.

User-friendly interface for initiating speech recognition and displaying transcribed text.

Documentation detailing system architecture, installation instructions, and usage guidelines.

Evaluation report outlining system performance metrics and user feedback.

**2. Work Schedule:**

The project will follow a structured work schedule, divided into phases:

**3. Budget Estimation:**

A budget estimation for the project include:

|  |  |  |
| --- | --- | --- |
| **S.N** | **Hardware Name** | **Price** |
| **1.** | **Raspberry Pi** | **Rs.14000** |
| **2.** | **Microphone** | **Rs.250** |

**Conclusion**

The proposal for a Nepali Speech-to-Text Recognition System marks a crucial step in overcoming linguistic barriers. By integrating Deep Learning on Raspberry Pi, the project aims to empower Nepali speakers with real-time speech transcription. With a focus on accessibility and inclusivity, the system promises accurate results, fostering technological empowerment within the community. Through meticulous planning and execution, this endeavor seeks to democratize access to advanced speech recognition technology, enhancing communication and participation in the digital sphere. Ultimately, the project envisions a future where language is no longer a barrier to technological advancement, promoting socio-economic growth and cultural preservation among Nepali speakers.

**References**

1. https://www.researchgate.net/publication/375139583\_Nepali\_Text-to-Speech\_Synthesis\_using\_Tacotron2\_for\_Melspectrogram\_Generation
2. <https://www.researchgate.net/publication/359950525_A_Novel_Deep_Learning_Based_Nepali_Speech_Recognition>
3. <https://www.researchgate.net/publication/362745838_Automatic_speech_recognition_for_the_Nepali_language_using_CNN_bidirectional_LSTM_and_ResNet>
4. https://www.researchgate.net/publication/334523238\_Nepali\_Speech\_Recognition\_using\_RNN-CTC\_Model
5. [Automatic Messaging System for Vehicle Tracking and Accident Detection | Request PDF (researchgate.net)](https://www.researchgate.net/publication/343438742_Automatic_Messaging_System_for_Vehicle_Tracking_and_Accident_Detection)
6. [(PDF) Epidemiology of Road Traffic Accidents in Nepal from 2009/10 to 2019/20: A 10 Year Study (researchgate.net)](https://www.researchgate.net/publication/354438644_Epidemiology_of_Road_Traffic_Accidents_in_Nepal_from_200910_to_201920_A_10_Year_Study)
7. [Road Safety in Nepal | Traffic accidents, crash, fatalities & injury statistics | GRSF (roadsafetyfacility.org)](https://www.roadsafetyfacility.org/country/nepal)

14

14

13