SAHYADRI
COLLEGE OF ENGINEERING & MANAGEMENT
An Autonomous Institution
MANGALURU

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
SYNOPSIS

1.	**Title of the Project**	**Voice Based Banking for Rural Areas**		
2.	**Group Number**	52		
3.	**Name of the Students and Guide**	Sushrutha Shanbhogue	4SF22CS225	
		Raghavendra SS	4SF22CS158	
		Sowndarya S	4SF22CS218	
		Iram A.K Shaikh	4SF23CS404	
		Mr. Raghavendra Sooda	Assistant Professor	

ABSTRACT

Accessing and utilizing banking services remains a significant challenge for rural populations, primarily due to language barriers, low digital literacy, and limited exposure to formal financial systems. This project, titled "Voice-Based Banking for Rural Areas" addresses these issues by developing an intelligent, voice-enabled chatbot that allows users to interact with banking systems using regional languages like Kannada and even unwritten languages such as Tulu and Konkani.

The system starts by taking the user's details, such as name, address, etc. as voice input in a regional language like Kannada. This voice input is recognized and processed using the OpenAI Whisper model, a multilingual speech recognition model trained on large-scale weakly labeled audio data. Whisper efficiently transcribes and translates the Kannada (or any regional or unwritten language) speech to English text. If the request involves filling out a banking form for account creation or availing loan, the chatbot extracts relevant information and automatically generates a filled-out text document. This ensures that the rural people feel encouraged to visit banks and follow banking procedures, making them more financially secure and educated.

INTRODUCTION

In recent years, India has witnessed a rapid expansion in digital banking services, contributing significantly to financial inclusion and economic development. However, despite these advancements, a significant segment of the population, especially in rural and semi-rural areas continues to face substantial barriers in accessing and utilizing banking services. The major challenges include low literacy levels, lack of digital fluency, and the dominance of regional languages like Kannada or unwritten languages like Tulu or Konkani, which are often not supported in conventional banking applications.

Traditional banking systems primarily rely on text-based interfaces and English-language communication, which can be intimidating and inaccessible for rural users. As a result, many individuals remain excluded from the formal banking sector, unable to independently perform basic financial transactions such as checking account balances, transferring funds, or filling out forms. To address these critical gaps, this project proposes an innovative, voice-based solution that leverages recent advancements in artificial intelligence and natural language processing to create a user-friendly banking interface tailored specifically for rural users.

The core objective of this project is to develop a voice-enabled banking chatbot that understands and processes user queries spoken in regional languages. The system utilizes OpenAI's Whisper model, a powerful speech recognition and translation framework, to transcribe the spoken language, such as a Kannada input into English. This translated input is then processed to identify the user's intent, enabling the system to perform the corresponding banking operation.

In addition to facilitating conversational interaction, the system automatically generates required banking documents, such as transaction records or application forms, with the relevant user information. Hence, the final output includes a text document ready for submission or printing.

By combining speech recognition, machine translation, chatbot interaction, and document automation, this project aims to build a practical, scalable, and accessible voice-based banking solution. It is particularly designed to empower rural communities by eliminating language and literacy barriers, enhancing trust in digital systems, and enabling independent financial operations. Ultimately, this solution has the potential to contribute meaningfully to digital financial literacy and rural empowerment.

LITERATURE SURVEY

Sl. No.	Paper Title & Year of Publication	Author/s	Key Findings	Methodologies Used
1	Robust Speech Recognition via Large-Scale Weak Supervision, *2022*	Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, Ilya Sutskever	Whisper delivers strong zero-shot performance, surpassing supervised and commercial ASR systems in robustness and long-form transcription. Scaling data and model size boosts multilingual accuracy, approaching human-level results.	Whisper is a Transformer-based speech model trained on 680k hours of multilingual, multitask audio-text data, using a unified token-based format and minimal preprocessing for tasks like transcription and translation.
2	UniSpeech-SAT: Universal Speech Representation Learning with Speaker-Aware Pre-training, 2022	Chen et al.	Improved speaker recognition and personalization in speech systems; useful for authentication in banking.	Speaker-aware pre-training; Fine-tuning for ASR and speaker verification; Self-supervised representation learning.
3	Maestro: Matched Speech Text Representations Through Modality Matching, 2022	Chen et al.	Combines speech and text learning to enhance ASR performance; supports seamless speech-to-text conversion in banking.	Joint training on aligned speech-text pairs; Modality-matching architecture; Fine-tuning for downstream tasks.
4	Unsupervised Speech Recognition, *2021*	Baevski et al.	Eliminated need for labeled audio data; supports voice apps in rural settings lacking transcription data.	Pretraining with wav2vec 2.0; Fine-tuning with small transcribed sets; Unsupervised decoding.
5	SpeechStew: Simply Mix All	Chan et al.	Demonstrated large-scale data aggregation improves	Dataset mixing; Transformer-based ASR

	Available Speech Recognition Data to Train One Large Neural Network, *2021*		ASR across domains; applicable for mixed-accent rural voices.	model; Joint training on multiple corpora; Robust to noise and accent variance.
6	VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation, *2021*	Wang, C. et al.	Introduced a large multilingual dataset (~400K hours, 23 languages) ideal for training robust voice systems in low-resource and rural language settings.	Self-supervised learning (SSL); Data collection from EU Parliament; Pretraining for ASR and multilingual tasks.
7	XLS-R: Self-Supervised Cross-Lingual Speech Representation Learning at Scale, *2021*	Babu et al.	Enabled multilingual speech recognition, useful for diverse rural languages; improved low-resource language support.	Cross-lingual self-supervised learning; pre-trained on 436 languages; Fine-tuning for specific domains.
8	Self-training and Pre-training are Complementary for Speech Recognition, 2021	Xu, Q. et al.	Demonstrated that combining SSL pretraining and pseudo-label self-training significantly improves performance, especially for low-resource domains.	wav2vec 2.0 + pseudo-labeling; Semi-supervised pipeline; ICASSP evaluation on LibriSpeech and CommonVoice.
9	Multitask Training with Text Data for End-to-End Speech Recognition, 2020	Wang, P. et al.	Showed that incorporating text-only data improves ASR performance, enabling better generalization where speech data is scarce.	Multitask training; CTC/attention-based hybrid models; Text-based language modeling integration.
10	CHiME-6 Challenge: Tackling Multi-	Watanabe, S. et al.	Addressed challenges of real-world, noisy, multi-speaker environments—	Distant microphone ASR; Speech diarization; End-to-end neural diarization

	Speaker Speech Recognition for Unsegmented Recordings, 2020		critical for rural deployment in public spaces.	(EEND); Realistic dataset collection.
11	SpecAugment: A Simple Data Augmentation Method for ASR, 2019	Park et al.	Improved ASR model robustness against background noise, crucial for rural environments.	Time warping, frequency/time masking; Integrated into RNN-based ASR models; Improved generalization.
12	Toward Domain-Invariant Speech Recognition via Large Scale Training, *2018*	Narayanan et al.	Large-scale training achieves domain generalization; helps adapt banking systems to rural usage contexts.	Domain-invariant training; Noise-robust feature learning; Evaluation across multiple acoustic domains.
13	Attention Is All You Need, *2017*	Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, Illia Polosukhin	The Transformer achieves state-of-the-art translation with reduced training time, generalizes to other tasks, and scales efficiently thanks to parallelism and shorter path lengths.	The paper proposes the Transformer model, which uses only self-attention mechanisms for efficient and parallel sequence transduction.
14	Librispeech: An ASR Corpus Based on Public Domain Audio Books, *2015*	Panayotov et al.	Provided a widely-used dataset for training and evaluating speech models; baseline for voice interfaces.	Data collection from audiobooks; Preprocessing pipeline; Used for ASR benchmarking.

PROBLEM STATEMENT AND DESCRIPTION

Despite India's rapid strides in digital banking and financial technologies over the past decade, a substantial portion of the rural population continues to face significant obstacles in accessing and utilizing formal banking services. This is mainly due to the language barrier which exists between many rural and urban populations, where all banking procedures are designed for English or Hindi speaking customers. This puts them at a financial disadvantage and leaves them largely unaware of bank rules and procedures.

In states like Karnataka, most rural inhabitants speak and understand only Kannada, a regional language that is often not supported effectively in mainstream banking interfaces, which mainly support English. This presents a major barrier for users who cannot read, write, or comprehend these languages. Additionally, many rural users are unfamiliar with standard banking procedures, and form-filling protocols. As a result, even basic banking operations such as opening an account and submitting an application to avail loan become intimidating and difficult.

While some voice-based systems have been introduced in the past to simplify customer interaction, these solutions are generally limited in scope. Existing voice assistants are often designed to respond to rigid, predefined commands and are rarely capable of understanding regional languages like Kannada or unwritten languages such as Tulu and Konkani. As a result, they fall short of meeting the practical, real-world needs of rural users who require end-to-end assistance in a language and mode of communication they are comfortable with.

In response to this pressing gap, the proposed project aims to develop a comprehensive, voice-based banking assistant designed for users in rural areas. This system will use OpenAI's Whisper model, a speech recognition and translation technology capable of accurately transcribing spoken language (such as Kannada) and translating it into English in real time. Once translated, the input will be interpreted based on the user's intent, such as account creation or availing loan and generates the corresponding document or form accordingly. The user's interaction will be entirely voice-driven, eliminating the need for reading, writing, or digital literacy.

By enabling rural users to speak naturally in their regional language, this solution addresses both linguistic and technical barriers. It encourages rural users to create bank accounts or avail loans without depending on any third parties, and fosters trust in digital financial systems, simplifying banking for rural communities.

OBJECTIVES

- To design a user-friendly interface that accommodates users with low or no digital literacy, minimizing the need for typing, reading, or understanding English
- To develop a voice-based chatbot system that allows rural users to interact with banking services using natural speech in a regional language like Kannada or unwritten languages like Tulu and Konkani.
- To integrate OpenAI's Whisper model for accurate speech recognition and real-time translation of spoken Kannada into English for further processing

PROPOSED METHODOLOGY

- The system starts by taking a spoken query in a regional language like Kannada from the user through a chatbot, which acts as an interface between the user and the system. This voice input is processed using the OpenAI Whisper model, a multilingual speech recognition model trained on large-scale weakly labeled audio data.
- Using a sequence-to-sequence Transformer-based architecture, Whisper not only transcribes the spoken Kannada into text but also translates it into English. Once the English text is generated, based on the user's request, which could be either opening an account or filling a loan application form, the chatbot extracts relevant information and automatically generates a filled-out text document.
- The document is formatted according to standard banking templates and filled with the extracted details, significantly reducing the time and effort required by the user. This document can be printed or submitted at the bank, greatly simplifying banking access for non-English-speaking rural users.

Step 1: Voice Input from User - Kannada

The user opens the application and speaks into the microphone in Kannada

Step 2: Speech Recognition using OpenAI Whisper

- Open AI Whisper is trained on 680k+ hours, supports low-resource languages and noisy environments [1]
- Whisper model takes Kannada speech and converts it to Kannada text
- Handles accent variation, background noise, and mixed language speech, which are important parameters for rural scenarios

Step 3: Language Translation using Transformer

The Kannada text is passed to a Transformer-based translation model:

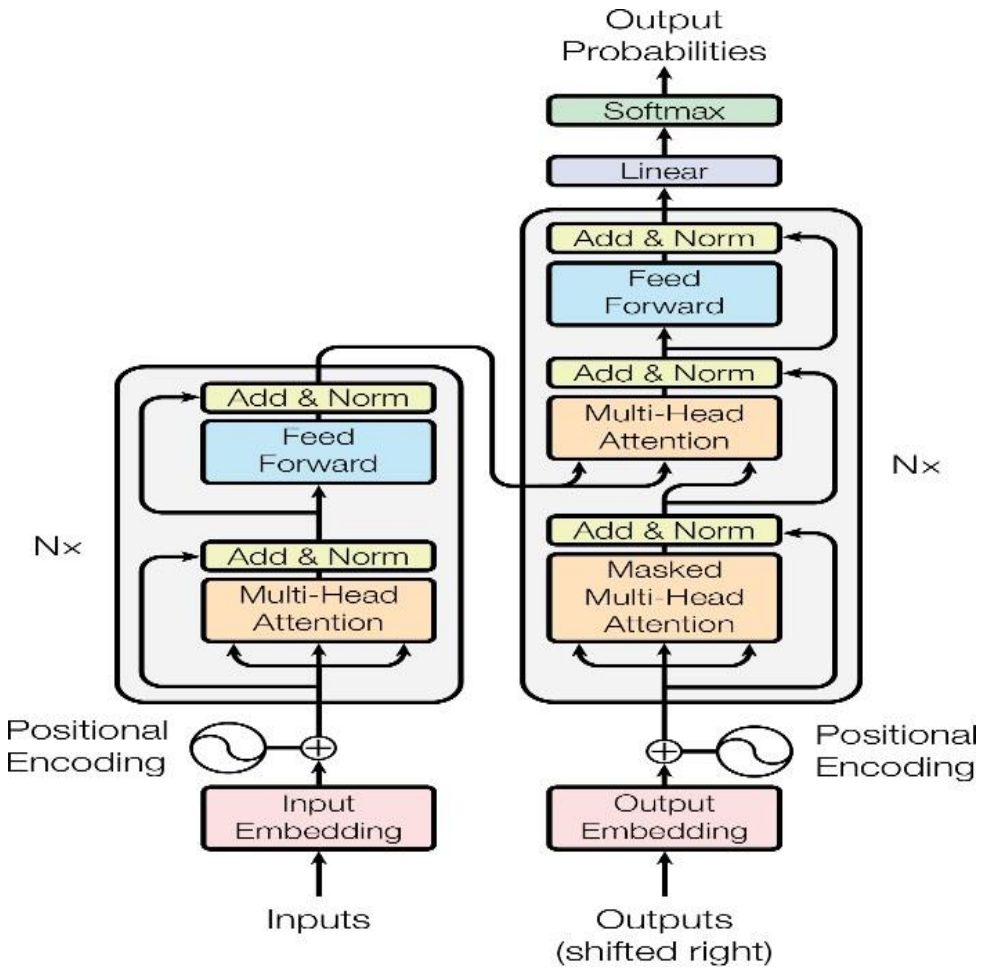


Figure 1: Transformer model architecture

Components of the Transformer model architecture:

1. Input Embedding

- Input Embedding maps each token ID to a dense vector using a learnable embedding matrix.
- It captures semantic meaning and converts discrete tokens into continuous representations.
- These embeddings are combined with positional encodings and passed to the Transformer layers.

2. Positional Encoding

- Positional Encoding adds information about token positions to the input embeddings.
- Since Transformers lack recurrence, this helps the model understand the order of tokens.
- It uses fixed or learnable vectors added to embeddings before passing them to the layers.

3. Multi-Head Attention

- Positional Encoding adds information about token positions to the input embeddings.

- Since Transformers lack recurrence, this helps the model understand the order of tokens.
- It uses fixed or learnable vectors added to embeddings before passing them to the layers.

4. Add & Norm

- Add & Norm applies residual connections by adding the input to the output of a sub-layer such as feed forward.
- This helps preserve input information and improves gradient flow during training.
- Layer normalization is then applied to stabilize and speed up the training process.

5. Feed Forward

- The Feed Forward layer applies two linear transformations with a ReLU activation in between.
- It processes each position independently to add non-linearity and enhance feature representation.
- This helps the model learn complex patterns beyond what attention alone can capture.

6. Output Embedding

- Output Embedding maps the decoder's output vectors to vocabulary logits using a linear layer.
- It converts high-dimensional features into a probability distribution over the target vocabulary.
- This allows the model to predict the next token during sequence generation.

7. Linear Layer

- The Linear layer applies a learned weight matrix to transform input vectors into a new feature space.
- It is used to project outputs, such as attention results or hidden states, into desired dimensions.
- This transformation is essential for tasks like generating logits over the vocabulary.

8. SoftMax Layer

- The SoftMax function converts raw logits into a probability distribution over the target classes.
- In Transformers, it is used in the attention mechanism and final output layer.
- It ensures the model assigns probabilities that sum to 1, enabling token prediction or attention weighting.

Our project achieves Kannada to English translation using sequence-to-sequence learning:

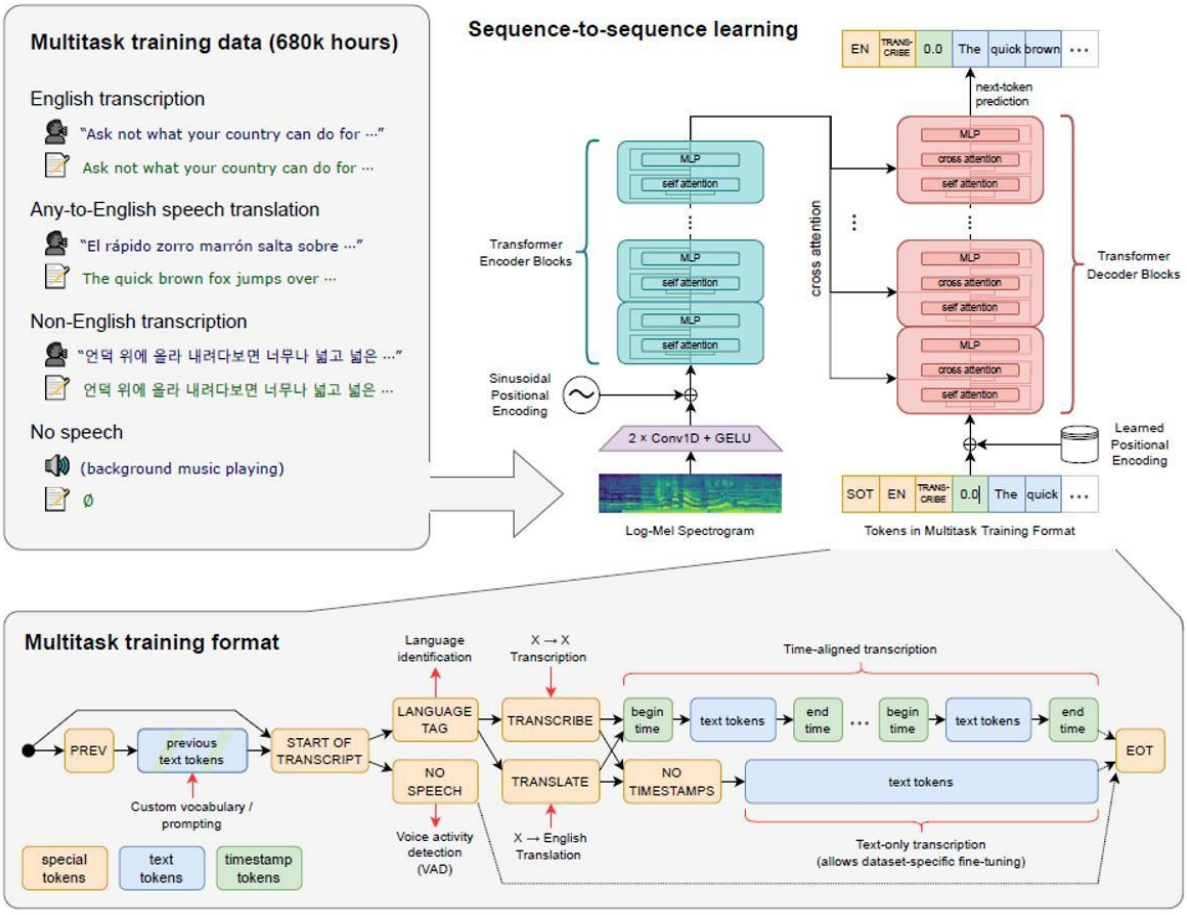


Figure 2: sequence -to-sequence Transformer model architecture

Components of the sequence – to – sequence Transformer model architecture:

1. **Log-Mel Spectrogram**

- A Log-Mel Spectrogram converts audio signals into a time-frequency representation using the Mel scale and logarithmic amplitude.
- In sequence-to-sequence Transformers, it serves as the input feature for the encoder.
- This representation captures important speech characteristics in a format suitable for learning.

2. **2 x Conv1D + GELU**

- The 2 x Conv1D + GELU block applies two 1D convolution layers with GELU activation in between.
- It captures local temporal patterns and non-linear features from the input sequence, such as audio or embeddings.

- This helps enhance feature extraction before feeding into the Transformer encoder.

3. Sinusoidal Positioning

- Sinusoidal Positioning encodes token positions using sine and cosine functions of different frequencies.
- It provides fixed positional information to input embeddings, enabling the model to understand token order.
- This is crucial in sequence-to-sequence Transformers, which lack built-in recurrence or convolution.

4. Self – Attention

- Self-Attention allows each token in a sequence to weigh and attend to other tokens using query, key, and value projections.
- It helps the model capture contextual relationships regardless of token position.
- This mechanism enables better understanding of sequence dependencies in both encoder and decoder.

5. Multi – Layer Perceptron

- The applies two linear layers with a non-linear activation in between.
- It processes each token's representation independently to enhance feature transformation.
- This helps the model learn complex patterns beyond attention mechanisms.

6. Cross – Attention

- Cross-Attention allows the decoder to attend to encoder outputs using queries from the decoder and values from the encoder.
- It helps the model align and extract relevant information from the input sequence.
- This mechanism is essential for generating context-aware outputs in sequence-to-sequence tasks.

7. Learned – Positional Encoding

- Learned Positional Encoding uses trainable embeddings to represent token positions in a sequence.
- These embeddings are added to input embeddings to provide the model with positional context.
- Unlike fixed sinusoidal encoding, they are optimized during training for the specific task.

Step 4: Translated Voice Reply

- The English translation is converted into speech using a Transformer based Text-to-Speech (TTS) engine
- The system speaks out the English reply to confirm understanding

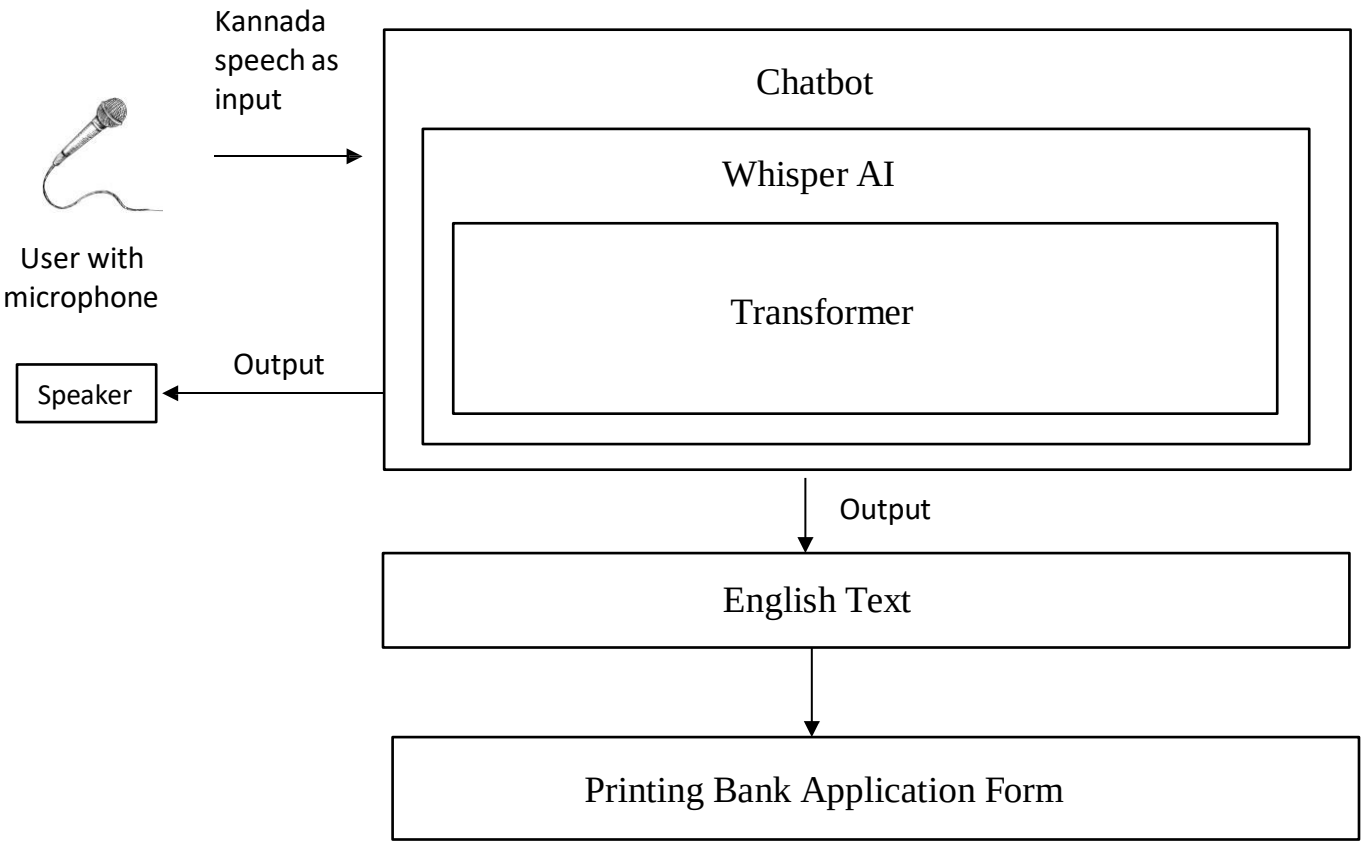
Step 5: Document Generation

If the query involves a task like depositing or withdrawing money:

- The system extracts relevant info (name, amount, account number, etc.).
- Fills a bank form template (e.g., withdrawal slip, application form).

Output: Text or PDF document using tools like FPDF or docx

Architectural Diagram



EXPECTED OUTCOME OF THE WORK

The expected outcome of this project is a complete voice-based banking assistant designed specifically to help rural users interact with banking systems using their native language, such as Kannada or other unwritten regional languages such as Tulu and Konkani. The system will allow users to speak their banking-related queries or requests, such as opening a bank account or availing loan.

Our project aims to bridge the gap between rural and urban bank customers, encouraging the rural users to visit banks, open accounts and be more financially secure and educated.

WORK PLAN

The development of the voice-based banking assistant will be carried out in the following structured phases:

Phase 1: Basic Chatbot Creation

- Create a chatbot which can recognise speech and respond effectively

Phase 2: Model translating regional language to English

- Find a Whisper model which can translate regional language like Kannada, to English
- Try to enhance the model to ensure efficient translation.

Phase 3: Model Integration and Development

- Integrate the chatbot with OpenAI Whisper for transcription and translation.

Phase 4: User Interface Development

- Design a simple, intuitive voice-first interface suitable for mobile or low-end devices.

Phase 5: Documentation Generation

- Develop a software which can write the translated text into a word document for filling the required bank application

GANTT CHART

Task	Week 1	Week 2	Week 3	Week 4	Week 5	Week 6	Week 7	Week 8	Week 9	Week 10	Week 11	Week 12
Research and Data Collection												
Data Preprocessing & Feature Eng.												
Model Development & Training												
Model Evaluation & Selection												
Web Application Development												
Testing and Deployment												

CONCLUSION

This project successfully addresses a critical gap in financial accessibility for rural users by developing a voice-based banking assistant tailored for regional language speakers. Traditional banking interfaces often do not consider language barriers which exclude a large segment of the rural population from independently managing their financial needs. By integrating OpenAI's Whisper model for robust speech recognition and translation, along with a natural language processing-based chatbot and automated document generation, the system enables users to interact with banking services entirely through voice in their native language.

Hence, this project enables rural users to speak naturally in their regional language, addressing both linguistic and technical barriers. It encourages rural users to create bank accounts or avail loans without depending on any third parties, and fosters trust in digital financial systems, simplifying banking for rural communities.

REFERENCES

[1] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust Speech Recognition via Large-Scale Weak Supervision," *OpenAI*, 2022.

[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is All You Need," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[3] A. Babu et al., "XLS-R: Self-Supervised Cross-Lingual Speech Representation Learning at Scale," *arXiv preprint arXiv:2111.09296*, 2021.

[4] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "Unsupervised Speech Recognition," *arXiv preprint arXiv:2105.11084*, 2021.

[5] W. Chan, C. J. Wang, and N. Jaitly, "SpeechStew: Simply Mix All Available Speech Recognition Data to Train One Large Neural Network," *arXiv preprint arXiv:2104.02133*, 2021.

[6] D. S. Park et al., "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition," in *Proc. Interspeech*, 2019.

[7] S. Chen et al., "UniSpeech-SAT: Universal Speech Representation Learning with Speaker-Aware Pre-Training," *arXiv preprint arXiv:2110.05752*, 2022.

[8] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. ICASSP*, 2015.

[9] A. Narayanan et al., "Toward Domain-Invariant Speech Recognition via Large Scale Training," in *Proc. Interspeech*, 2018.

[10] S. Chen et al., "Maestro: Matched Speech Text Representations through Modality Matching," *arXiv preprint arXiv:2210.06674*, 2022.

[11] C. Wang et al., "VoxPopuli: A Large-Scale Multilingual Speech Corpus for Representation Learning, Semi-Supervised Learning and Interpretation," *arXiv preprint arXiv:2101.00390*, 2021.

[12] P. Wang, Y. Zhang, Y. Wu, and Z. Yang, "Multitask Training with Text Data for End-to-End Speech Recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2391–2403, 2020.

[13] S. Watanabe et al., "CHiME-6 Challenge: Tackling Multi-Speaker Speech Recognition for Unsegmented Recordings," in *Proc. CHiME Workshop*, 2020.

[14] Q. Xu et al., "Self-Training and Pre-Training Are Complementary for Speech Recognition," in *Proc. ICASSP*, 2021.