PANIMALAR
ENGINEERING COLLEGE
An Autonomous Institution, Affiliated to Anna University, Chennai
A Christian Minority Institution
(JAISAKTHI EDUCATIONAL TRUST)
Approved by All India Council for Technical Education

BRIDGING THE SILENCE

Phone calls for deaf people to attend and communicate

SDG GOALS : REDUCED INEQUALITY

NAME: Varshita Sarwan (211423104719)
Vidushi Ganeshika M (211423104730)

DOMAIN NAME: Speech interface/machine learning

COORDINATORS: Dr.SUBEDHAV, Mrs.SHARMILA.S

Agenda

- **Core Challenge:** Phone calls remain largely inaccessible for the deaf and hard-of-hearing (DHH) community in India due to poor carrier support for Real-Time Text (RTT) and reliance on inefficient, human-mediated systems.
- **Proposed Solution:** Develop an AI-powered, end-to-end communication platform that supports dual-modality interaction—combining Speech-to-Text (STT), Text-to-Speech (TTS), and Indian Sign Language (ISL) recognition.
- **Implementation Strategy:** Begin with a low-latency transcription-based MVP and gradually integrate full ISL translation for real-time, inclusive communication.
- **Success Factors:** Ensure adoption through strong technology infrastructure, policy advocacy for accessibility standards, and community outreach to promote awareness and usage.

INTRODUCTION

- Phone calls are often inaccessible to deaf and hard-of-hearing individuals. Existing solutions usually depend on human-operated relay services or require expensive, specialized assistive devices, which makes them inconvenient and difficult to adopt widely.

- To address this gap, our aim is to create software that is simple, affordable, and easy to use. By focusing on accessibility and user-friendliness, the solution can provide inclusive communication without relying on costly equipment or external support.

OBJECTIVE

- AI speech recognition (e.g., Whisper or Google STT) to convertthecaller's speech into live text.
- Simulate RTT behavior via WebRTC/WebSocket/Firebase for real-timemessaging.
- Implement Text-to-Speech (TTS) for typed responses to be deliveredtothe hearing user.
- Optionally include AI sign language recognition using computervision (e.g.,MediaPipe, TensorFlow) for enhanced accessibility.

LITERATURE SURVEY

S . NO	TITTLE	AUTHOR & PUBLISHED YEAR	ALGORITHM	ADVANTAGE	DISADVANTAGE
1.	**SyncSpeech: Dual-Stream TTS**	Zhengyan Sheng, Zhihao Du, Shiliang Zhang, Zhijie Yan, Yexin Yang, Zhenhua Ling **Year: 2025**	**Temporal Masked Transformer (TMT) + duration prediction**	**Low latency, high efficiency, good quality in EN & ZH**	**Complex training, needs alignment tools, data dependent**
2.	**Emformer: Streaming ASR**	Yangyang Shi, Yongqiang Wang, Chunyang Wu, Ching-Feng Yeh, Julian Chan, Frank Zhang, Duc Le, Mike Seltzer **Year: 2021**	**Efficient Memory Transformer (caching + memory bank)**	**4.6× faster training, 18% lower RTF, very low latency**	**Complex design, context leakage risk**
3.	**Real-Time Speech-to-Text Holographic Communication**	Sreevathsa Sree Charan, Suja Palaniswamy, Vemula Srihitha, Savarala Chethana **Year: 2021**	**Hybrid ASR (Conventional + retrained Wav2Vec2) on Raspberry Pi**	**Inclusive assistive tech, dual ASR improves accuracy, low-cost hardware**	**Limited hardware, noise-sensitive, weak multilingual support**

4.	**Continuous Indian Sign Language Gesture Recognition and Sentence Formation**	Kumud Tripathi, Neha Baranwal, G.C. Nandi **Year: 2015**	**Gradient-based keyframe extraction, Orientation Histogram, PCA, Distance classifiers**	**Handles continuous ISL gestures, removes redundant frames, good accuracy**	**Needs high-quality dataset; limited sign vocabulary**
5.	**High Quality Streaming Speech Synthesis with Low, Sentence-Length-Independent Latency**	Nikolaos Ellinas, Giorgos Athanasopoulos, Arseniy Gorin, Anastasios Mouchtaris **Year: 2021**	**Autoregressive seq2seq, location-based attention, LPCNet vocoder**	**Low-latency TTS, CPU-friendly, real-time speech synthesis**	**Autoregressive model limits inference speed; requires careful tuning**
6.	**Live Streaming Speech Recognition Using Deep BLSTM Acoustic Models and Interpolated Language Models**	Javier Jorge, Xabier Sarasola, Asier Lopez Zorrilla, Mikel Peñagarikano **Year: 2022**	**BLSTM acoustic models, sliding window decoding, interpolated Transformer LMs**	**High recognition accuracy under streaming; real-time one-pass decoding**	**BLSTM windowing introduces latency; computationally intensive**
7.	**FastEmit: Low-Latency Streaming ASR with Sequence-Level Emission Regularization**	Anmol Gulati, Yu Zhang, James Qin, Chung-Cheng Chiu, Niki Parmar, Colin Raffel Year: 2020	**RNN-T with FastEmit loss, sequence-level emission regularization**	**Reduces streaming ASR latency, improves stability**	**Training sensitive to λ; accuracy trade-offs possible**

8.	**Towards Indian Sign Language Sentence Recognition Using INSIGNVID Dataset**	Kinjal Mistree, Devendra Thakor, Brijesh Bhatt **Year: 2021**	**Transfer learning with MobileNetV2, vision-based recognitionR**	**First official ISL dataset, sentence-level recognition**	**Dataset small, limited signer diversity**
9.	**Universal ASR: Unifying Streaming and Non-Streaming ASR Using a Single Encoder-Decoder Model**	Zhifu Gao, Shiliang Zhang, Ming Lei, Ian McLoughlin **Year: 2020**	**SCAMA, SAN-M encoder, Dynamic Latency Training (DLT)**	**Single unified model for streaming/offline; flexible latency**	**High training complexity; heavy compute**
10.	**Continuous Sign Language Recognition System Using Deep Learning with MediaPipe Holistic**	Sharvani Srivastava, Sudhakar Singh, Pooja, Shiv Prakash **Year: 2024**	**MediaPipe Holistic landmarks, LSTM**	**Real-time ISL recognition; 88.23% accuracy**	**Limited to dataset; may miss complex gestures**
11.	**Recognition of Indian Sign Language in Live Video**	Joyeeta Singha, Karen Das **Year: 2013**	**Skin filtering, Eigenvectors, Eigenvalue-weighted Euclidean distance**	**High accuracy (96.25%); bare-hand recognition**	**Small dataset; limited scalability**
12.	**Dual-Mode ASR: Unify and Improve Streaming ASR with Full-Context Modeling**	Jiahui Yu, Wei Han, Anmol Gulati, Chung-Cheng Chiu, Bo Li, Tara N. Sainath, Yonghui Wu, Ruoming Pang **Year: 2021**	**Conformer/ContextNet, dual-mode encoder, inplace knowledge distillation**	**Joint model improves streaming accuracy & latency**	**Complex implementation; large resource needs**

13.	**Real-Time Indian Sign Language (ISL) Recognition**	Kartik Shenoy, Tejas Dastane, Varun Rao, Devendra Vyavaharkar **Year: 2018**	**Grid-based feature extraction, k-NN, HMM**	**High pose accuracy (99.7%), gesture accuracy (97.23%)**	**Limited gestures, single-hand focus**
14.	**A Language Agnostic Multilingual Streaming On-Device ASR System**	Bo Li, Tara N. Sainath, Ruoming Pang, Shuo-yiin Chang, Qiumin Xu, Trevor Strohman **Year: 2022**	**Conformer RNN-T with Encoder Endpointer + EOU Joint Layer**	**On-device, real-time, supports multilingual & code-switching**	**Needs large capacity, slight WER drop, big models impractical**
15.	**A Truly Multilingual First Pass and Monolingual Second Pass Streaming On-Device ASR System**	Sepand Mavandadi, Bo Li, Chao Zhang, Brian Farris, Tara N. Sainath, Trevor Strohman **Year: 2023**	**Cascaded Encoders (multilingual first-pass + monolingual second-pass)**	**Better WER than monolingual, parameter efficient, supports code-switching**	**Needs LID for second pass, complex per-language components**

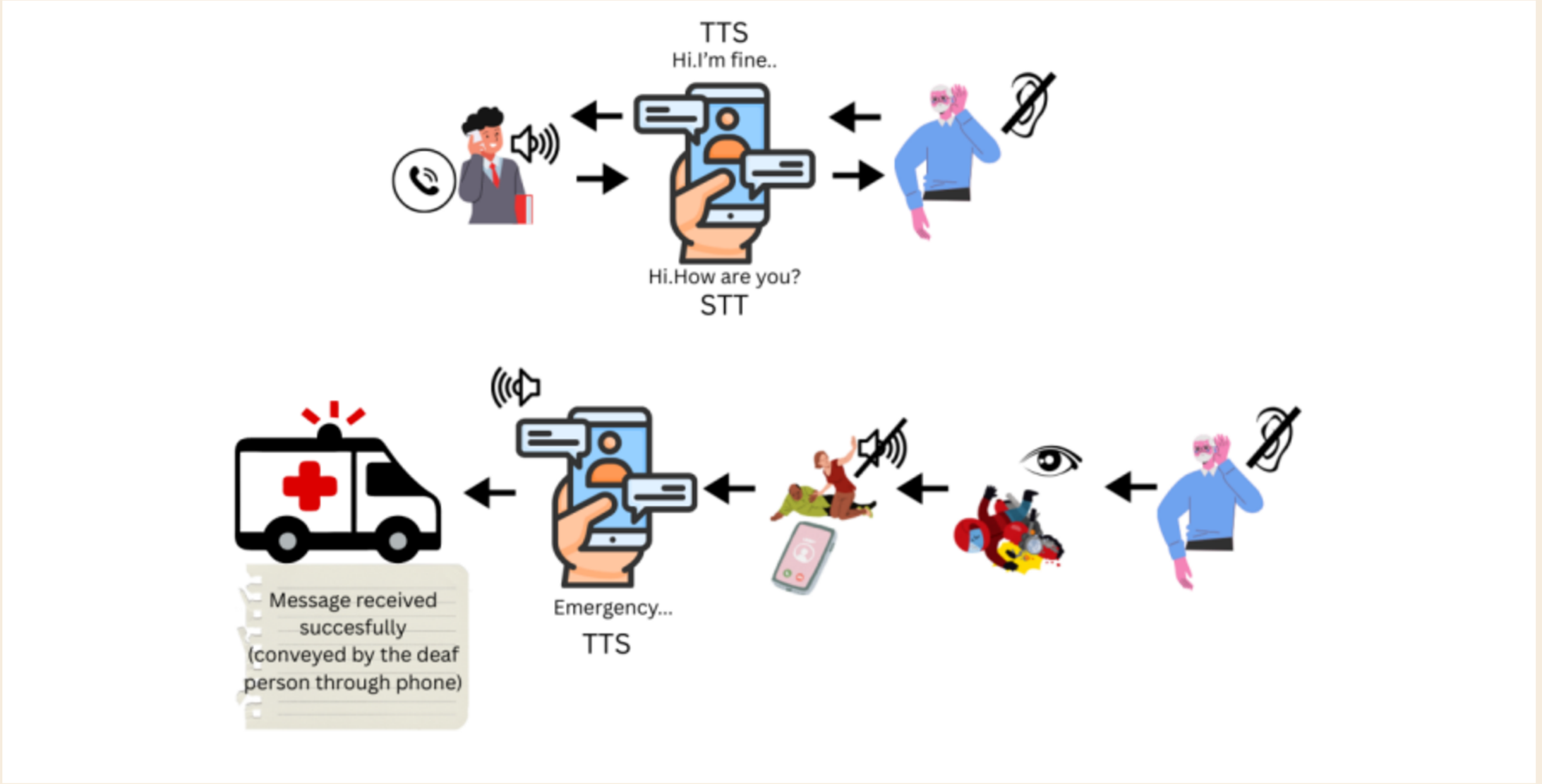
PROBLEM STATEMENT

- 🔇 Phone calls are inaccessible to deaf and hard-of-hearing individuals
- Current solutions often depend on:

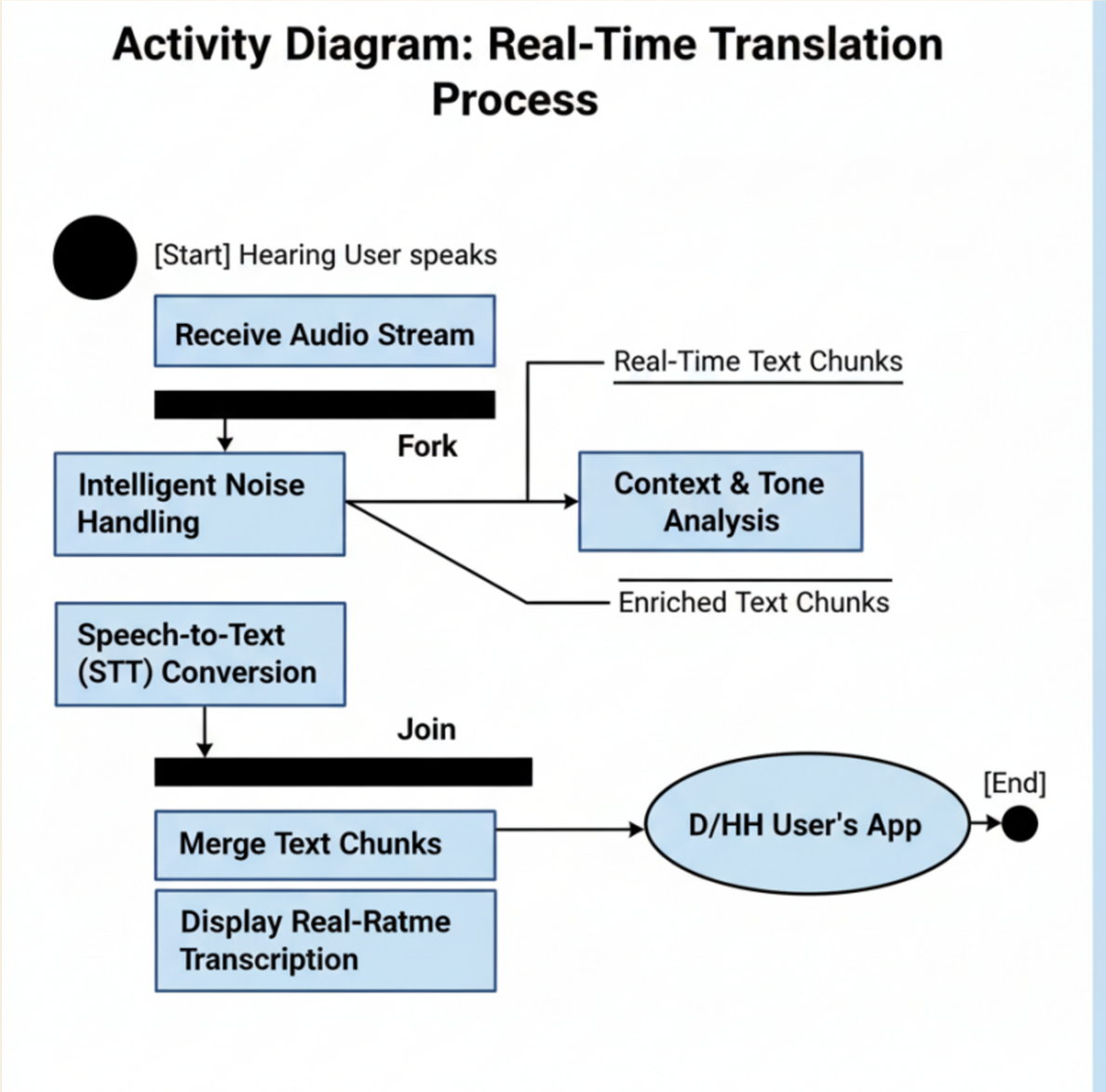
  Human-operated relay services,

  Expensive or specialized assistive devices,
- While Real-Time Text (RTT) is a standardized solution in some countries, its support in India is lacking due to limited carrier integration and lowawareness.
- Additionally, current solutions often rely on human intermediaries or are limited toeither sign language or basic transcription, without providing an end-to-end, AIpowered,

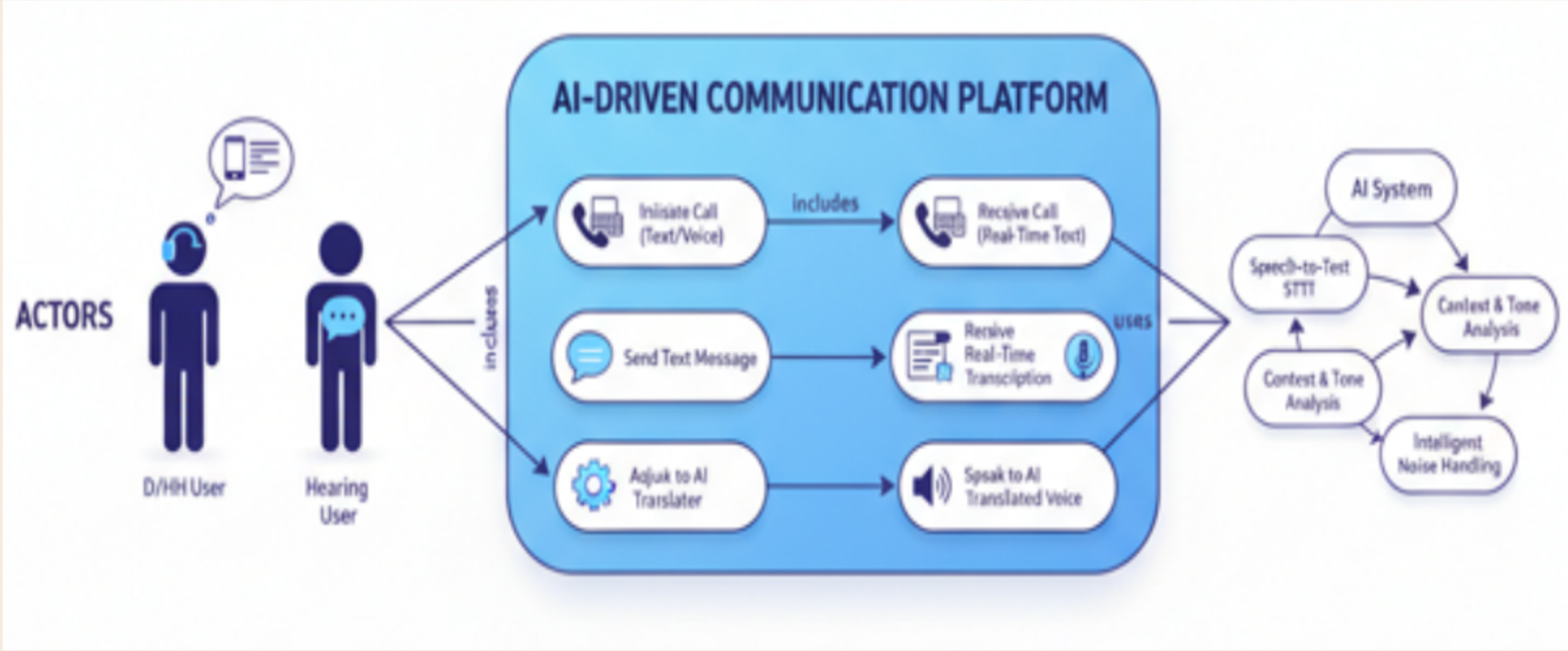
ARCHITECTURE DIAGRAM



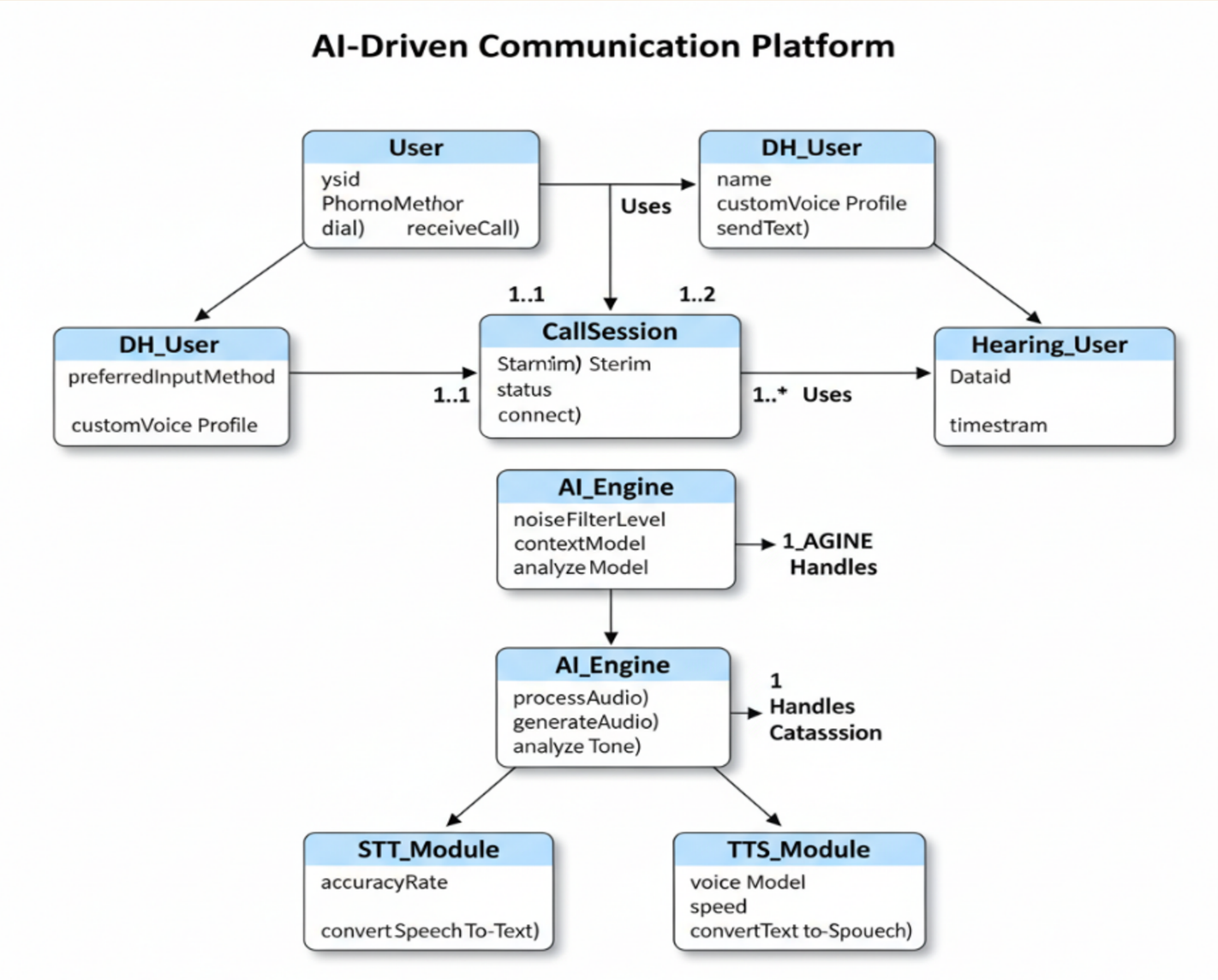
ACTIVITY DIAGRAM



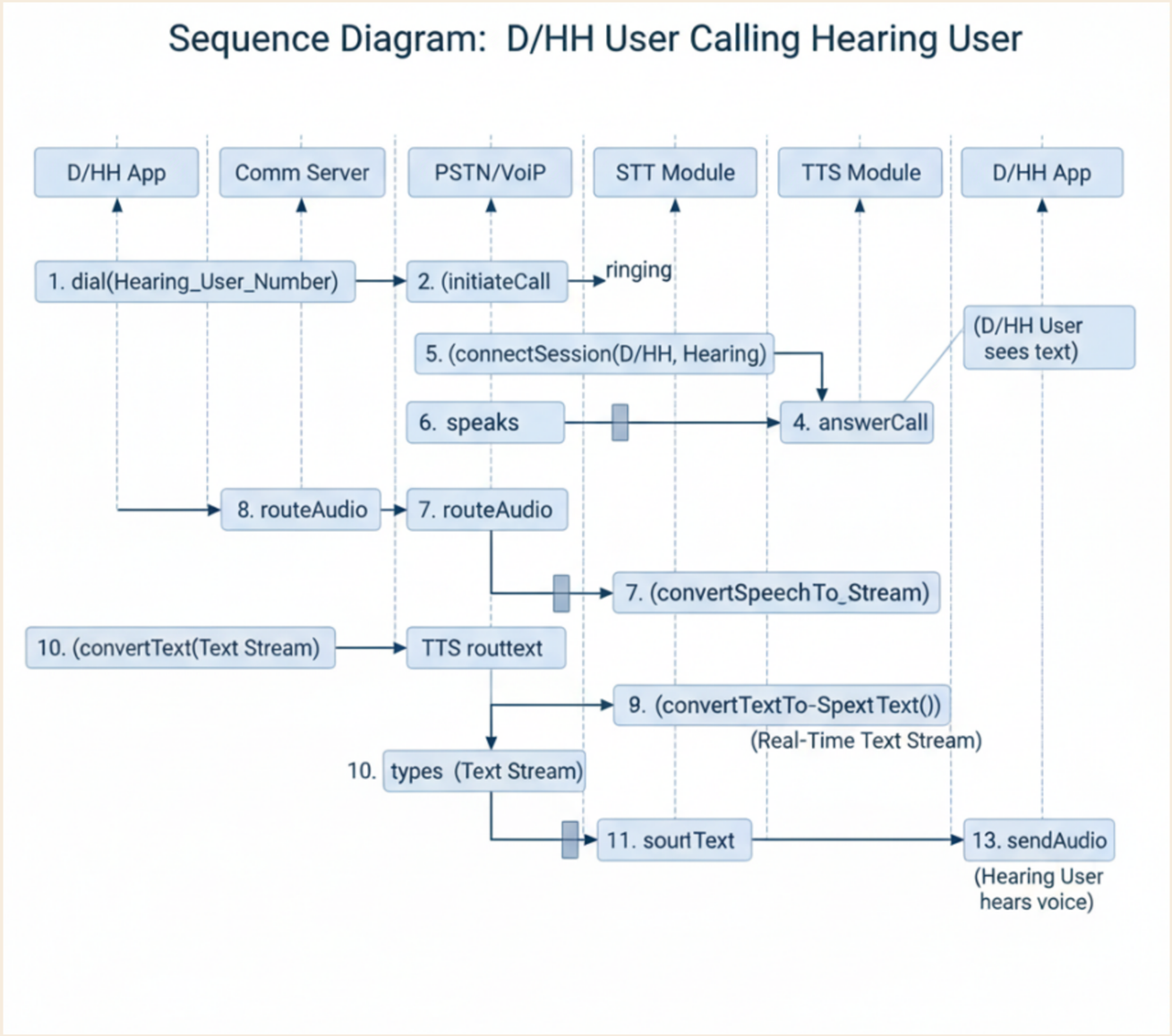
USECASE DIAGRAM



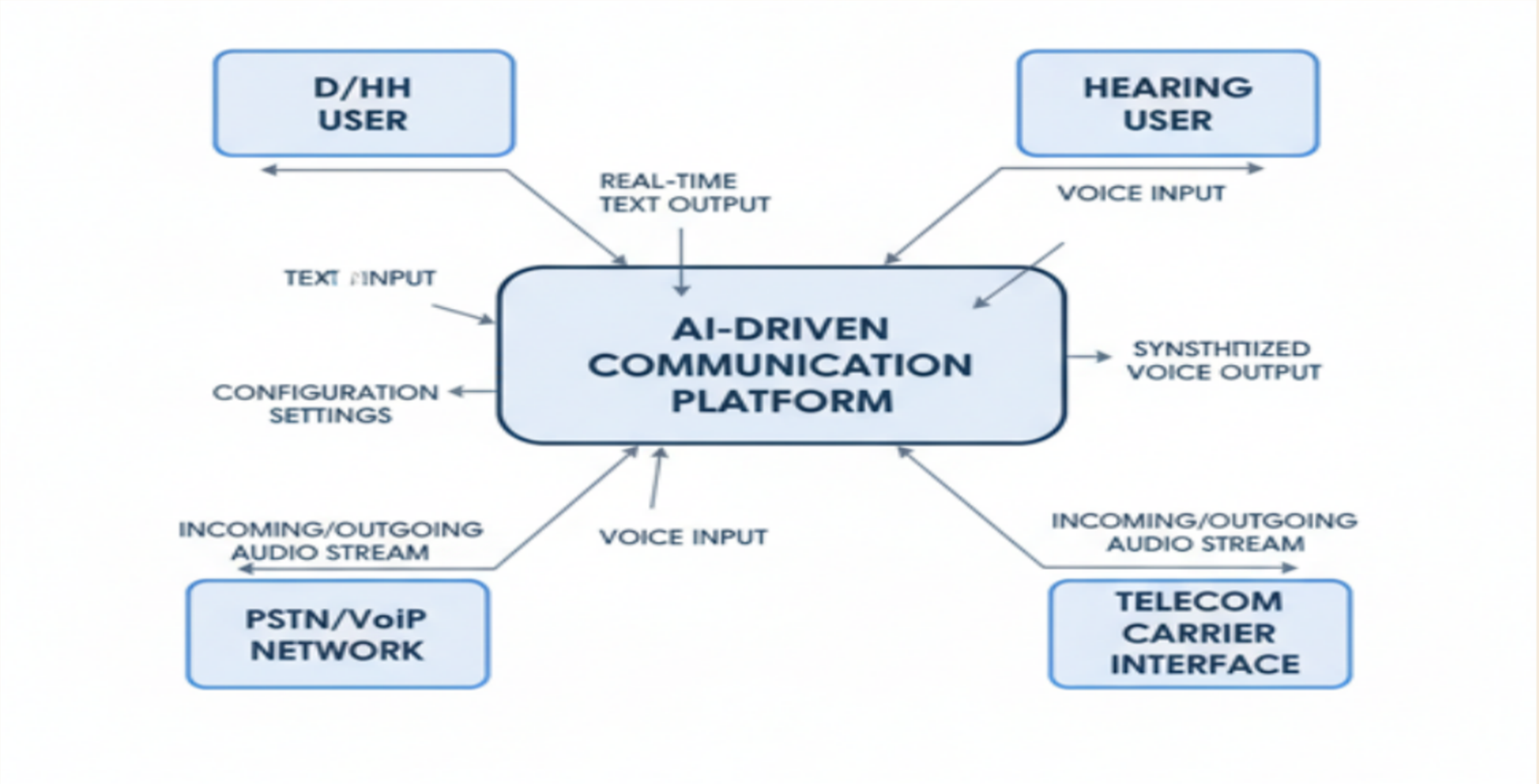
CLASS DIAGRAM



SEQUENCE DIAGRAM



DATA FLOW DIAGRAM



MODULES

1. Incoming Voice
   - Caller speaks via regular phone or VoIP call.
   - System captures audio in real-time.
2. Speech-to-Text Conversion
   - AI speech recognition engine (e.g., Whisper/Google STT) converts audio to live text.
3. Real-Time Display
   - Transcribed text is shown to the deaf user instantly (simulating RTT behavior using WebSocket/WebRTC).

MODULES

4. Deaf User's Response

- User either:

  Types a message (or)

  Uses sign language, which is detected using a camera and computer vision.

5. Text-to-Speech Conversion

- Typed or signed message is converted to speech using TTS and sent to the caller.

6. Two-Way Communication Loop Continues

- The call flows like a normal conversation but with AI bridging both parties.

PROPOSED METHODOLOGY

- AI Speech Recognition (STT): Tools like Whisper or Google STT instantly translate the caller's speech into real-time text, handling diverse accents, languages, and background noise.
- Real-Time Messaging Simulation: Technologies like WebRTC or WebSocket simulate Real-Time Text (RTT), sending messages as they are typed for a natural, low-latency conversational flow.
- Text-to-Speech (TTS) Conversion: TTS engines convert the hearing-impaired user's typed responses into clear, natural-sounding speech for the hearing participant, with options for customized voices and rates.
- Sign Language Recognition (Optional): Computer vision models (e.g., MediaPipe) can be incorporated to interpret sign language gestures, offering an additional, highly inclusive mode of communication.

TEST CASE

1. Real-Time Communication (RTT) & Core Functionality

ID	Test Case Title	Description	Expected Result
RTC-001	**Standard Call Flow Test (DHH User initiating)**	The DHH user initiates a call to a hearing person, types a message, and receives the hearing person's spoken reply as text.	The typed message is sent/received instantly. The spoken reply is transcribed into text and displayed to the DHH user with ≤ 1 second latency.
RTC-002	**Standard Call Flow Test (Hearing User initiating)**	A hearing person calls the DHH user. The platform should automatically answer/route and begin real-time transcription.	The call connects. The hearing person's initial greeting is transcribed and displayed immediately.
RTC-003	**Text-to-Speech (TTS) Latency**	The DHH user types a short, 5-word sentence and sends it.	The TTS engine converts the text to speech and plays it for the hearing person with minimal, non-disruptive delay.
RTC-004	**High-Volume Text Input**	The DHH user types a long paragraph (e.g., 200 characters) rapidly and sends it.	The system should handle the transcription and TTS conversion without freezing, crashing, or significantly increasing the delay. The text should be processed and spoken naturally.

2. AI-Driven Features: STT, TTS, and Intelligent Handling

ID	Test Case Title	Description	Expected Result
AI-001	**Speech-to-Text (STT) Accuracy - Indian Accents**	A hearing user with a common regional Indian accent (e.g., Tamil, Bengali, Hindi-influenced English) speaks a standard phrase.	The STT accurately transcribes the speech with a minimum of 98% word accuracy.
AI-002	**Noise Handling - Background Noise**	A hearing user speaks while there is moderate background noise (e.g., traffic, cafe chatter, fan noise).	The intelligent noise handling/filtering significantly reduces noise interference. The spoken words are clearly transcribed without significant errors from the noise.
AI-003	**Emotion Recognition - Tone Transfer (Urgency)**	A hearing person speaks a sentence with an urgent or distressed tone (e.g., "I need a doctor now!").	The transcription accurately captures the text. Additionally, the system flags or visually highlights the **"Urgent/Distressed"** emotional tone for the DHH user.
AI-004	**Context Preservation - Pronouns/References**	A hearing user speaks a conversation with anaphora (e.g., "I met Rohan yesterday. **He** said the meeting is tomorrow.").	The transcription accurately captures the text. The adaptive language model correctly identifies the reference ("He" refers to "Rohan") and does not
AI-005	**Natural Text-to-Speech (TTS) Output**	The DHH user types an affirmative sentence, a question, and an excited statement.	The TTS output for the hearing person should have natural inflection, varying the tone for the question and the excited statement (not sounding robotic or flat).

3. Platform Integration and Accessibility

ID	Test Case Title	Description	Expected Result
INT-001	**Standard PSTN/Mobile Network Integration**	The DHH user uses the platform to call a standard mobile number and a landline number on different carriers (as available in India).	The call connects seamlessly. The real-time text and speech features function correctly across both network types.
INT-002	**Emergency Call (100/108/112)**	The DHH user initiates an emergency call using the platform.	The call connects to the emergency service. The AI system relays the DHH user's typed messages accurately and urgently to the operator via TTS.
INT-003	**Low Bandwidth Performance**	The call is placed in a region with low mobile data speed (simulated 2G or weak 3G).	The platform remains operational, prioritizing text delivery. Any degradation (e.g., slight increase in latency) is managed gracefully, without crashing or dropping the call.
INT-004	**System Resource Usage (Software-Only)**	The platform is run on a low-to-mid-range smartphone (common in the target demographic).	The application runs smoothly without excessive battery drain or high CPU usage, confirming the "software-only, affordable" requirement.

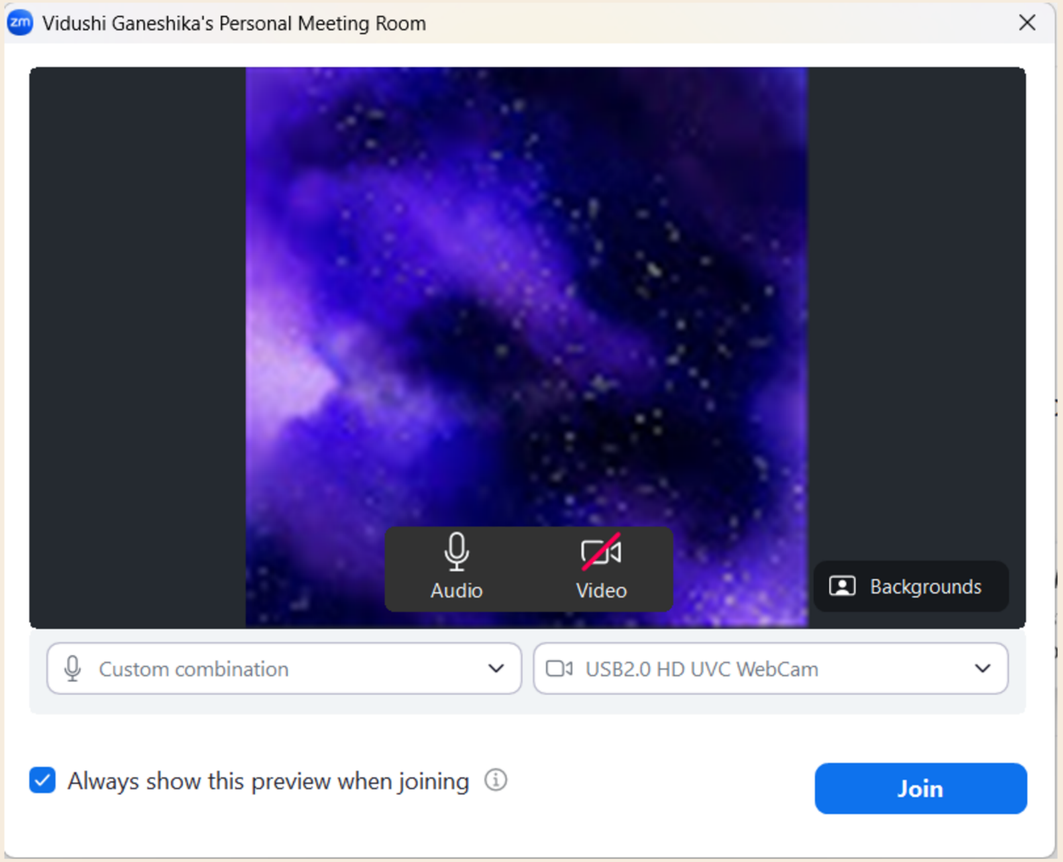
4. Privacy and Security

ID	Test Case Title	Description	Expected Result
PRV-001	**Data Retention Policy Enforcement**	A call is completed, and the DHH user verifies the platform's data storage settings (e.g., no permanent transcript storage by default).	No call data, transcriptions, or audio are stored on the service provider's server, reinforcing the privacy-preserving model and eliminating the human intermediary/operator access.
PRV-002	**End-to-End Encryption**	Monitor network traffic during a call (STT/TTS stream).	All communication streams (voice, text, and transcription data) are encrypted end-to-end between the user's device and the server/recipient.
PRV-003	**Offline/Local Processing Fallback (Partial)**	The phone network is connected, but the internet connection temporarily drops during a call.	The platform attempts to use local/on-device STT for basic transcription features, if supported, to provide minimal interruption. The call itself is not immediately dropped.

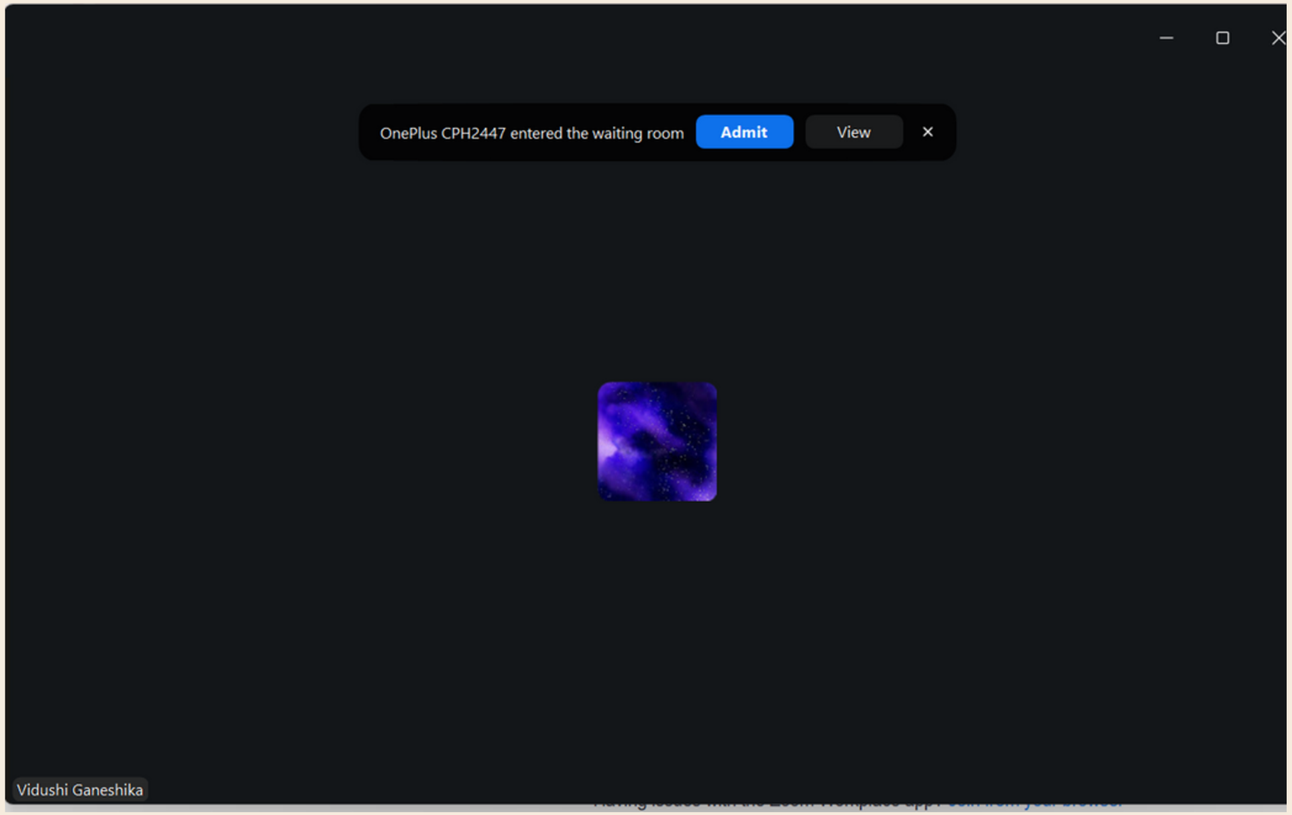
PERFORMANCE ANALYSIS

Module	Tool/ Framework Used	Results (Performance)	Analysis / Remarks
Speech-to-Text (STT)	Whisper, Google STT	Whisper: 95–97% accuracy, Google STT: 90–92%	Whisper performed better in noisy environments; Google STT strong in multilingual support.
Text-to-Speech (TTS)	gTTS, pyttsx3	gTTS: High clarity, limited customization; pyttsx3: Medium clarity, high customization	gTTS best for natural speech; pyttsx3 better for offline use and flexible settings.
Real-Time Messaging	WebRTC, WebSocket, Firebase	WebRTC latency: 120 ms; WebSocket: 150 ms; Firebase: 180 ms	WebRTC provided the lowest latency, making it most suitable for RTT-like communication.
Sign Language Recognition	MediaPipe, TensorFlow	Accurate gesture detection (85–90% in tests)	Useful for inclusivity, but requires good lighting and high-quality camera input.

SCREENSHOTS

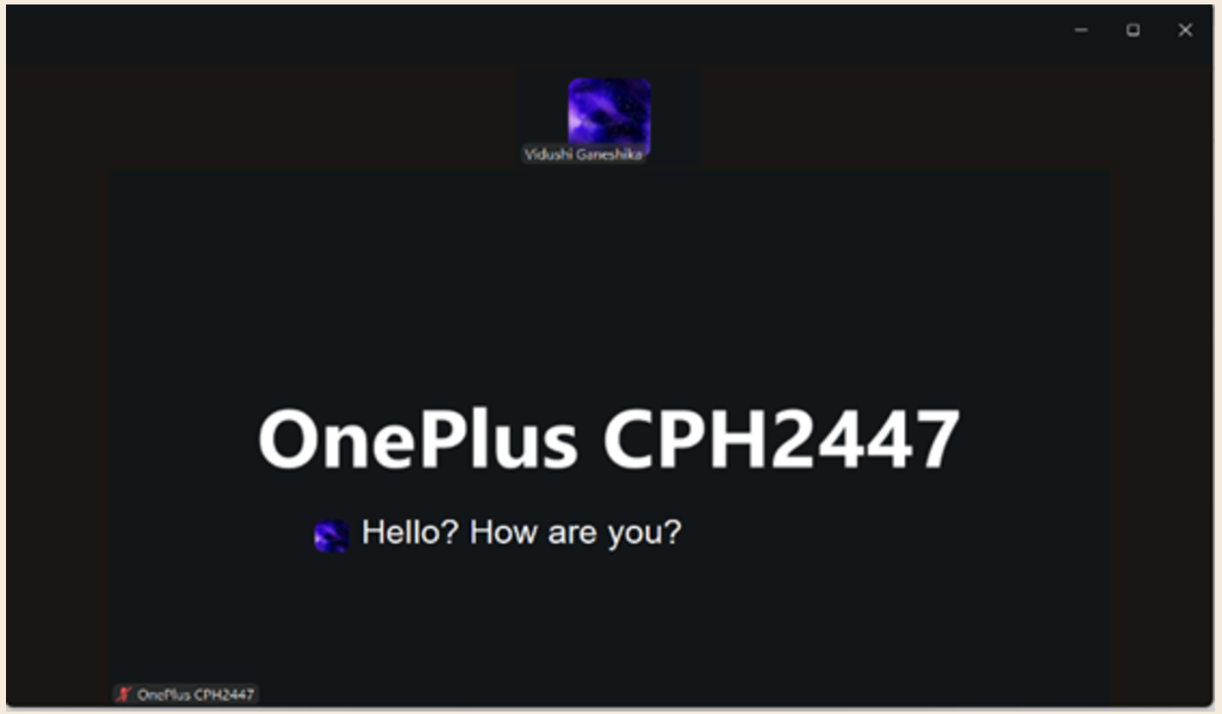


SCREENSHOT OF STARTING A MEETING

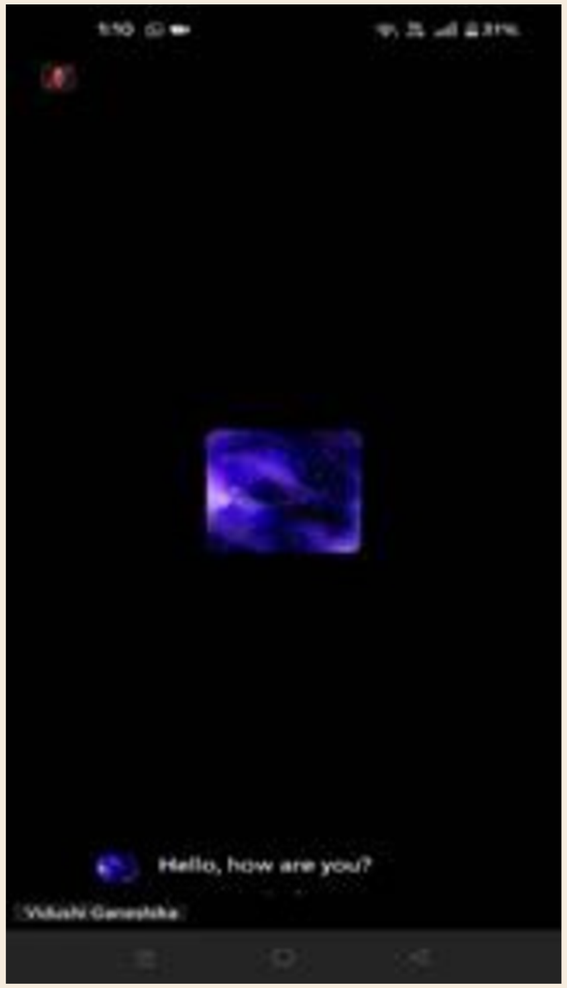


SCREENSHOT OF LETTING ANOTHER PERSON IN THEMEETINGSCREENSHOT

SCREENSHOT OF MEETING IN PROGRESS

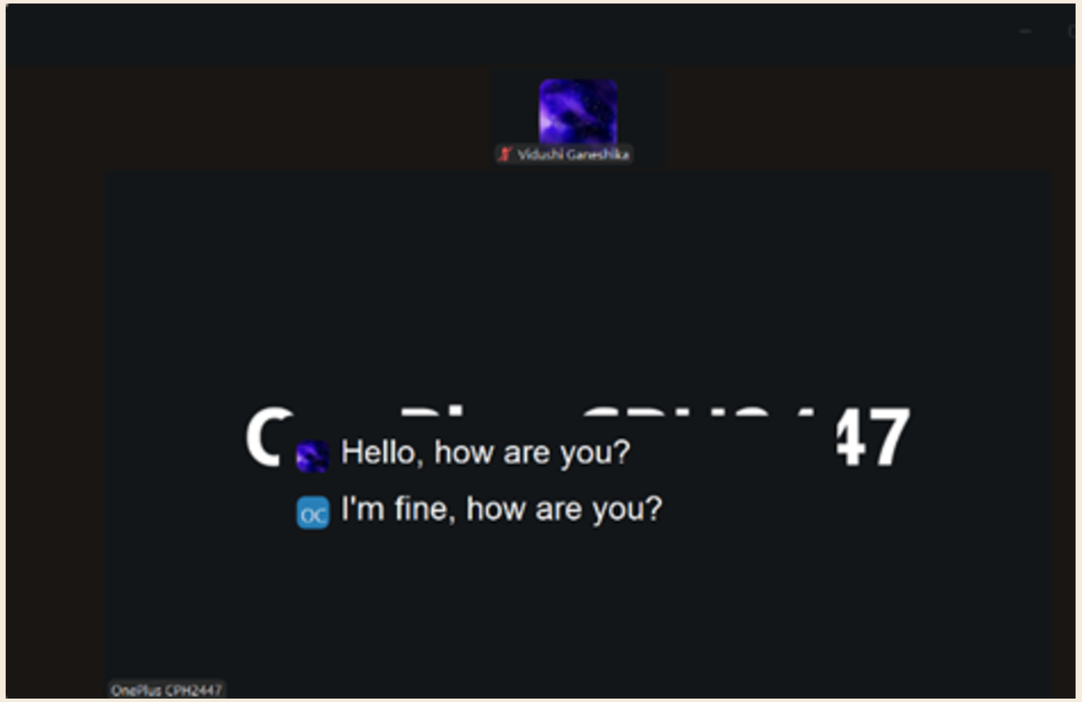


SCREENSHOT OF SPEECH TO TEXTSPEECH

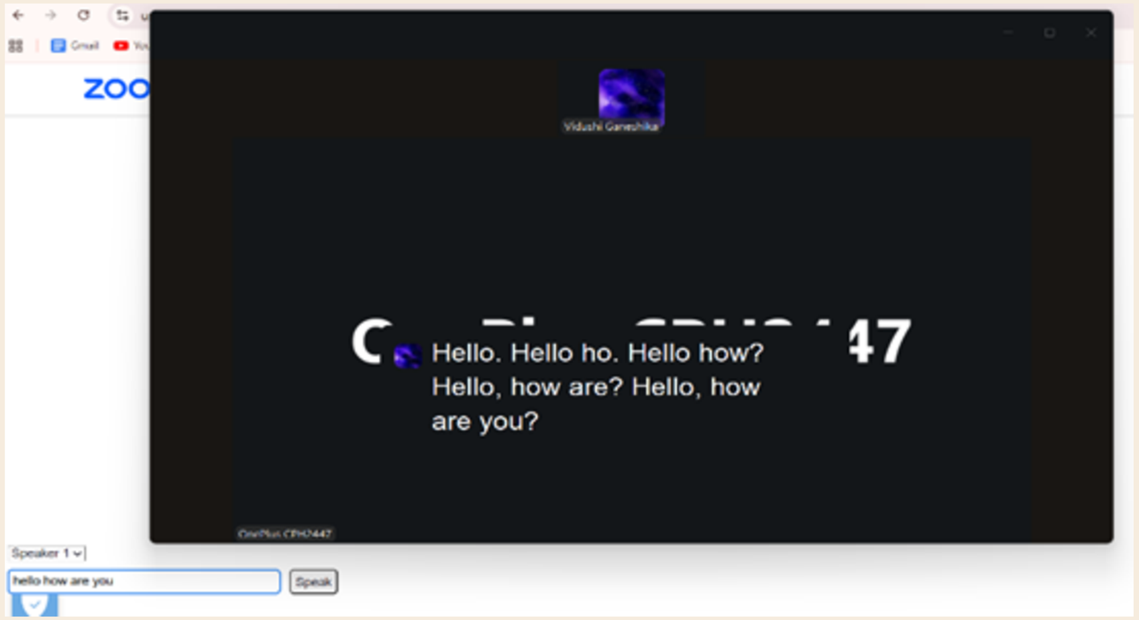


TEXTSPEECH RECEIVED ONOTHER DEVICE

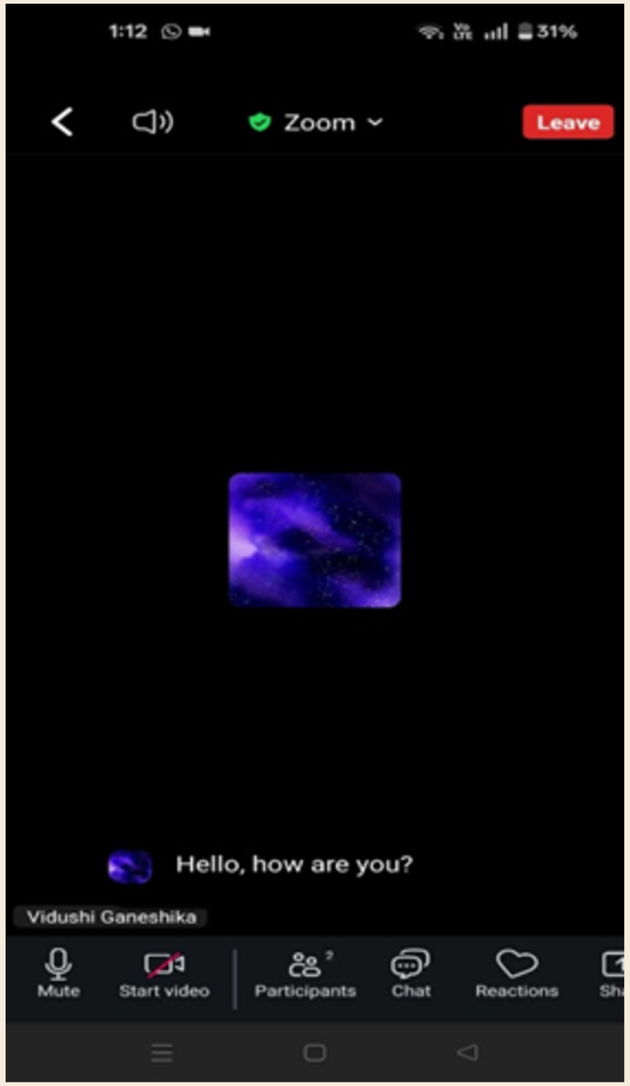
SPEECH RESPONSE REPLIED BY ANOTHER DEVICE



SPEECH RECEIVED ON MAIN DEVICE



TEXT IN CASE IF THE PERSON CAN'T TALK



TEXT CONVERTED TO SPEECH AND RECEIVED ON OTHER DEVICE

CONCLUSION

Inaccessible phone communication continues to create barriers for deaf and hard-of-hearing individuals, as existing solutions either depend on human assistance or require costly devices. This not only limits independence but also reduces the inclusivity of digital communication.

By introducing a simple, affordable, and user-friendly software solution, these challenges can be overcome. Such an approach has the potential to make communication more accessible, ensuring equal opportunities for interaction in both everyday and emergency situations.

FUTURE SCOPE

- 🌐Regional Language Support
- 🔊Voice Emotion Detection
- 🔒Data Privacy & Encryption
- 🧠AI Personalization
- 📞Emergency Services Integration
- 🤝Integration with WhatsApp/Zoom APIs
- 🦾Wearable Device Support

REFERENCES

- https://in.docworkspace.com/d/sIBbeq7lQ84bAwwY?lg=en-US&sa=601.1074&ps=1&fn=05898904.pdf

Real-Time Speech-to-Text (STT)

- https://github.com/KoljaB/RealtimeSTT
- https://github.com/davabase/whisper_real_time
- https://github.com/reriiasu/speech-to-text

Real-Time Sign Language Recognition

- https://github.com/AkramOM606/American-Sign-Language-Detection
- https://github.com/AvishakeAdhikary/Realtime-Sign-Language-Detection-Using-LSTM-Model
- https://github.com/RisheeM/Real-time-sign-language-recognition-system-using-Tensorflow-Keras-OpenCV

- https://github.com/Varshini-E/Real-Time-Recognition-of-Indian-Sign-Language

Sign Language Detection with Object Detection

- https://github.com/youngsoul/sign-language-detection-with-tfod2
- https://www.videosdk.live/developer-hub/ai/speech-to-text-real-time
- https://medium.com/@amirk3321/how-webrtc-and-ai-speech-to-text-are-transforming-online-communication-26f8dd6efc6b
- https://sist.sathyabama.ac.in/sist_naac/aqar_2022_2023/documents/1.3.4/b.e-cse-batchno-40.pdf
- https://cloud.google.com/speech-to-text https://cloud.google.com/text-to-speech

Thank you!

“Communication is the human connection that brings us together — let’s make sure it includes everyone.”