Greet - A Video Conferencing Platform

Team Members:
- Sarthak Singhal – sarthak.2125cse1157@kiet.edu
- Aman Bhatt – aman.2125cse1083@kiet.edu
- Kanishk Chaudhary – kanishk.2125cse1196@kiet.edu

Guide: Mr. Vijay Patidar (Assistant Professor, CSE)
Institution: KIET Group of Institutions, Ghaziabad

PROJECT OUTCOME-
Research Paper
(accepted and submitted)

SDG MAPPING

- SDG 4: Quality Education
- - Supports inclusive learning through multilingual communication.
- - Bridges language gaps in virtual classrooms and online education.
- - Enables accessibility and participation for linguistically diverse students.
-
- SDG 9: Industry, Innovation & Infrastructure
- - AI-driven platform using Google Cloud APIs, Agora SDK, and Firebase.
- - Demonstrates scalable, innovative communication solutions.
- - Encourages development of intelligent multilingual systems.
-
- SDG 10: Reduced Inequalities
- - Empowers users from diverse linguistic backgrounds.
- - Reduces barriers to participation in global conversations.
- - Promotes equity in education, employment, and collaboration.
-
- SDG 17: Partnerships for the Goals
- - Enables international collaboration across industries and academia.
- - Fosters global dialogue through real-time language translation.
- - Builds partnerships by eliminating language-based limitations.

Abstract

- Traditional video conferencing tools struggle with language barriers, limiting effective global communication.
- Greet is an AI-powered platform enabling real-time, cross-lingual speech-to-speech translation.
- Combines:
- - Automatic Speech Recognition (ASR)
- - Neural Machine Translation (NMT)
- - Speech Synthesis (TTS)
- Users can speak in their native language, and others receive translated speech in their preferred language—instantly.
- Designed for diverse settings: from technical meetings to casual chats.
- Enhances accessibility with live transcription and customizable language settings.
- Integrates with existing tools for seamless adoption.

Problem Statement

- Existing platforms lack real-time multilingual support, limiting their accessibility for non-native speakers.
- Language barriers hinder collaboration in international meetings, classrooms, and content creation.
- Most systems focus on text-based translation and fail to offer natural speech-to-speech interaction.
- Background noise and network issues further reduce the clarity and efficiency of communication.
- There's a pressing need for a solution that enables inclusive, context-aware, and real-time voice translation in virtual environments.

Objective

- Develop a real-time voice-to-voice translation system to eliminate language barriers in virtual meetings.
- Integrate automatic speech recognition (ASR), neural machine translation (NMT), and text-to-speech (TTS) for seamless communication.
- Ensure multilingual accessibility with customizable language settings for both input and output.
- Provide live transcription for enhanced accessibility, including support for hearing-impaired users.
- Maintain low latency and high translation accuracy, even in noisy or low-bandwidth environments.
- Support secure and user-friendly meetings using tools like Firebase, Agora SDK, and Google Cloud APIs.

Literature Review

- WaveNet (Oord et al.)
- Introduced deep generative models for audio synthesis, improving speech naturalness and quality over traditional TTS systems.
- Tacotron & Deep Voice (Wang et al., Arik et al.)
- Enabled expressive and high-quality speech synthesis directly from text using sequence-to-sequence models and attention mechanisms.
- Stream-Speech (Zhang et al.)
- Demonstrated real-time speech-to-speech translation using multi-task learning, reducing translation latency and improving accuracy.
- Conformer Model (Gulati et al.)
- Hybrid architecture combining CNN and Transformer, improving speech recognition through better modeling of acoustic patterns and context.
- Multi-Speaker Neural TTS (Deng et al.)
- Enabled adaptive speaker embeddings and real-time multilingual voice synthesis for dynamic conversational translation.

System Architecture

- Automatic Speech Recognition (ASR)
- ➤ Captures spoken input and converts it into text in real-time.
- ➤ Robust against background noise using AI-based noise suppression.
- Neural Machine Translation (NMT)
- ➤ Translates transcribed text into the desired target language.
- ➤ Uses attention mechanisms for context-aware, fluent translation.
- Text-to-Speech (TTS) Synthesis
- ➤ Converts translated text back into natural-sounding speech.
- ➤ Uses models like Tacotron & WaveNet for human-like audio output.
- User Interface Layer
- ➤ Includes meeting features: join, schedule, view recordings.
- ➤ Language selection, live transcription, and sentiment-aware feedback.
- Security & Performance Enhancements
- ➤ MFA (Multi-Factor Authentication), voice/facial recognition.
- ➤ Real-time analytics panel (latency, bandwidth, quality stats).
- ➤ Adaptive feedback loop for continuous learning and improvement.

Model Comparison – Technical

Feature	SMT	NMT	Hybrid (SMT + NMT)
Translation Approach	Phrase based, probabilistics	Deep learning with attention	Combined SMT rules + NMT fluency
Context Understandiing	Limited	Strong	Enhanced through NMT integration
Fluency	Moderate	High	Balanced, context-aware
Accuracy	Domain-dependent	General-purpose, high	Optimized for real-time
Latency	High	Low (real-time capable)	Optimized for real-time
Adaptability	Good for specific domains	Broad multilingual support	Best of both worlds
Data Requirement	Moderate	High	Moderate to High
Use Cases	Rare language pairs	Common/global language sets	Real-time multilingual platforms

Results Summary

Platform Performance Highlights:

- Real-Time Translation Accuracy: High-quality, low-latency speech-to-speech translation in multiple languages.
- Noise Handling: AI-powered suppression ensures clarity even in noisy environments.
- Multilingual Accessibility: Dynamic language selection and interface localization.
- User-Friendly Design: Intuitive UI with easy access to meeting tools, settings, and recordings.
- Performance Monitoring: Live call stats dashboard with latency, jitter, and packet loss metrics.
- Recording & Review: Auto-saved sessions for future playback and documentation.

Demonstration Insight:

- Audio frequency analysis showed clear adaptation to linguistic differences (e.g., English vs. Hindi stress/rhythm patterns).
- Real-time processing maintained natural flow across languages.

Conclusion

- Greet revolutionizes multilingual video conferencing with real-time speech-to-speech translation.
- Utilizes advanced tools like Google Cloud API, Agora SDK, and Firebase for seamless integration and scalability.
- Empowers inclusive communication across languages, cultures, and regions—no human translators required.
- Successfully addresses common challenges:
- Language barriers
- Background noise
- Usability across diverse user groups
- Future Enhancements:
- Smarter noise reduction
- Expanded dialect and accent support
- Edge computing for lower latency

"Greet bridges linguistic gaps and brings people closer in the digital world."

Thank You