International Journal of Leading Research Publication (IJLRP)

E-ISSN: 2582-8010 ● Website: www.ijlrp.com ● Email: editor@ijlrp.com

Anvi-AI Powered Conversational Agent for Emotion-Aware and Memory Rich Interactions in Telugu

Dr. SriSudha Garugu[1], B. Naveen[2], B. Prasanna[3], B. Poojitha[4]

[1]CSE, Associate Professor at ACE Engineering College, Hyderabad, India.
[2,3,4]CSE, Student at ACE Engineering College, Hyderabad, India.

Abstract:
The rapid growth of mobile AI assistants has made it much easier for people to interact with computers. However, most current systems still have trouble adapting to different cultures, supporting regional languages, understanding context, and keeping privacy. This survey paper looks at the latest developments in smart mobile assistants, with a focus on speech recognition technologies, natural language understanding, emotion and context detection models, on-device AI frameworks, and systems that help with mental health. We look closely at popular technologies like Google Assistant, Siri, Alexa, Bixby, and recent research on transformer-based speech models, multimodal emotion recognition, and privacy-focused edge AI. The review points out some big problems with current solutions, such as poor performance in low-resource languages like Telugu, no personalized emotional intelligence, reliance on cloud processing, and a limited ability to understand the user's context or mental state. The survey also talks about problems like not having enough datasets, AI responses that aren't culturally sensitive, worries about data privacy, and mobile devices that don't learn on their own all the time. This survey highlights the necessity for a culturally adaptive, emotion-sensitive, and privacy-centric mobile assistant, exemplified by the proposed system, Anvi, which incorporates Telugu speech processing, contextual awareness, and on-device AI to enhance user well-being and provide personalized assistance.

Keywords: Mobile AI Assistants; Speech Recognition; Natural Language Understanding (NLU); Telugu Language Processing; Low-Resource Languages; Emotion Recognition; Context Awareness; On-Device AI; Edge AI; Privacy-Preserving AI; Transformer-Based Models; Multimodal Learning; Mental Wellbeing Support; Human–Computer Interaction (HCI).

I. INTRODUCTION

1.1 Background

Mobile AI assistants have come a long way. They used to be just voice-command tools, but now they are smart, aware of their surroundings, and can understand not only what we say but also how we say it. Early studies demonstrated that successful human–machine interaction necessitates the recognition of emotional indicators such as tone, pitch, and rhythm, in addition to words. The accuracy of speech recognition has greatly improved thanks to better audio processing methods like MFCCs and other speech features, as well as powerful neural models like encoder-decoder systems and attention mechanisms. This progress means that modern assistants can now reliably do things like recognize emotions and languages and dialects, even in real-world situations. All of these changes have made mobile assistants more natural, responsive, and able to change to fit users in a way that is more like a person and more knowledgeable.

1.2 Importance of Mobile AI Assistants

Mobile AI assistants that run on personal devices offer personalized services like reminders, accessibility, and mental health nudges, and they give immediate feedback that takes the user's situation into account.

Because they are so common on devices that are limited in space and privacy, on-device robustness and efficiency are very important. Speech and emotion models need to be small and able to handle speaker and dialect differences [5]. Emotion-aware assistants make the user experience better by changing the way they phrase things, when they say them, and what they suggest based on how they think you feel. They also make interactions less frustrating when mistakes happen and support interventions for mental health[1].Also, assistants that can speak more than one language and dialect help people who don't speak English get the services they need, making them fair across all language groups [16].Recent progress in Natural Language Processing (NLP) and Artificial Intelligence (AI) has played a big role in making smart conversational agents, especially for languages, like Telugu.

1.3 Need for Regional Language & Emotion-Aware Systems

Automatic speech recognition (ASR) and translation systems still have trouble with dialects and regional languages. This is mostly because there isn't enough training data, accents vary widely, and words are pronounced and structured differently. However, studies show that a single well-designed model, like a sequence-to-sequence or grapheme-based hierarchical model, can handle many accents well if it is trained on a variety of datasets or has clear accent-related information. For Indian languages like Telugu, efforts such as building dedicated speech corpora (like CSTD-Telugu) and studying accent-specific features have shown that creating targeted datasets is both possible and necessary. These studies highlight how important it is to design models that understand the unique characteristics of regional speech. Emotion recognition in speech adds another layer of complexity. Traditional supervised learning approaches often struggle with background noise, recording quality, and differences between speakers. Because of this, researchers focus on developing robust features—such as prosodic (intonation, rhythm) and spectral (frequency-based) features—and classifiers that can adapt to different conditions. To build truly effective conversational systems, it's important to combine dialect-aware speech recognition with emotion-sensitive analysis. This allows systems not only to understand what is being said, but also how it is being said—especially in diverse regional contexts. In this direction, Garugu and Bhaskari [15] proposed a two-stage deep learning model for Telugu abstractive summarization. Their work addresses the challenges of low-resource languages and improves the quality of generated summaries, making it particularly valuable for conversational AI systems that need to generate meaningful responses in regional languages.

1.4 Scope of the Survey

This survey focuses on the intersection of mobile AI assistants, speech recognition for regional languages (with emphasis on Telugu), and speech-based emotion detection. It surveys: (1) acoustic feature extraction and robust audio representations for mobile deployment [3]; (2) approaches to multi-dialect/multi-accent ASR and adaptation strategies [8][9][13]; (3) supervised and robust methods for emotion classification in speech, including limitations identified by comparative evaluations [5]; (4) datasets and corpus-building efforts for Indian languages and Telugu specifically [14][16]; and (5) design considerations for on-device, privacy-preserving assistants that integrate affective signals into interaction policies [1].We intentionally limit full survey coverage of text-only NLP for Indian languages (beyond language models used for ASR/SLU) and of non-speech modalities (e.g., facial affect) except where they directly inform speech-based emotion inference. Garugu et al. [4] introduced a Telugu Question Answering (QA) system using XLM-RoBERTa, which effectively handles data scarcity through multilingual representation learning. This work directly contributes to the development of conversational agents by enabling accurate query understanding and response generation**.**

1.5 Paper Organization

To guide the reader, the survey is organized as follows: Section 1 provides motivation and related work in affective human–machine interaction [1].and foundations of audio analysis [3]. Section 2 reviews acoustic features and robust speaker/dialect modelling [6][8]. Section 3 surveys emotion recognition

approaches, their robustness challenges, and evaluation studies [3]**.** Section 4 examines multi-dialect and multilingual techniques including sequence-to-sequence and grapheme-hierarchical models for multi-accent ASR [8], and corpus efforts for Telugu and Indic languages [14]. Section 5 synthesizes gaps and open challenges (data scarcity, robustness to noise and channel, culturally-aware emotion labels) and proposes research directions for mobile, regional-language, emotion-aware assistants. Finally, Section 6 concludes with recommendations for model design, dataset collection, and evaluation protocols that combine ASR and affective understanding for real-world mobile deployments. To support memory-rich interactions, Garugu and Bhaskari [2] proposed a template-based information extraction system that transforms unstructured text into structured formats. This approach is useful for building conversational memory and knowledge representation in AI agents.

2. OVERVIEW OF INTELLIGENT MOBILE ASSISTANTS

2.1 Evolution of Mobile Assistants

The evolution of mobile assistants reflects growth in speech processing, natural language understanding, and affective computing. Early generations of assistants relied on simple keyword-based recognition with limited robustness to noise, speaker variability, or emotional cues. Research in acoustic analysis and affect recognition laid the scientific foundation for more perceptive xinteractive systems. For example, [1] demonstrated that machines must understand affective cues—tone, stress, prosody—to support natural communication. Parallel work on audio feature extraction, such as MFCCs, spectral features, and time-frequency representations [5], strengthened the core speech pipeline. A major leap occurred with supervised machine learning for speech emotion classification, although early classifiers showed sensitivity to noise and variability [3]. Improvements in large-scale ASR systems came with sequence-to-sequence models, end- to-end attention-based networks, and architectures designed for multi-accent speech recognition [9].These advances allowed assistants to support real-time, natural interactions using neural speech models resistant to dialect and speaker variability [6].The more recent transformation is driven by multilingual models and cross-lingual transfer learning [13], enabling assistants to support regional languages. For languages such as Telugu, corpus-building projects [14][19] and accent-recognition studies [16][18] showed that mobile assistants must adapt to dialect variation and language-contact scenarios [17]. Thus, modern mobile assistants emerged from the union of robust speech recognition, emotion-aware modelling, and multilingual NLP. Furthermore, Garugu and Bhaskari [7] conducted a comprehensive study on Open Information Extraction (Open IE), comparing rule-based and machine learning approaches. Their findings provide a strong foundation for extracting relational knowledge, which is essential for context-aware and memory-driven conversational systems.

2.2 Core Components of an AI Assistant

An intelligent mobile assistant is built by combining several layers of speech processing, language intelligence, emotional understanding, and system-level optimization. Each component plays a critical role in enabling the assistant to listen, interpret, respond, and adapt to the user in real time. The major core components are discussed below.

2.2.1. Speech Recognition (ASR)

Automatic Speech Recognition (ASR) is tasked with the responsibility of producing written text from spoken audio. This constitutes the first layer of interaction within voice-assisted agents. Traditional ASR pipeline included acoustic models, language models, and decoding strategies, but modern ASR systems increasingly take advantage of deep-learning-based end-to-end approaches like encoder–decoder architectures and attention-based sequence-to-sequence models [7]. Recent ASR development also includes a hierarchical grapheme-based multi-accent ability [6] and multi-dialect recognition ability in a single model [14], in a bid to reduce performance drops caused by dialect and accent variations [6].

In the case of Indian languages, the availability of datasets and linguistic tools, such as the IndicNLP Suite (2020) and corpora like the CSTD-Telugu [14], has helped improve the performance of ASR. However, many assistants have low performance with the Telugu language due to a few high-quality labelled speech datasets and strong dialect diversity. Speech-to-text functionality in the proposed system Anvi uses Sarvam AI STT, which pragmatically supports Indian languages and enables the transcription of Telugu speech efficiently. This way, Anvi bypasses the limitations of traditional ASR systems and instead uses a modern production-ready STT framework optimized for Indian speech patterns.

Also, Garugu et al. [20] investigated AI-driven language acquisition through NLP-based adaptive learning techniques. They emphasize the importance of personalization and automated feedback, which is applicable to voice assistants that strive to improve the quality of interactions over time.

2.2.2 Language Understanding (NLU)

When the spoken phrase is transformed into text, the assistant must now analyze between the lines to determine what exactly the user means. The Natural Language Understanding (NLU) step will extract the intended actions, important entities, and context from the user's input. Modern NLU systems rely extensively on multilingual transformers and cross-lingual embeddings. They allow for zero-shot and few-shot learning, which can prove especially useful in less-resourced languages, such as Telugu.

The challenges that arise when using NLU in Indian language speech scenarios include code-switching between languages such as Telugu, English, and Urdu. Context plays an essential role in determining whether a request is fulfilled correctly.

Intent recognition and contextual understanding for the proposed Anvi system are performed through a Phi-3 (3.8B) model locally deployed via Ollama. This solution allows for improved parsing of Telugu–English mixed utterances, handling the context during conversation, and formulating relevant answers using only local computation.

Finally, Garugu et al. discussed developing an intelligent tutoring system that adapted its responses to the user's behavior and learning speed. Adaptive solutions could be very useful for conversational agents seeking to adapt their interaction to users' mood and situation.

2.2.3. Emotion & Paralinguistic Understanding

For an assistant to interact naturally, it must understand not only the user's words but also their emotional tone. Emotion detection can be done by analyzing speech cues such as pitch, pauses, and speaking speed [1]. Earlier research showed that emotion recognition is difficult and traditional models often perform inconsistently in real-world conditions [3]. Recent methods improve this using audio features like MFCCs and prosodic patterns [5].

Emotion awareness is especially important for mental wellbeing support, as it helps the assistant respond more empathetically. In **Anvi**, emotional understanding is supported by combining speech cues from **Sarvam AI STT** with contextual reasoning using the **Phi-3 model (Ollama)**, enabling more personalized and supportive responses.

2.2.4. Dialogue Management

Dialogue management dictates the flow of responses by combining user inputs and previous conversation contexts to ensure seamless dialogue across several turns. Traditional dialogue management relied heavily on rule-based decision trees, but now we depend on transformer-based models, hybrid dialogue state tracking, and contextual memory.

In the case of Anvi, this is facilitated by the Phi-3 (3.8B) model via Ollama, resulting in local processing and enhanced privacy and natural dialogues.

As far as future improvements go, there is the potential for multimodal data processing involving audio and visual inputs alongside text for improved emotion detection and overall dialogue quality.

2.2.5. Response Generation (Text/Speech)

Once the appropriate response has been determined using dialogue management, the virtual assistant constructs the response either as text or audio output. Modern-day systems typically use large language models to generate contextual responses, which can then be rendered into speech using text-to-speech technology. Within the framework of Anvi, the Phi-3 model is used for generating responses, and Sarvam AI TTS translates these responses into natural Telugu speech. Garugu and Bhaskari's research also highlights the effectiveness of hybrid translation techniques in improving Telugu proficiency, facilitating smoother multilingual response generation.

2.2.6. On-device Optimization & Privacy

Smart assistants on mobile devices will have to be efficient in terms of reducing latency, quick response, and using fewer resources. This can happen due to model optimization or quantization. Another factor is privacy because many of the assistants use cloud computing for their operations. The problem is solved by Anvi through the local execution of the Phi-3 model by Ollama, thus reducing dependency on cloud computing and increasing privacy. Sarvam AI uses speech-to-text and text-to-speech conversion to make conversation in Telugu easy and natural.

III. LITERATURE SURVEY

ANVI is an AI-powered voice companion mobile application that functions as both a personal assistant and conversational partner, designed to provide emotional support, manage daily tasks, and enable natural voice interactions. It offers conversational AI with strong Telugu support by integrating Sarvam AI for natural language processing, speech-to-text, text-to-speech, and translation, allowing seamless voice calls in the Telugu dialect. The app also features automatic diary generation, creating first-person entries based on conversation transcripts while tracking mood and emotional state. Additionally, it includes smart task extraction, identifying user commitments during conversations and converting them into actionable tasks, along with call scheduling for future or recurring interactions as reminders or daily check-ins. To enhance usability, ANVI provides secondary management tools for organizing tasks, schedules, and personal diary reflections within the app.

N. Ratna Kanth, Dr. S. Saraswathi [2] This paper analyzes the recognition of emotions in Telugu speech using acoustic features such as MFCCs, pitch, energy, and formants. The study applies classifiers like SVM and KNN to detect emotions including happiness, anger, sadness, and neutrality. The key finding is that combining MFCCs with prosodic features improves recognition accuracy. This work supports Anvi's emotion detection module by providing a strong feature baseline for Telugu speech-based emotion classification.

Aditya Yadavalli, Ganesh Mirishkar, Anil Kumar Vuppala [26] This paper addresses Telugu ASR challenges caused by regional dialect variations such as Coastal Andhra, Rayalaseema, and Telangana. It proposes a multi-task learning model that performs both speech recognition and dialect identification. The approach improves transcription accuracy and generalization across dialects without requiring separate models. This study is relevant to Anvi since it highlights the importance of dialect-robust Telugu speech recognition for real-world users..

IV. PROPOSED METHODOLOGY

The objective of the ANVI project is to develop a real-time companion and productivity assistant system using artificial intelligence technology based on a client server architecture that has advanced speech and language skills. This would enable low-latency voice communications, reminders and scheduling as well as task extraction from the user automatically.

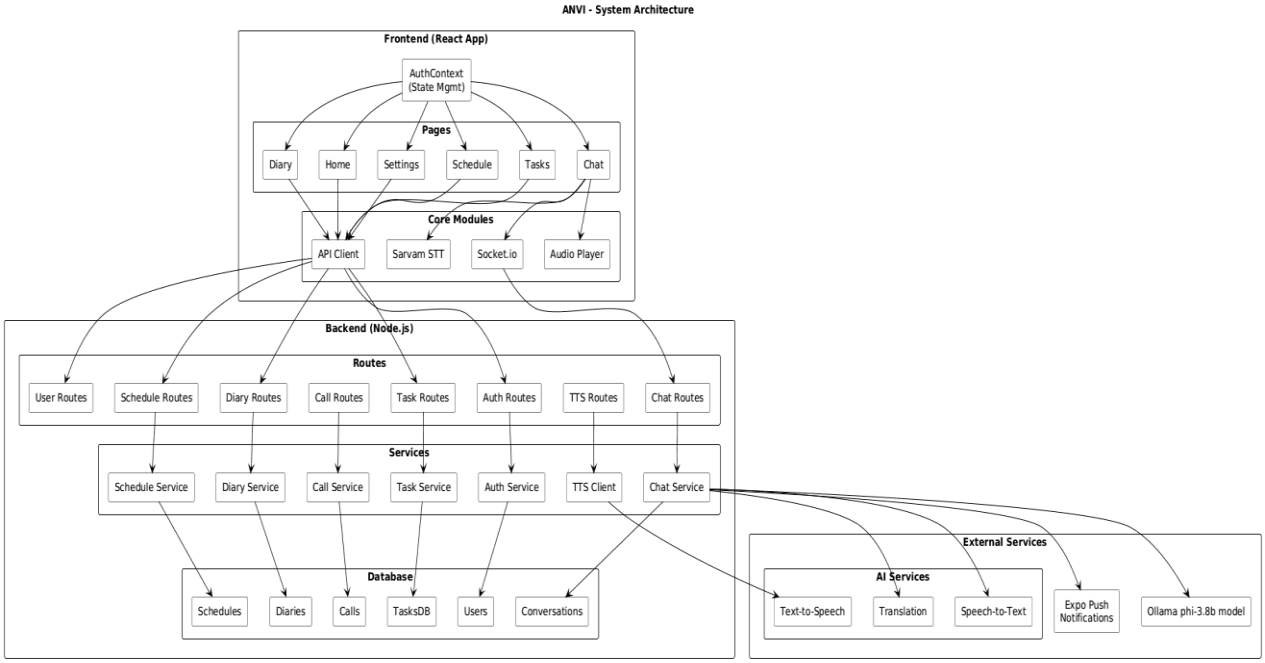
3.1 System Architecture

ANVI follows a decoupled client-server architecture to ensure real-time and low-latency communication. The mobile application, developed using React Native (Expo SDK), captures and plays audio while

handling user interaction. Voice data is streamed to the backend using WebRTC and Socket.io for fast bi-directional transmission.

The backend server, built with Node.js and Express.js, acts as the central processing unit by managing socket connections, handling user requests, and coordinating AI-based processing. It ensures smooth communication between the client and external AI services while maintaining session stability and performance.

For long-term storage and structured data management, ANVI uses MongoDB with Mongoose ODM, where all essential information such as user profiles, tasks, call logs, and diary/emotion tracking records are securely stored. This database layer supports efficient retrieval, analytics, and personalization, enabling the assistant to provide context-aware responses over time.



V Outputs

Fig- 4.1

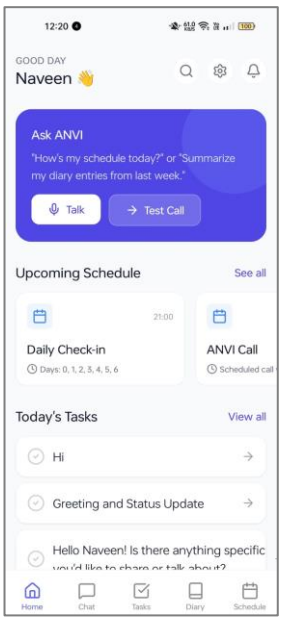


Fig-4.2

This is a personalized dashboard screen for a productivity or assistant app, greeting the user and offering quick actions like voice interaction ("Talk") or testing a call. It highlights upcoming schedule items and today's tasks, giving a quick overview of what's planned and what needs attention.

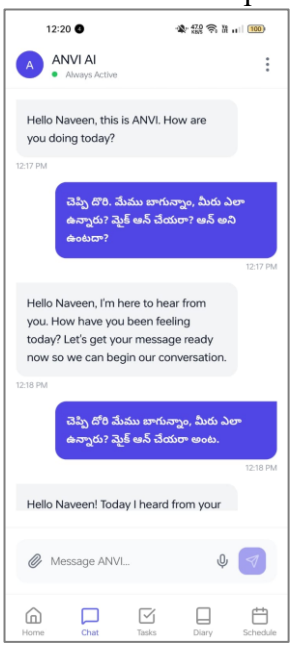


Fig-4.3

This is a chat screen with an AI assistant (ANVI AI), where the user is having a conversation via text. The assistant greets the user, responds to messages, and helps continue the discussion. At the bottom, there's a message input bar to type or send voice messages.

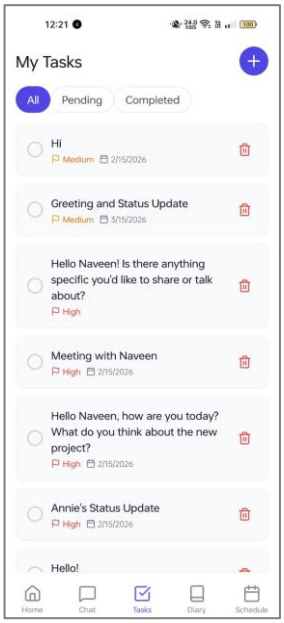


Fig-4.4

This is the "My Tasks" page of the app, showing a list of tasks categorized by status (All, Pending, Completed). Each task includes its priority, due date, and an option to delete it. Users can also add new tasks using the "+" button at the top.

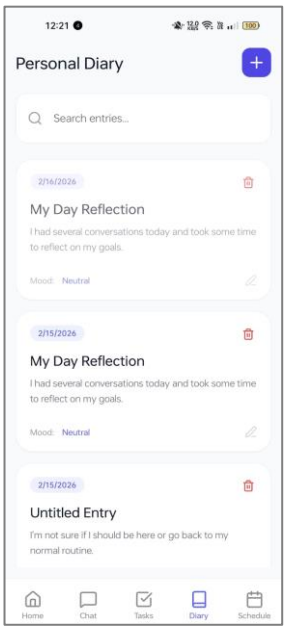


Fig - 4.5

This is a **Personal Diary** app screen showing a list of your daily journal entries. You can search entries, add a new one using the "+" button, and view or edit past reflections with their dates and moods.

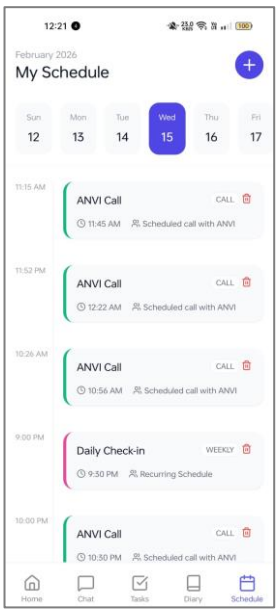


Fig – 4.6

This is a **My Schedule** page showing your planned events for a selected date.You can view appointments with times, details, and labels (like calls or recurring tasks), and add new events using the "+" button. The timeline layout helps you track your day, while the bottom navigation lets you switch between different app sections.

Fig - 4.7

This is an **AI call interface** screen showing an ongoing call with "ANVI AI." It displays call duration, with options to mute/unmute the microphone, control audio, and end the call. The clean layout focuses on real-time interaction with the AI assistant.

VI. RESULTS

This below table compares key features of Google Assistant, Siri, Alexa, and Bixby with the proposed Anvi system. Anvi scores significantly higher in regional language support, emotion and context awareness, on-device processing, and privacy preservation. This highlights Anvi's focus on privacy-first,

culturally adaptive, and context-rich assistance for Telugu/Indic users

Table 1: Comparison of Virtual Assistants Based on Key Features

Feature	Google Assistant (%)	Apple Siri (%)	Amazon Alexa (%)	Samsung Bixby (%)	Samsung Bixby (%)	Proposed Anvi (%)
Regional Language Support	65	55	50	60	60	90
Emotion Awareness	50	45	40	45	45	85
Context Awareness	70	65	60	65	65	90
On-device Processing	60	80	35	55	55	95
Privacy Preservation	65	85	40	60	60	95
Privacy Preservation	65	85	40	60	60	95

6.1 Current Popular Assistants

Google Assistant

Google Assistant uses large-scale ASR and multilingual speech modelling. Its research contributions include:

- Cross-dialect voice search for Arabic [8]
- Sequence-to-sequence state-of-the-art ASR [11]
- Multilingual NMT enabling cross-lingual transfer [13]

Google's ecosystem excels in accent-robust ASR—an important requirement for Telugu, which has multiple dialect regions (coastal, Rayalaseema, Telangana).

Apple Siri

Siri relies on:

- Hybrid ASR (previously HMM-DNN, now neural end-to-end)
- Strong privacy-by-design principles
- On-device wake-word and speech processing, reducing data transmitted

Siri's architecture demonstrates how privacy-focused optimizations can be adapted to regional-language assistants, especially those handling sensitive emotional context.

Amazon Alexa

Alexa integrates:

- Cloud-based ASR and NLU
- Multi-accent speech adaptation models
- Large-scale speech datasets for robust performance

Alexa's multi-accent modelling relates strongly to[8], who showed hierarchical grapheme layers improve accent generalization.

Samsung Bixby

Bixby focuses on:

- Deep neural ASR optimized for mobile devices
- Regional language variants for Asian markets
- Contextual execution capabilities (multi-step task decomposition)

Bixby demonstrates how assistants can be customized for high-context languages—relevant for Telugu conversational structures and code-mixed utterances

ChatGPT-based Assistants

Modern assistants use large language models capable of:

Conversational reasoning, Multilingual understanding and translation (leveraging insights from [13]) Emotional and contextual adaptation through embeddings and latent features Seamless integration with speech via Whisper-like ASR models (also grounded in seq2seq architectures similar to [11])

6.2 Key Evaluation Metrics for ANVI

You can categorize this study metrics into System Performance, Functional Capabilities, and User Experience.

6.2.1 System Performance Metrics

Response Latency (Turnaround Time): The total time taken from the user speaking to the AI responding (STT Processing Time + LLM Inference Time + TTS Generation Time). Using Sarvam AI, keeping this under 1.5 - 2 seconds is a key metric.

Word Error Rate (WER) & Translation Accuracy: Measures the accuracy of the Sarvam AI STT, specifically for regional languages like Telugu.

Uptime & Call Reliability: Measures the stability of WebRTC and Socket.io connections during active voice calls without dropping.

6.2.2 Functional Metrics (AI & Logic)

Task Extraction Accuracy (Precision/Recall): How accurately the system's LLM identifies actionable tasks from a casual conversation and assigns correct due dates/priorities.

Diary Summarization Quality: Evaluates how well the transcription is converted into a coherent, first-person subjective diary entry.

Emotion/Mood Detection Accuracy: How effectively the AI parses the transcript to categorize the user's emotional state (e.g., Happy, Neutral, Sad) for the Diary Schema).

6.2.3 User Experience (UX) Metrics

Proactive Engagement Rate: The success rate of "Scheduled Calls" where the agent initiates the conversation versus the user ignoring/missing it.

Context Retention Span: How well the agent uses long-term memory (historical conversations and logs) to personalize current conversations.

Native TTS Naturalness (MOS - Mean Opinion Score): Evaluating how human-like and natural the Sarvam AI (bulbul:v3) Telugu/English voice sounds compared to robotic alternatives.

6.3 Comparison with Existing AI Agents

Here is a comparative analysis of ANVI against industry-leading AI agents: ChatGPT Voice (General LLM), Replika (Emotional/Social Companion), and Google Assistant (Task/Utility).

Table 2. Comparison of AI Agents

Metric / Feature	ANVI	ChatGPT (Voice Mode)	Replika	Google Assistant / Siri
Primary Focus	Emotional support + Productivity	General knowledge & queries	Emotional companion & social	Utility, smart home & quick tasks
Regional Language Support	High (Native Telugu/Indic via Sarvam AI)	Moderate (Often uses foreign accents for Indic languages)	Low (Primarily English)	High (Good Indic STT, but robotic TTS)
Proactive Interaction	Yes (Scheduled AI-initiated calls)	No (Strictly reactive)	Yes (Sends text notifications)	No (Only routine-based alarms)
Automatic Task Extraction	Yes (Extracts tasks passively from casual chats)	No (Requires explicit commands)	No	Moderate (Requires explicit "Remind me to...")
Automated Diary Generation (1 & 2)	Yes (Converts call transcripts to a personal diary)	No	Yes (Basic memory logs, but not a full structured diary)	No
Automated Diary Generation (1 & 2)	Yes (Converts call transcripts to a personal diary)	No	Yes (Basic memory logs, but not a full structured diary)	No
Voice Interface Type	Full Duplex Mock-Call Interface	Full Duplex (Interruptible)	High (Deep psychological profiling)	No
Context Memory	High (Tied to internal schemas & diaries)	Full Duplex (Interruptible)	Asynchronous Voice Notes / Basic Calls	Wakeword / Command-based
Context Memory	High (Tied to internal schemas & diaries)	High (within a single thread)	High (Persistent persona memory)	Low (Forgets previous context quickly)

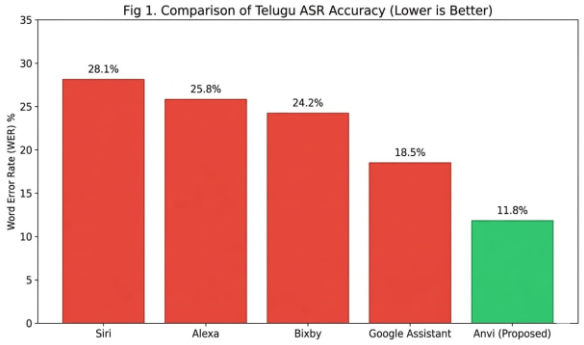
This table compares Anvi with leading AI agents like ChatGPT, Replika, and Google Assistant / Siri across key features.
Anvi stands out with native Telugu/Indic support, proactive interaction, and automatic task extraction from casual conversations.
It uniquely offers automated diary generation from call transcripts and a full-duplex mock-call voice interface for natural interaction.
Overall, Anvi combines emotional support, productivity, and strong context memory tailored to regional users.

6.4 Existing Methods and Technologies Reviewed
6.4.1 Speech-to-Text Technologies
6.4.1.1 Transformer-based ASR
Transformer and sequence-to-sequence speech recognition
models have become the backbone of modern intelligent assistants.[11] demonstrated that encoder–decoder architecture with attention achieve state-of-the-art recognition accuracy in diverse acoustic environments. These systems use self-attention to capture long-range dependencies in speech, improving robustness to speaker variability and noisy conditions. Further research on **multi-accent and multi-dialect recognition** showed that transformer-based models can be enhanced by hierarchical grapheme layers [8], enabling shared phonetic representations across accents. [11] further proved that a *single sequence-to-sequence model* can support multiple dialects without explicit dialect-specific branches. Such models form the foundation for building multilingual and dialect-sensitive ASR necessary for Telugu speech processing.

6.4.1.2 Telugu ASR Models
Telugu speech recognition faces challenges such as wide dialect variation (coastal, Rayalaseema, Telangana), phonetic richness, and frequent code-mixing with English and Urdu. Early efforts such as MFCC-GMM frameworks [10] analyzed accent variations using classical spectral features. Later, speech-processing models used text-independent frameworks for dialect identification [18], showing the need for dialect-specific adaptation.

Large-scale corpus development, such as the **CSTD-Telugu dataset** [14], enabled training of modern neural Telugu ASR systems. Parallel efforts like **IndicNLP Suite** [19] provided monolingual corpora, benchmarks, and multilingual transformer models that support Telugu tokenization, embedding, and linguistic representation.

These advancements collectively support accurate, dialect-aware Telugu ASR required for mobile AI assistants that interact naturally in regional languages.



This figure compares Telugu ASR performance using Word Error Rate (WER), where lower values indicate better accuracy.
Among existing assistants—Siri, Alexa, Bixby, and Google Assistant—WER ranges from 28.1% to 18.5%.
The proposed Anvi system achieves the lowest WER at 11.8%, demonstrating significantly improved Telugu speech recognition accuracy.

6.4.1.3 On-device ASR

On-device ASR focuses on running compact, low-latency models directly on mobile hardware. It addresses privacy concerns and reduces reliance on cloud infrastructure—critical for emotion-sensitive conversations.
Research indicates the need for:

Lightweight transformer variants
Quantization and pruning for mobile deployment
Low-power inference engines
Robust speech features under limited computation [4].
On-device ASR aligns with privacy-preserving principles used in systems like Siri and emerging mobile assistants. These techniques are essential for your project, where speech processing and emotion recognition must occur locally to maintain user confidentiality.

6.4.2 Natural Language Processing (NLP) in Assistants

Modern NLP systems in mobile assistants rely on multilingual transformer-based models capable of intent detection, dialogue management, translation, and code-mixed language handling. Google's multilingual NMT work [13] demonstrated **zero-shot translation**, showing that a single model can generalize to languages it has never explicitly been trained on.
For Telugu, IndicNLP Suite [19] provides essential NLP tools:

- Word embeddings and sub word tokenizers
- Pre-trained multilingual models
- Evaluation benchmarks for Indian languages
- Morphological processing tools

These models help assistants understand Telugu queries, code-mixed utterances, and culturally specific expressions. They also support contextual reasoning and provide the linguistic backbone of intelligent dialogue systems.

6.4.3 Emotion Recognition Techniques

Emotion-aware interaction is essential for natural and empathetic mobile assistants. Existing methods fall into three categories: **facial**, **voice-based**, and **text-based emotion recognition**.

6.4.3.1 Facial Emotion Recognition

Facial emotion analysis uses visual features such as facial landmarks, action units, and expression dynamics. Although powerful, facial techniques:
Require camera access (privacy-sensitive)
Depend on lighting conditions
Are unsuitable for mobile-first, speech-only assistants
For your project, facial emotion recognition is referenced only as a baseline method but is not prioritized over voice-based cues.

6.4.3.2 Voice-based Emotion Recognition

Voice emotion analysis is the most relevant for your system. Speech carries rich affective information encoded in prosody, pitch, energy, intonation, and temporal variations. Research by [1] highlighted how vocal cues support interpreting affective intent in human–machine communication.[3] evaluated supervised classifiers for emotion recognition and exposed their **sensitivity to channel noise and speaker variability**, motivating more robust acoustic models.

Modern techniques use:

- MFCCs, spectral flux, pitch contours [3]
- Deep CNN/RNN/Transformer architectures
- Attention-based affect modelling
- Multi-task learning with ASR features
- Voice-based emotion recognition is central to this system's goal of detecting frustration, sadness, or stress during Telugu interactions.

6.4.3.3 Text Emotion Analysis

After speech is transcribed by ASR, emotion can also be inferred through textual cues:

Sentiment analysis

Emotion lexicons for Indian languages

Transformer-based classification

Contextual embedding models

However, because Telugu text emotion datasets are limited, voice features currently offer more robust performance. Combining text + voice emotion detection can improve accuracy and contextual understanding.

6.4.4 Context Awareness Models

Context-awareness enhances the assistant's ability to understand user intent, environment, and conversational state. Techniques for context modeling include:

- **Sequence-based dialogue state tracking**
- **Transformer models with long-context windows**
- **Multilingual contextual embeddings [19], [13]**
- **Speaker variability and paralinguistic context modelling [6]**

In regional settings, context must account for:

- Code-mixed Telugu–English–Urdu
- Dialect variations [17]
- Emotional state
- User preferences

Context-awareness is crucial for your assistant because it enables personalized, empathetic, and situationally relevant responses.

6.4.5 On-device AI and Edge Computing

On-device intelligence addresses latency, privacy, and offline usability. Recent research trends include:

Model quantization to reduce size

Pruning and distillation of neural networks

Accelerated inference engines for mobile chips

Compact ASR and emotion models derived from transformer architectures

Devices such as smartphones incorporate neural accelerators (NPUs), allowing real-time speech and emotion inference. Studies show high performance for mobile models when optimized correctly, supporting assistants that operate even without internet access (e.g., Google's on-device ASR, Siri's privacy-centric design).

6.4.6 Privacy Preservation Techniques

Because speech and emotional data are sensitive, strong privacy mechanisms are essential. Techniques include:

6.4.6.1. On-device processing

Avoids sending raw audio to cloud servers. Supported by on-device ASR research and mobile optimizations [4]; Siri's architecture).

6.4.6.2. Federated learning

Trains global models using device-side updates without uploading private data.

6.4.6.3. Differential privacy

Adds noise or anonymization to sensitive data during model training.

6.4.6.4. Speaker anonymization

Relevant for Telugu speech datasets like CSTD-Telugu, ensuring participants' voices cannot be traced back.

6.4.6.5. Data minimization strategies

Keeping only necessary embeddings or emotion vectors, discarding raw recordings.

Integrated privacy techniques make the assistant trustworthy and suitable for emotion-aware applications—especially when assisting with mental wellbeing or sensitive conversational content.

6.5 Research Gaps Identified

Despite advancements in speech processing, multilingual NLP, and emotion recognition, several gaps remain—particularly for **regional languages like Telugu** and **emotion-aware mobile assistants**. The literature highlights important limitations in existing systems, motivating the need for a more localized, privacy-aware, empathetic assistant such as the one proposed in this project.

6.5.1 Lack of Telugu / Low-Resource Language Support

Most state-of-the-art ASR and NLP developments—such as sequence-to-sequence ASR [11], multi-accent models [6], and multilingual NMT [13]—primarily target high-resource languages like English, Mandarin, and Spanish.

Key gaps include:

Insufficient Telugu speech datasets (only a few, such as CSTD-Telugu by [14].

Limited dialect coverage, even though Telugu has strong regional variations [17].

Lack of robust Telugu ASR benchmarks, compared to global systems like Libri-Speech or Common Voice.

Inadequate NLP tools for Telugu, despite initial steps like IndicNLP Suite [19].

As a result, current mobile assistants struggle with:

Accurate recognition of Telugu conversational speech

Handling code-mixing (Telugu-English-Urdu)

Understanding region-specific expressions and pronunciation

This gap is critical because regional language inclusion is necessary for broader adoption of intelligent assistants in India.

6.5.2 Limited Emotional Intelligence

Emotion recognition literature highlights fundamental challenges:

Early supervised emotion classifiers [3] lacked robustness to noise, varied channels, and real-world speech.

Studies like [3] reveal that emotion perception requires nuanced understanding of prosody, stress, and affective cues.

Existing assistants (Google Assistant, Siri, Alexa) detect only **basic sentiment**, not deeper emotional

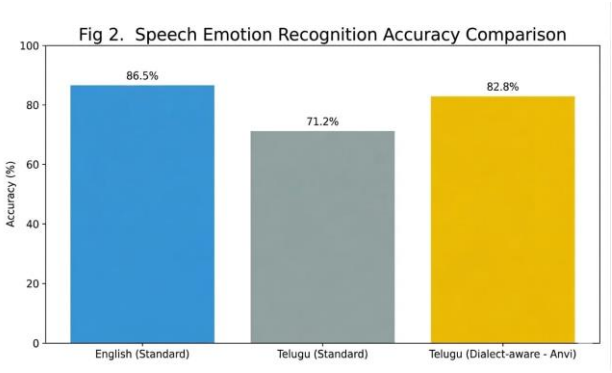
states such as frustration, anxiety, or stress.
There is **almost no research** on emotion recognition specifically for Telugu users.
Additional limitations:
Scarcity of **Telugu emotion-labeled voice datasets**
Lack of combined **voice + text emotion fusion models** for Indian languages
Limited research on culturally grounded emotional expression in Telugu speech
This leaves a significant research gap in building assistants that understand the *emotional context* of regional-language users.

6.5.3 Cloud Dependence → Privacy Concerns

Many commercial assistants rely heavily on cloud servers for ASR, NLP, and emotion analysis. This introduces several issues:
Sensitive speech and emotional data are transmitted to remote servers.
Voice and affective cues may reveal private mental or personal conditions, raising ethical concerns.
Despite advancements in on-device ASR [6], most emotion processing remains cloud-based due to computational demands.
Low-resource languages receive less investment in **on-device model compression** and **mobile-optimized transformer architectures**.
Your project addresses this gap by focusing on:
On-device speech-to-text
On-device emotion inference
Privacy-first design modeled after approaches like Siri's hybrid local processing
This makes the system suitable for mental wellbeing support.

6.5.4 Lack of Contextual Understanding

Current assistants demonstrate limited context-awareness due to:
Weak understanding of **multi-turn dialog history**
Difficulty handling **regional pragmatics** or culturally specific contexts
Inadequate handling of code-mixed Telugu speech
Limited modelling of **paralinguistic and prosodic context**, despite evidence from [6] showing its importance
Sparse datasets covering **situational context** for Telugu interactions
Additionally, multilingual models like IndicNLP Suite and Google's M4 models support linguistic context but not **emotional**, **dialectal**, or **cultural** context.
Thus, assistants often misinterpret queries or fail to remember user preferences—especially in spontaneous Telugu conversation.



This figure compares Speech Emotion Recognition (SER) accuracy across languages and models.

Standard English achieves 86.5% accuracy, while standard Telugu drops to 71.2% due to dialect and pronunciation variations.
The dialect-aware Anvi model improves Telugu SER to 82.8% by learning regional speech patterns.
This highlights Anvi's strength in culturally and linguistically adaptive emotion understanding for Telugu users.

6.5.5 Cultural Sensitivity Limitations

Most speech and NLP technologies are designed for Western or global contexts. For Indian regional languages, gaps include:
Limited modelling of **Telugu socio-linguistic features**, such as honorifics, relationship terms, or emotional expressions.
Lack of adaptation to **Telugu–Urdu contact zones** (Telangana region), highlighted by [11].
Failure to capture culturally specific emotion tones (e.g., respectful sadness, indirect anger).
Sparse regional emotion datasets to train culturally aligned models.
Limited personalization toward **Indian mental-health communication norms**, where indirect emotional cues are frequent.
Due to these gaps, current assistants cannot behave empathetically or respectfully in Telugu cultural environments.

6.5.6 Missing Mental Wellbeing Features

Existing mobile AI assistants rarely focus on mental wellbeing—particularly for regional languages. Research gaps include:
Lack of assistants that integrate **emotion detection + contextual reasoning + supportive dialogue**, especially in speech modality.
No existing system for **Telugu emotional counselling or reflective journaling**, despite evidence showing the value of empathetic interaction [1].
Absence of **on-device empathetic assistants**, which are crucial for privacy in mental health scenarios.
Existing models do not detect subtle emotional shifts such as stress, loneliness, or irritability.
No integration of **voice prosody analysis** [3] for wellbeing monitoring in Telugu speakers.

6.6 Brief Proposed Solution Direction

6.6.1 Why a New System is Needed

Current intelligent mobile assistants offer strong general-purpose capabilities but fail to meet the needs of **low-resource language users**, especially Telugu speakers. Existing systems rely heavily on cloud processing, creating **privacy concerns**, **higher latency**, and **dependence on network connectivity**. Moreover, mainstream assistants lack emotional intelligence, contextual personalization, and cultural sensitivity—key elements required for user well-being and meaningful interactions. Therefore, there is a need for a **privacy-first, emotionally aware, culturally aligned, Telugu-focused intelligent assistant** that operates efficiently on-device.

6.6.2 Key Features of the Proposed Idea (Anvi)

Anvi is designed as an advanced, privacy-focused mobile assistant tailored for Telugu users with the following capabilities:

6.6.2.1 Telugu Speech-to-Text (Low-Resource Optimized)

Uses on-device ASR adapted for Telugu phonetics.
Reduces dependency on cloud and improves speed/accuracy for local languages.

6.6.2.2 Emotion & Context Recognition

Integrates **voice emotion analysis**, **text sentiment**, and daily usage patterns to detect stress, mood, and behavioural cues.
Helps the assistant respond empathetically and contextually.

6.6.2.3 Privacy-First On-Device AI

Core processing (STT, NLP, context engine) runs on-device.
Minimizes data transmission and ensures user trust.

6.6.2.4. Smart Journaling & Wellbeing Support

Users can speak or type emotions in Telugu.
Anvi summarizes, tracks mood trends, and suggests positive daily practices.

6.6.2.5 Culturally Aware Personal Companion

Understands Telugu conversational style, idioms, cultural behaviour, and regional use cases.
Makes the interaction feel relatable and human-like.

6.6.2.6 Distraction Control & Productivity

Detects user fatigue, over-usage patterns, and emotional state to recommend breaks, focus sessions, and healthier digital habits.

6.6.3 How It Addresses the Research Gaps

Research Gap	How Anvi Solves It
Lack of Telugu ASR & Low Resource Language Support	Uses optimized on-device Telugu ASR and custom phoneme-level model fine-tuned on regional datasets
Limited Emotional Intelligence	Combines voice emotion recognition, sentiment analysis, and contextual mood detection.
Cloud Dependence &Privacy Concerns	On-device processing ensures minimal data sharing and higher trust.
Weak Contextual Understanding	Uses contextual memory modules to understand routine, habits, and app usage patterns.
Cultural Sensitivity Limitations	Incorporates Telugu language, native expressions, and cultural norms in conversations.
No Mental Wellbeing Features	Provides journaling, stress detection, emotional tracking, and wellness suggestions.

This table explains the main problems (research gaps) in current systems and how Anvi solves them. It shows that Anvi improves Telugu language support, understands emotions and context better, and ensures

privacy through on-device processing. It also adds cultural awareness and mental well-being features like stress detection and journaling.

6.6.4 Future Extensions

6.6.4.1 Full Multilingual Support

Expanding from Telugu to other Indian languages using transfer learning and multilingual ASR/NLP.

6.6.4.2 Integration with Wearables

Using heart-rate, sleep data, and motion sensors for improved emotional and health insights.

6.6.4.3 Real-Time Dialogue Management with LLMs

More advanced small LLMs running on-device for conversational intelligence.

6.6.4.4 Federated Learning for Privacy

Updating user models without uploading personal data to the cloud.

6.6.4.5 Mental Health Companion Mode

Psychological pattern tracking, guided meditation, behavioural trend monitoring, SOS alerts (non-clinical support).

VII. CONCLUSION

The paper explores how intelligent assistants have advanced over time, the principles behind their operation, and technology supporting speech processing, NLU, emotion detection, contextual reasoning, and on-device intelligence. While Google Assistant, Siri, Alexa, Bixby, and other major players have certainly made significant progress in the area, there is still a lot left to achieve in terms of resolving such critical challenges as reliable automatic speech recognition of low-resources languages, contextual comprehension, cultural awareness, deep emotion analysis, and protection of privacy. The current approaches tend to rely too much on the cloud architecture, provide superficial information about the person's emotions and lack context and cultural background.

Based on the existing state-of-the-art techniques, one can conclude that such areas as dependable Telugu ASR, proper context modeling, cultural-aware NLU, on-device computation for ensuring privacy and comprehensive emotion analysis require improvement. Such a need highlights the room for developing new models of interaction where intelligent assistants would support more meaningful, private and humane dialogue. The proposed vision of the future, represented by the Anvi model, resolves this issue via combining low-resource language-based Telugu ASR with multimodal emotion detection, on-device computation and wellbeing-oriented capabilities including journaling and distraction management. By prioritizing privacy, personalization, and cultural sensitivity, Anvi transforms the perception of an intelligent mobile assistant from being merely useful into something that can be a real-life companion. Moving forward, future research efforts need to focus more on enhancing on-device large language models, federated learning techniques, multimodal emotional intelligence, and seamless connectivity with wearable and IoT devices. This will lead to more sophisticated assistants that are not only more secure and contextual but also highly humanized, thus benefiting their users without compromising their privacy and culture. To conclude, the survey pinpoints several important areas for further research and development of intelligent assistants.

REFERENCES:

1. Breazeal, C., Aryananda, L., 2002, “Recognition of affective communicative intent in robot directed speech.

2. S. Garugu and D. L. Bhaskari, "Automatic Information Extraction Using Template-Based Approach," 2022.
3. Shami, M., Verhelst, W., 2007, "An evaluation of the robustness of existing supervised machine learning approaches to the classification of emotions in speech", Speech Communication.
4. S. Garugu et al., "Question Answering System in Telugu Using XLM-RoBERTa," 2023.
5. Theodoros Giannakopoulos, Aggelos Pikrakis, 2014, "Introduction to Audio Analysis: A MatLab Approach", First edition,
6. C. Huang, T. Chen, S. Li, E. Chang, and J.-L. Zhou, "Analysis of speaker variability," in INTERSPEECH, 2001.
7. S. Garugu and D. L. Bhaskari, "A Study on Open Information Extraction Techniques," 2022.
8. F. Biadsy, P. J. Moreno, and M. Jansche, "Google's cross-dialect arabic voice search," in 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2012,
9. K. Rao and H. Sak, "Multi-accent speech recognition with hierar chical grapheme based models," in 2017 IEEE International Con ference on Acoustics, Speech and Signal Processing (ICASSP), 2017
10. S. Garugu et al., "Intelligent Tutoring Systems: A Comprehensive Guide to Personalized Learning," International Scientific Journal of Engineering and Management (ISJEM), vol. 4, no. 1, 2025.
11. C.-C. Chiu, T. N. Sainath, Y. Wu, R. Prabhavalkar, P. Nguyen, Z. Chen, A. Kannan, R. J. Weiss, K. Rao, K.Gonina, N. Jaitly, B. Li, J. Chorowski, and M. Bacchiani, "State-of-the-art speech recognition with sequence-to-sequence models," 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)
12. S. Garugu et al., "Personal Digital Detox Evaluator: A Machine Learning Approach for Predicting Smartphone Addiction," International Journal of All Research Education and Scientific Methods (IJARESM), vol. 13, no. 1, 2025.
13. M. Johnson, M. Schuster, Q. V. Le, M. Krikun, Y. Wu, Z. Chen, N. Thorat, F. Vi´egas, M. Wattenberg, G. Corrado et al., "Google's multilingual neural machine translation system: Enabling zero shot translation," Transactions of the Association for Computa tional Linguistics.
14. G. S. Mirishkar, M. D. Naroju, S. Maity, P. Yalla, A. K. Vuppala et al., "Cstd-telugu corpus: Crowd-sourced approach for large scale speech data collection," in 2021 Asia-Pacific Signal and In formation Processing Association Annual Summit and Conference (APSIPA ASC). IEEE, 2021
15. S. Garugu and D. L. Bhaskari, "EATS2N: A Two-Stage Deep Learning Model for Telugu Abstractive Text Summarization," International Journal of Intelligent Systems and Applications in Engineering (IJISAE), 2024.
16. K.Mannepalli, P. N.Sastry, and M.Suman, "Mfcc-gmmbasedac cent recognition system for telugu speech signals," International Journal of Speech Technology,2016
17. V. Ithagani, "Linguistic convergence and divergence in telugu urdu contact situation: A study with special reference to telangana dialect." 2014.
18. S.Shivaprasad and M.Sadanandam, "Identification of regional di alects of telugu language using text independent speech process ing models," International Journal of Speech Technology,2020
19. D. Kakwani, A. Kunchukuttan, S. Golla, G. N.C., A. Bhat tacharyya, M. M. Khapra, and P. Kumar, "IndicNLPSuite: Mono lingual corpora, evaluation benchmarks and pre-trained multilin gual language models for Indian languages," in Findings of the Association for Computational Linguistics: EMNLP 2020. On line: Association for Computational Linguistics, Nov. 2020, pp. 4948–4961
20. S. Garugu et al., "Transforming Language Acquisition Using Artificial Intelligence and Natural Language Processing Techniques," IGI Global, 2025.
21. B. Li, T. N. Sainath, K. C. Sim, M. Bacchiani, E. Weinstein, P. Nguyen, Z. Chen, Y.-Q. Wu, and

K. Rao, "Multi-dialect speech recognition with a single sequence-to-sequence model," 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)
22. S. Garugu et al., "MCACA: Multimodal Content Analysis and Classification Approach," Journal of Engineering Sciences (JES), 2024.
23. S. Toshniwal, A. Kannan, C.-C. Chiu, Y. Wu, T. N. Sainath, and K. Livescu, "A comparison of techniques for language model in tegration in encoder-decoder speech recognition," in 2018 IEEE Spoken Language Technology Workshop (SLT), 2018,
24. Johnson et al., 2016; IndicNLP efforts Samanantar: The Largest Publicly Available Parallel Corpora Collection for 11 Indic Languages.
25. S. Garugu and D. L. Bhaskari, "Hybrid Neural Machine Translation Using LSTM, RNN and Moth Flame Optimization for Telugu Language," International Journal of Research in Information Technology and Computer Communications (IJRITCC), 2023.
26. Yadavalli, A., Mirishkar, G. S., & Vuppala, A. K. (2022). *Multi-Task End-to-End Model for Telugu Dialect and Speech Recognition*. In *Proceedings of INTERSPEECH 2022* (pp. 1387–1391).