A

Project Report

on

Greet – A Video Conferencing Platform

submitted as partial fulfillment for the award of

BACHELOR OF TECHNOLOGY DEGREE

SESSION 2024-25

in

Computer Science and Engineering

By

Sarthak Singhal (Roll No.: 2100290100148)
Aman Bhatt (Roll No.: 2100290100023)
Kanishk Chaudhary (Roll No.: 2100290100078)

Under the supervision of

Mr. Vijay Patidar

KIET Group of Institutions, Ghaziabad

Affiliated to

Dr. A.P.J. Abdul Kalam Technical University, Lucknow

(Formerly UPTU)

May, 2025

DECLARATION

We hereby declare that this submission is our own work and that, to the best of our knowledge and belief, it contains no material previously published or written by another person nor material which to a substantial extent has been accepted for the award of any other degree or diploma of the university or other institute of higher learning, except where due acknowledgment has been made in the text.

Signature

Names: Sarthak Singhal

Roll No.: 2100290100148

Signature

Names: Aman Bhatt

Roll No.: 2100290100023

Signature

Names: Kanishk Chaudhary

Roll No.: 2100290100078

Date:

CERTIFICATE

This is to certify that Project Report entitled “Greet – A Video Conferencing Platform” which is submitted by Student name in partial fulfillment of the requirement for the award of degree B. Tech in Department of Computer Science & Engineering of Dr. A.P.J. Abdul Kalam Technical University, Lucknow is a record of the candidates own work carried out by them under my supervision. The matter embodied in this report is original and has not been submitted for the award of any other degree.

.

Mr. Vijay Patidar **Dr. Vineet Sharma**

(Assistant Professor) **(Dean CSE)**

Date:

ACKNOWLEDGEMENT

It gives us a great sense of pleasure to present the report on the B. Tech Project undertaken during B. Tech Final Year. We owe special debt of gratitude to our project guide Mr. Vijay Patidar, Department of Computer Science & Engineering, KIET, Ghaziabad, for his constant support and guidance throughout the course of our work. His sincerity, thoroughness and perseverance have been a constant source of inspiration for us. It is only his cognizant efforts that our endeavors have seen light of the day.

We also take the opportunity to acknowledge the contribution of Dr. Vineet Sharma, Dean of Computer Science & Engineering, KIET, Ghaziabad, for his full support and assistance during the development of the project. We also do not like to miss the opportunity to acknowledge the contribution of all the faculty members of the department for their kind assistance and cooperation during the development of our project.

We also do not like to miss the opportunity to acknowledge the contribution of all faculty members, especially faculty/industry person/any person, of the department for their kind assistance and cooperation during the development of our project. Last but not the least, we acknowledge our friends for their contribution in the completion of the project.

Signature
Name: Sarthak Singhal
Roll No.: 2100290100148

Signature
Name: Aman Bhatt
Roll No.: 2100290100023

Signature
Name: Kanishk Chaudhary
Roll No.: 2100290100078

Date:

ABSTRACT

Greet is more than just a translation tool — it's your voice, understood anywhere in the world. Whether you're tuning into a meeting from Tokyo or taking an online class from São Paulo, Greet makes sure you're heard — and understood — in your own language. It breaks down language barriers so people can connect, collaborate, and communicate freely, no matter where they are.

At its core, Greet is a real-time, speech-to-speech multilingual translation platform that works effortlessly with today's video conferencing tools. It uses advanced technology like Automatic Speech Recognition (ASR), Neural Machine Translation (NMT), and Text-to-Speech (TTS) to make live conversations flow smoothly. You speak, it listens. It translates your words and then speaks them out loud in another language — all in real time, and with a natural-sounding voice that feels human, not robotic.

But Greet isn't just about translating words. It's about making communication easier in real-world situations. For instance, let's say you're joining a meeting from a noisy café or a home with kids playing in the background — no problem. Greet comes with noise suppression, so your voice stays clear even if your surroundings aren't.

There's also live transcription built right in. While you're talking, your words appear as subtitles on the screen — in the language of your choice. That means even if you can't hear well at the moment, or you're in a quiet place and need to mute audio, you can still follow the conversation just by reading.

Greet is built with flexibility in mind. You can set your preferred language for speaking and listening, choose specific dialects, or let the system automatically detect the language you're using. That's especially helpful in meetings with participants from different regions or linguistic backgrounds. It takes the guesswork out of multilingual conversations.

Security is a top priority, too. Greet uses strong multi-factor authentication (MFA) and end-to-end encryption, so your conversations stay private and protected. If you're using Greet in a company or classroom, you'll appreciate the role-based access controls — they make sure that only the right people can access important features and content.

Another smart feature is Greet's built-in analytics. After a meeting, you can see insights like how long it lasted, who spoke the most, which languages were used, and more. It's a helpful way to keep track of engagement, especially for educators, team leaders, and anyone managing remote communication.

And here's the best part: Greet is designed to work everywhere. Whether you're on a laptop, desktop, tablet, or smartphone, the experience is smooth and reliable. As more people join and more languages are added, the platform continues to perform — no lag, no drop in quality.

Greet is built for all kinds of people and situations. It's ideal for international business meetings, virtual classrooms, global freelance projects, creative content creators, and multinational teams. No matter your role, Greet helps you stay connected with the people who matter.

At the end of the day, Greet isn’t just a piece of software — it’s a bridge. A bridge between people, cultures, and ideas. It brings us closer together in a world where language should never be a barrier to connection.

TABLE OF CONTENTS

Page No.

LIST OF FIGURES AND TABLES

LIST OF ABBREVIATION

AI	Artificial Intelligence
ASR	Automatic Speech Recognition
CNN	Convolutional Neural Network
MNCs	Multinational Companies
NMT	Neural Machine Translation
SMT	Statistical Machine Translation
TTS	Text-to-Speech

CHAPTER 1
INTRODUCTION

In today's rapidly evolving digital landscape, the ability to communicate across language and cultural boundaries is more crucial than ever. Whether it's fostering international collaboration, conducting global business, or sharing ideas across diverse communities, effective communication plays a foundational role. Technology has made connecting easier and more immediate, especially with the rise of video conferencing platforms. This shift became even more pronounced during the COVID-19 pandemic, when video calls transformed from a convenience into an essential tool for work, education, and maintaining social ties. However, despite these advances, most video conferencing platforms still fall short when it comes to supporting multilingual conversations in real time. This limitation creates significant obstacles for participants from different linguistic backgrounds, hampering clear and smooth communication.

Research by Aksenova et al. highlights how the lack of adequate multilingual support remains a persistent challenge in current video conferencing systems, often leading to miscommunication, delays, and user frustration. Recognizing this gap, we developed **Greet**—a comprehensive and user-friendly video conferencing platform designed to seamlessly integrate real-time speech-to-speech translation. By combining cutting-edge technologies such as automatic speech recognition (ASR) and neural machine translation (NMT), Greet enables highly accurate, natural, and fluid live translations, helping users engage effortlessly despite language differences.

The importance of integrating speech recognition with translation technologies has been well documented by experts like Zhang et al., who underscore how this combination can dramatically improve cross-lingual interactions, reducing misunderstandings and making conversations feel more natural. Greet takes this insight and puts it into practice, creating a platform that listens, understands, translates, and speaks—all in real time—allowing participants to communicate as if they were speaking the same language.

But Greet goes beyond just translation. We designed it with accessibility and user experience in mind. Real-time transcription displays the spoken content visually, providing essential support for individuals who are hard of hearing or who prefer reading along with audio. Users can easily customize their preferred input and output languages, ensuring each person enjoys a personalized and comfortable communication experience. Additionally, Greet's compatibility

with popular existing video conferencing tools means users can adopt the platform without overhauling their current workflows, allowing for a smooth transition and wider accessibility.

Security and privacy have been prioritized throughout the platform's development. Leveraging Firebase's robust authentication and data management services, Greet ensures that user data and conversations remain secure and confidential, instilling trust in all participants. The system is also designed for scalability, so it can grow with its user base without compromising performance or reliability.

The practical applications of Greet are broad and impactful. In international business settings, teams can collaborate more efficiently when language is no longer a barrier. Educational institutions can create inclusive virtual classrooms where students from various linguistic backgrounds learn together seamlessly. In healthcare, accurate communication between doctors and patients speaking different languages can improve outcomes and patient satisfaction. Content creators and freelancers can expand their reach globally, and customer support teams can offer assistance in multiple languages without needing extensive multilingual staff.

Moreover, Greet's design considers the social and emotional nuances of conversation, which are often lost in translation. Language is more than just words—it carries culture, tone, and context. Our platform uses advanced natural language processing techniques to preserve meaning and intent as much as possible, so conversations don't just translate literally but also feel authentic and respectful. This focus on nuance helps build trust and connection, which are essential for any meaningful interaction.

From a technical perspective, Greet's architecture is built to handle the demanding needs of real-time communication. The platform uses cloud-based processing to ensure that heavy computational tasks like speech recognition and translation happen quickly and accurately, minimizing any lag or delay. This means that users experience conversations that feel immediate and natural, rather than disjointed or robotic.

The user interface is another crucial aspect of Greet's success. We prioritized simplicity and ease of use so that people of all technical backgrounds can benefit from the platform. Whether you're a busy professional juggling multiple calls or someone new to video conferencing technology, Greet's intuitive design helps you get started quickly and communicate effectively. Features like automatic language detection, one-click mute and unmute, and real-time subtitles make it easy to stay engaged and focused on the conversation.

In addition, Greet supports a wide range of languages, including widely spoken global languages as well as regional dialects and less commonly supported tongues. This inclusivity allows more people to participate in global conversations and ensures that no one is excluded due to language limitations.

Looking ahead, the future of Greet is bright and full of exciting possibilities. We plan to incorporate features like emotion detection, which will analyze vocal tone and facial expressions to better understand a speaker's feelings and intentions. This would allow the platform to provide more context-aware translations and even offer suggestions for clearer communication. Gesture recognition is another planned enhancement, enabling users to communicate non-verbally in ways that transcend language.

Ultimately, Greet is not just a tool; it's a bridge between cultures, a facilitator of understanding, and a catalyst for collaboration. In a world that is more connected than ever, breaking down language barriers is key to unlocking the full potential of human interaction. Greet strives to make every conversation inclusive, accessible, and meaningful, because communication is the foundation of all relationships, growth, and progress.

1.1 Project Description

To bring real-time speech translation with text message display to life, our system relies on a carefully chosen set of powerful technologies that work together seamlessly behind the scenes. At the heart of this solution is **Flutter**, a versatile front-end framework known for creating beautiful, fast, and responsive apps that work smoothly on both mobile devices and desktops. We chose Flutter because it allows us to build an intuitive and user-friendly interface that feels natural to use, no matter what device you're on. The programming language powering this framework is **Dart**, which pairs perfectly with Flutter to enable rapid development and easy maintenance.

For the actual real-time audio and video communication, we depend on the **Agora SDK**. Agora is a trusted name in the field of live streaming and video conferencing, and it provides the essential backbone that captures and transmits the user's voice and video without lag or loss of quality. When you speak during a video call, Agora SDK is responsible for picking up your audio clearly and delivering it instantly to the system so that it can be processed in real time.

Once the audio stream reaches the backend, the magic of converting spoken words into text happens thanks to **Google Speech-to-Text**. This service is incredibly advanced and can recognize speech from various accents, dialects, and even noisy environments. It takes your voice and transcribes it into written text with impressive accuracy and speed. This transcription is the crucial first step in the translation process, transforming your spoken language into a format that a computer can understand and work with.

Following speech recognition, the text is handed over to **Google Cloud Translate API**. This API is a powerhouse of machine translation technology, using sophisticated neural machine translation models to convert the recognized text from the speaker's language into the target

language chosen by the listener. The goal here is not just to translate words literally but to preserve the meaning, tone, and context so that conversations feel natural and engaging. Google Cloud Translate supports a wide range of languages and continually improves its accuracy based on real-world usage, which means users get better and more nuanced translations over time.

Meanwhile, to keep the system organized, secure, and synchronized, we utilize **Firebase**, a comprehensive platform by Google that handles several key functions. Firebase manages user authentication, ensuring that each participant is verified and their data remains private and secure. It also powers the real-time database that keeps translated messages and transcriptions flowing instantly to all users involved in the call. This means that as soon as the system translates a piece of speech, the text is immediately pushed to every participant's screen, allowing everyone to read along or catch anything they might have missed. Firebase also helps manage message storage and retrieval, making sure conversations are smooth and consistent across different devices.

All these components—the Flutter app, Agora SDK, Google Speech-to-Text, Google Cloud Translate API, and Firebase—work together like a well-rehearsed orchestra. When a user speaks, their audio travels through Agora to the backend, is transcribed by Google Speech-to-Text, translated by Google Cloud Translate, and then simultaneously displayed on screen and sent back as speech via text-to-speech synthesis (handled by another service). The user interface built with Flutter then displays the translated text clearly and in real time, giving users a seamless experience where language barriers feel like they've simply melted away.

This integration allows our platform not only to provide accurate and fast translations but also to do so in a way that feels natural and accessible. Whether you're a business professional connecting with overseas colleagues, a student attending an international class, or just chatting with friends across the globe, the system is designed to make communication effortless, enjoyable, and inclusive.

Once the speech is converted into text, the next crucial step is translation. The text is sent over to the **Google Cloud Translate API**, a powerful and intelligent service designed to convert text from one language into another almost instantly. This API doesn't just swap words one-for-one; instead, it uses advanced neural machine translation technology, which helps maintain the original meaning, tone, and context of what was said. This ensures that the translated message sounds natural and accurate, making conversations smooth and understandable across different languages.

After the translation is complete, the translated text is quickly sent back to the user's app. This allows the app to display the translated words on the screen in real-time, so participants can read along as the conversation happens. This immediate visual feedback is especially helpful for

users who might have hearing difficulties, are in a noisy environment, or simply prefer reading the translation alongside hearing it. It helps ensure that everyone stays on the same page, no matter the language they speak.

But the experience doesn't stop there. To make communication even more natural and immersive, the system also uses **Google's Text-to-Speech (TTS)** technology. The translated text is sent to this service, which converts it back into spoken words using voices that sound clear, natural, and human-like. This audio output is then played back to the user in their chosen language, giving the impression of a live conversation without any language barriers. Users can even choose from different voice options, such as male or female voices, to make the experience feel more personalized and comfortable.

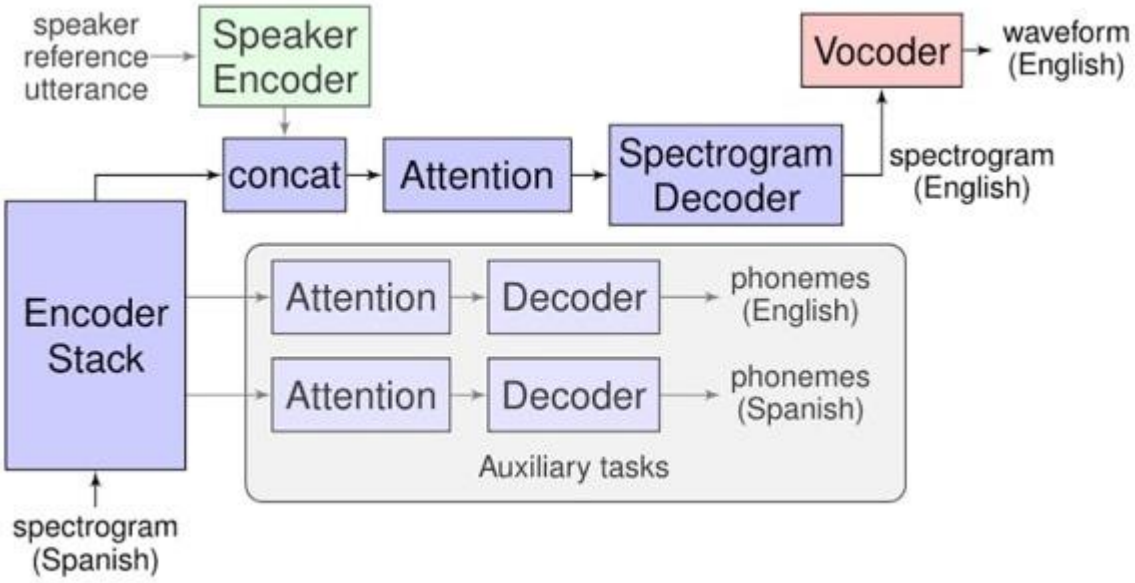
To keep everything running smoothly and synchronized across all users in the session, the system relies on **Firebase**, a real-time database and cloud service. Firebase acts like the glue that holds all the pieces together, sending translated text and messages instantly to every participant's device. This means when one person speaks and their words are translated, the message doesn't just appear on their own screen—it shows up on everyone else's screens at the same time. This real-time synchronization is essential for maintaining the flow of conversations without awkward delays or confusion.

Firebase also plays a key role in managing user sessions and security. Using **Firebase Authentication**, the system makes sure that each user is properly identified and authenticated before joining a call. This protects user privacy and keeps the conversation secure, which is especially important in professional or sensitive settings. By handling authentication and database management, Firebase allows the app to focus on delivering a seamless and reliable communication experience.

On the front end, the entire user interface is managed by the **Flutter** app, which provides a clean and intuitive environment for users to interact with. From the moment someone logs in, the app lets them easily select their source language (the language they'll speak) and their target language (the language they want to hear or read). The translated text then appears clearly on the screen, updated live as the conversation progresses. This thoughtful design makes the technology accessible to people of all ages and tech backgrounds, allowing users to focus on their conversation without worrying about complex settings or technical glitches.

All these components work together behind the scenes to create an effortless, natural experience where language differences no longer pose a barrier. By combining real-time translation, synchronized text display, natural-sounding speech synthesis, and secure user management, the system enables users from anywhere in the world to connect, understand, and engage meaningfully with each other.

An illustration of the setup and call flow can be seen in Fig. 1



Speech Recognition is a key part of our system, and for this, we rely on **Google Speech-to-Text** technology. This service is highly advanced and trusted worldwide for converting spoken language into written text accurately and quickly. Specifically, Google provides acoustic models for different languages—in our case, English and Spanish—that have been trained on massive amounts of audio data from diverse sources. This extensive training helps the system recognize a wide variety of voices, accents, dialects, and even different speaking speeds, making it reliable for users from all walks of life.

What makes Google Speech-to-Text particularly impressive is that it uses **deep learning**, a form of artificial intelligence that mimics how humans learn patterns. Because of this, the system can understand natural speech even in challenging conditions, such as noisy environments or when people speak with strong regional accents. This means that when someone talks during a video call on our platform, their words are captured and transcribed with remarkable accuracy, no matter where they're from or how they speak.

In our setup, the first step happens on the user's device. When a person speaks, their voice is captured by the **Agora SDK**, which is responsible for managing real-time audio and video communication. Agora acts as the messenger, streaming the audio smoothly and without delay to the backend servers. The importance of Agora lies in its ability to handle live conversations efficiently, ensuring minimal lag and high-quality sound transmission.

Once the audio reaches the backend, it's processed by the **Google Speech-to-Text API**. This API takes the audio stream and immediately starts converting the spoken words into text, sentence by sentence, in real-time. As the conversation flows, the system continuously updates the transcription, keeping pace with the speaker. This ongoing transcription is essential because it forms the foundation for the next stages—translation and speech synthesis.

An important detail is that our implementation does not require any additional training or customization of the speech recognition models. The Google pre-trained models are already fine-tuned to work well across many scenarios, so we can simply plug them into our system and start getting accurate results right away. This saves a lot of time and resources, and it ensures consistent performance no matter the user.

The transcribed text is more than just words on a screen—it's the bridge that connects spoken communication to translation. Because the text is generated almost instantly, it can be fed directly into the machine translation component of the system, which then converts it into the desired language. This seamless handoff from speech to text to translation is what allows our platform to provide smooth, real-time multilingual conversations.

Moreover, the system's ability to recognize speech accurately even with background noise or overlapping voices means that users don't have to worry about speaking perfectly or finding quiet rooms. The technology adapts to real-world situations, making communication feel natural and effortless.

In summary, by combining the robust audio capture capabilities of Agora SDK with the intelligent, AI-powered transcription of Google Speech-to-Text, our platform delivers a speech recognition experience that is fast, reliable, and highly accurate. This strong foundation sets the stage for all the other exciting features of our system, helping break down language barriers and bring people closer together through clear and meaningful conversations.

Segmentation is a crucial yet often overlooked part of how real-time speech translation systems work. When someone speaks continuously during a conversation, their words come out as one long stream of sound. But for us humans—and for any good translation system—to understand that stream properly, it needs to be broken down into smaller, meaningful pieces, or "segments." These segments help clarify the speaker's intended pauses, breaths, and the natural rhythm of language, which are vital for both reading the text and for producing clear, natural-sounding translations.

In our system, we use the **Google Speech-to-Text API** not only to transcribe speech but also to help with this segmentation process. One of the remarkable features of this API is its ability to predict punctuation—such as commas, periods, question marks, and other sentence-ending

symbols—based purely on the way the speaker talks. This might sound simple, but it's actually quite complex. The system analyzes speech patterns, tone, pauses, and changes in pitch to infer where natural breaks in the sentence should be.

For example, when someone says, "I went to the store and bought apples, oranges, and bananas," the system knows to place commas between "apples" and "oranges," as well as before "and," just like a human would in writing. Likewise, when someone finishes a sentence, the system predicts a period or question mark depending on the tone of voice and context. This punctuation helps transform a continuous string of words into sentences and phrases that make sense on their own.

Why is this important? Without proper segmentation, the text would just appear as one endless line of words, which would be difficult to read, understand, or translate correctly. Imagine trying to translate a paragraph where all punctuation is missing—the meaning could become confused, and the translated text might lose its natural flow. Proper punctuation and segmentation give the system clear boundaries, so it knows exactly where one thought ends and the next begins.

Once the text is segmented by these punctuation markers, each segment is sent individually to the **Google Cloud Translate API**. This approach allows the translation process to happen smoothly and efficiently, translating smaller chunks of text one at a time rather than waiting for an entire paragraph or long speech to finish. This is critical in a real-time translation system like ours, where every millisecond counts.

By working on smaller segments, the translation engine can focus on maintaining the meaning and nuances of each phrase without being overwhelmed. This segmented approach helps preserve the speaker's original intent and tone, making the translated output feel natural and easy to follow. It also means the system can start showing translations on the user's device almost immediately, without long delays.

On the user's screen, the translated text appears alongside the original, segmented text, allowing users to follow along easily. This visual feedback is especially helpful for people who may have hearing difficulties or who prefer to read along as they listen. It also adds a layer of clarity, confirming that the system understood the speaker correctly and is providing an accurate translation.

The segmented text is also used to guide the **speech synthesis** process. When the translated text is converted back into spoken words through text-to-speech (TTS) technology, maintaining punctuation is essential for creating natural-sounding speech. Punctuation marks such as commas and periods tell the TTS engine where to pause, change intonation, or emphasize certain words, just like a human speaker would. Without this, the spoken translation would sound robotic, rushed, or difficult to understand.

For example, a well-placed comma can introduce a short pause, giving listeners a moment to process what was said before moving on. A period signals the end of a thought, often with a slight drop in pitch and a longer pause, helping listeners recognize the sentence boundary. Question marks prompt a rising intonation at the end, indicating that the speaker is asking something. All these subtle cues are vital for making the translation feel natural and engaging.

Our system takes great care to preserve these punctuation cues exactly as predicted by Google Speech-to-Text, ensuring the translated speech has the right rhythm and flow. This attention to detail makes a big difference in how comfortable and intuitive conversations feel, especially when participants speak different languages.

Another advantage of segmenting the text is that it helps handle potential errors or ambiguities more gracefully. Since the system translates one segment at a time, if there is an issue with one part—such as a misheard word or unclear phrasing—it won't disrupt the entire conversation. The system can continue working on the following segments independently, providing a smoother overall experience.

Segmentation also supports advanced features we plan to implement in the future, such as highlighting specific phrases on screen or offering alternative translations for tricky sentences. By breaking down speech into manageable pieces, our system can offer users more control and clarity over their conversations.

To sum up, segmentation is a foundational step in turning raw speech into clear, understandable text and then into accurate, natural-sounding translations. By leveraging Google Speech-to-Text's punctuation prediction capabilities, our system can automatically identify where sentences begin and end, and where important pauses should occur. This enables smoother, more efficient translations, better on-screen text display, and more lifelike speech synthesis.

Without proper segmentation, a real-time translation platform would struggle to keep up with the flow of conversation, leaving users confused or frustrated. Thanks to this thoughtful approach, **Greet** offers a truly seamless experience where language barriers dissolve, and conversations feel as natural and connected as if everyone were speaking the same language.

Machine Translation is the heart of any multilingual communication system like Greet. After the system has carefully listened to a speaker, turned their voice into text, and neatly broken that text into manageable segments, the next crucial step is to convert those segments into another language that the listener understands. This is where machine translation comes in — the magical process that bridges language barriers in real time.

For Greet, we rely on the powerful **Google Cloud Translate API** to perform this critical task. This API uses what's known as **Neural Machine Translation (NMT)**, a cutting-edge technology that's transforming how computers translate languages. Unlike older translation systems that translated word-by-word or phrase-by-phrase, NMT looks at entire sentences and even paragraphs to understand context and deliver much more natural, accurate translations.

To put it simply, the neural machine translation model functions somewhat like the human brain. It "learns" languages by studying vast amounts of text data — books, articles, websites, and conversations in many languages — to recognize patterns and nuances. This allows it to generate translations that sound less robotic and more fluent, capturing the subtleties of meaning, tone, and style.

Let's imagine a simple example: If someone says in English, "It's raining cats and dogs outside," a traditional word-for-word translator might produce a confusing phrase in another language that literally mentions animals falling from the sky. But Google's NMT understands that this is an idiom meaning "It's raining heavily" and translates it accordingly. This deeper understanding of language is exactly what makes real-time communication through Greet feel natural and effortless.

When the segmented text from the speech recognition step reaches the Google Cloud Translate API, the system quickly analyzes it to grasp the intended meaning and tone. It considers grammatical structure, word order, cultural context, and more to craft an accurate translation in the target language. Because this happens in real time, users don't have to wait long—usually just a fraction of a second—before they see or hear the translation.

One of the impressive things about Google's NMT technology is that it is continually improving. It learns from millions of users worldwide who use translation services daily. This ongoing feedback loop helps the system adapt to new slang, evolving language use, and rare phrases that might not have been common before. So, when you use Greet today, you're benefiting from a translation engine that's smart, adaptable, and constantly learning to serve you better.

Moreover, Greet supports a wide range of languages, not just English and Spanish. Whether it's Hindi, Mandarin, French, Arabic, or Japanese, Google's translation models cover over a hundred languages and dialects. This wide coverage means Greet can connect people from almost anywhere in the world, regardless of the languages they speak. For global businesses, educators, healthcare providers, or families spread across continents, this opens up endless possibilities for connection and collaboration.

After the translation is complete, the text doesn't just sit there—it gets passed immediately to the next stage of the system: **Text-to-Speech (TTS) synthesis**. But before we get to that, it's important to understand that the quality of machine translation directly impacts the whole user

experience. If the translation is off, confusing, or sounds unnatural, it can break the flow of conversation and make communication frustrating.

That's why the NMT models are so valuable: they reduce errors and awkward phrasing by understanding the bigger picture of the conversation, not just individual words. They can handle complex sentences, idiomatic expressions, and subtle nuances that older systems often struggled with.

Sometimes, of course, no machine translation is perfect—languages are incredibly complex, and context can be ambiguous. To help with this, Greet shows the translated text on screen alongside the original speech, so users can compare and understand if something doesn't sound quite right. This transparency builds trust and helps users adapt quickly during conversations.

Another strength of Google's translation system is its ability to customize translations for specific needs. For example, businesses can train the model to use industry-specific vocabulary, healthcare professionals can ensure medical terms are translated accurately, and educators can tailor translations for classroom settings. This flexibility means Greet isn't just a one-size-fits-all tool; it can evolve with its users.

It's also worth noting how fast all this happens. Real-time machine translation might sound like magic, but behind the scenes, it involves immense computing power, complex algorithms, and vast language databases all working together in milliseconds. Google's cloud infrastructure ensures that no matter where users are located, translations happen quickly and reliably.

To sum up, machine translation is more than just swapping words from one language to another—it's about conveying meaning, emotion, and intent as faithfully as possible. Google's Neural Machine Translation technology makes this possible by understanding the context, learning continuously, and producing translations that feel natural and smooth.

In Greet, this means you can join conversations across languages without feeling like you're talking to a robot. Instead, it feels like you're having a real, meaningful exchange, breaking down barriers and bringing people closer together—no matter what language they speak.

Text-to-Speech Synthesis plays a crucial role in the overall user experience of a real-time multilingual communication platform like Greet. After the spoken words have been recognized, translated, and converted into text, the next step is to bring those words back to life through speech—so that users can listen to translations as naturally and clearly as if they were speaking face to face. This is exactly what the text-to-speech (TTS) module is designed to do.

For Greet, we use **Google Text-to-Speech** technology, which is one of the most advanced TTS systems available today. This technology transforms the translated text into smooth, natural-sounding speech that users can easily understand. Instead of robotic or monotone voices that can feel distracting or tiring, Google's TTS produces voices that sound warm, expressive, and lifelike, enhancing the quality of conversations across languages.

One of the key reasons why Google's TTS sounds so natural is because it employs a method called **unit selection synthesis**. This technique works by piecing together small, pre-recorded fragments of real human speech, called "units," to form complete sentences. These units are carefully selected to fit the desired pronunciation, intonation, and rhythm, creating speech output that mimics how a human would naturally speak.

What makes this approach effective is its ability to capture the nuances of human speech — the way our tone rises and falls, how we pause between phrases, and how we emphasize certain words for meaning or emotion. This creates a listening experience that's more engaging and easier to follow, which is especially important during longer conversations or meetings.

Another important feature of Greet's TTS system is its sensitivity to **punctuation markers**. Thanks to the earlier speech recognition step, which predicts where commas, periods, question marks, and other punctuation should be placed, the TTS engine knows exactly when to pause, slow down, or change tone. This means that sentences sound natural and clear, with appropriate breaks that help listeners process the information easily.

For example, consider the sentence: "I'll meet you at 5 p.m., but if you're late, please call me." The pause after "5 p.m." signals a slight break, just like in normal speech, helping the listener understand the structure of the sentence. These small details make a big difference in clarity and comprehension, especially in a multilingual setting where listeners might be processing both translation and new vocabulary simultaneously.

Flexibility is another hallmark of Greet's TTS system. The platform allows users to choose between **male and female voices**, providing options that cater to personal preferences or cultural expectations. Some users might find a male voice more authoritative and clear, while others might prefer a female voice for its warmth and friendliness. Offering multiple voice options helps make communication more comfortable and relatable.

Furthermore, Greet supports **multiple voice variations** within languages like English and Spanish. This means that users don't just get a generic "English voice," but can select from a variety of accents and speech styles that better match their context. For instance, a user in Mexico might prefer a Mexican Spanish voice, while a user in Spain might choose a Castilian Spanish voice. This level of customization makes the conversation feel more familiar and personalized.

The choice of voice in the TTS system is also intelligently linked to the **source language** of the speaker. For example, if the original speaker is talking in English, Greet will select an English voice for the TTS output in the translated language, preserving a natural and logical flow. If the original speaker is in Spanish, the system switches accordingly. This dynamic selection ensures that the speech output matches the context and doesn't feel disjointed or out of place.

Behind the scenes, all this happens incredibly fast. The text is sent almost instantly to the TTS engine, which converts it to speech with minimal delay. This speed is essential to maintain a smooth and natural flow of conversation, especially when multiple people are talking back and forth in different languages.

One of the challenges in real-time speech synthesis is managing background noise, accents, and speech clarity. Google's TTS system has been trained on massive datasets from diverse speakers, helping it produce clear and understandable speech even when the input text varies in quality or complexity. This robustness is vital for Greet users, who might be in noisy environments, have different speaking styles, or use slang and idioms.

Another exciting aspect of Google Text-to-Speech is its ability to convey subtle emotions and emphasis through voice modulation. While current implementation focuses on clear and neutral speech to ensure universal understanding, the underlying technology can be enhanced in the future to include emotional tone, excitement, or urgency in speech. Imagine a business call where the TTS system can express enthusiasm or concern, making the interaction more human and engaging despite the digital barrier.

Additionally, Greet's TTS module contributes significantly to **accessibility**. For users with visual impairments or those who find it difficult to read text quickly, hearing the translation is a much more practical and inclusive way to participate in conversations. It also benefits those who are multitasking or on the move, enabling them to stay connected without needing to focus on the screen.

For individuals with **hearing impairments**, the system's real-time transcription complements the TTS output, ensuring that everyone in the conversation can access the content in a way that suits their needs. By combining visual and auditory information, Greet fosters a more inclusive communication environment.

From a technical perspective, Google's TTS API integrates smoothly with the rest of the Greet platform, allowing developers to easily incorporate new voices or languages as they become available. This means that as Google continues to expand and refine its TTS offerings, Greet users will benefit from ongoing improvements without needing to update their apps or settings manually.

In summary, the text-to-speech synthesis component of Greet is much more than just a voice reading out words. It is a sophisticated, thoughtfully designed system that brings translated text to life with clarity, warmth, and naturalness. It respects the rhythm and punctuation of speech, offers customizable voices to suit different preferences, and delivers output quickly to keep conversations flowing smoothly.

By transforming written translations back into spoken language that feels human, Greet removes one more barrier between people, making cross-lingual conversations feel seamless and genuine. Whether it's a business meeting, a virtual classroom, a healthcare consultation, or a friendly chat across continents, the TTS technology helps ensure that every word is heard and understood — just as if you were speaking in the same room.

CHAPTER 2

LITERATURE REVIEW

2.1 Offline Speech Translation

In the early days of speech translation technology, the systems that enabled one language to be spoken, translated, and then spoken back in another language were quite basic compared to what we have today. These early speech translation systems typically operated in an **offline** manner. What that means is that the entire process happened step-by-step, one after the other, rather than simultaneously or in real-time.

Here's how it worked: first, the spoken language was converted into text—a process called **speech recognition**. Once the spoken words were transformed into text, the system would then translate that text from the source language into the target language. Finally, the translated text was turned back into spoken words using **text-to-speech synthesis**.

While this approach was groundbreaking at the time, it had some significant drawbacks, especially when it came to real-time communication. Because the system was **sequence-based**, meaning it processed one step completely before moving to the next, there was an inevitable delay. Imagine speaking into a microphone and waiting several seconds—or sometimes even minutes—to hear the translated version of your speech. This delay made such systems impractical for natural conversations or live settings like meetings, classrooms, or casual chats. Real-time dialogue depends on quick responses to maintain flow and engagement, and the lag caused by this sequential processing simply didn't cut it.

Another challenge these early systems faced was tied to the **methods they used to analyze and process speech**. Most relied on fixed-length analysis windows—think of it as slicing the speech signal into small, uniform chunks and trying to interpret each chunk independently. While this helped simplify the analysis, it was far from perfect. Speech, after all, is fluid and continuous, with natural variations in speed, tone, and rhythm. Breaking it into rigid, fixed pieces often caused unnatural pauses and reduced the system's ability to capture context effectively.

Additionally, many systems assumed the speech signal could be modeled using Gaussian processes. While mathematically convenient, this assumption limited how well the models could handle the complex, varied nature of human speech. These limitations resulted in less accurate translations and robotic-sounding synthesized speech that made conversations feel artificial and stilted. Oord et al. [3] notably highlighted these shortcomings, emphasizing the need for more flexible, data-driven approaches to speech synthesis.

This is where **WaveNet**, a breakthrough developed by DeepMind and described by Oord et al., changed the game entirely. WaveNet introduced a radically different way of generating speech, shifting from traditional techniques to a **deep generative model** that operates directly on raw audio waveforms. This might sound a bit technical, but the implications are huge.

Traditional speech synthesis methods, like vocoders or concatenative synthesis, worked by piecing together pre-recorded fragments of speech or by applying filters and rules to manipulate sound parameters. These methods often produced speech that sounded mechanical or unnatural, and they struggled to convey the subtle nuances of human speech like emotion, intonation, and rhythm.

WaveNet, on the other hand, uses an **autoregressive model** to generate speech sample-by-sample, predicting one tiny piece of audio at a time based on all the previous pieces it generated. This approach mimics how sound waves actually flow in the real world and allows WaveNet to create speech with astonishing detail and naturalness. Instead of relying on pre-recorded snippets or overly simplified models, WaveNet learns from vast amounts of real human speech and is able to recreate the intricate patterns that make speech sound human.

What does this mean in practice? The speech generated by WaveNet sounds much more like a natural human voice, with realistic pauses, emphasis, and emotional undertones. Listeners find it easier to understand and more pleasant to listen to, which is a major step forward for any application relying on speech synthesis—especially live translation systems like Greet.

Furthermore, because WaveNet operates at the **waveform level**, it removes many of the constraints that earlier methods had. It doesn't rely on fixed analysis windows or overly simplistic assumptions about speech structure. Instead, it can adapt to the unique characteristics of each individual speaker's voice, handle different languages and accents more gracefully, and produce speech that flows smoothly in a variety of contexts.

The introduction of WaveNet marked a new era for speech technology. It allowed developers to finally create systems capable of producing **high-quality, natural-sounding speech in real time**, which is crucial for applications like video conferencing, virtual assistants, language translation apps, and more. The ability to generate speech dynamically, sample by sample, without needing to pre-record every possible phrase, also opened up the door to much greater flexibility and scalability.

Still, it's important to remember that WaveNet's remarkable capabilities come with increased computational demands. Generating speech sample by sample requires substantial processing power, especially when aiming for real-time output. However, ongoing improvements in hardware, cloud computing, and algorithm optimization have helped to mitigate these challenges, making WaveNet-style synthesis more accessible than ever.

In summary, the evolution from early, offline speech translation systems to advanced deep learning models like WaveNet has transformed how we communicate across languages. Where once conversations were slow and stilted due to sequential processing and simplified models, modern systems now offer near-instantaneous, natural-sounding speech translation that feels truly human. This evolution is a critical foundation for platforms like Greet, which rely on seamless, real-time communication to break down language barriers and connect people around the world.

Thanks to WaveNet and similar innovations, the dream of effortless, natural multilingual conversations is no longer just a futuristic idea—it's happening right now.

Speech Samples	**Subjective 5-scale MOS in Naturalness**	
Speech samples	NA English	Mandarin Chinese
LSTM-RNN parametric	3.67 ± 0.098	3.79 ± 0.084
HMM-driven concatenative	3.86 ± 0.137	3.47 ± 0.108
WaveNet (L+F)	4.21 ± 0.081	4.08 ± 0.085
Natural (8-bit μ-law)	4.46 ± 0.067	4.25 ± 0.082
Natural (16-bit linear PCM)	4.55 ± 0.075	4.21 ± 0.071

Table 1

Mos scores for different speech synthesis methods.

Subjective 5-scale mean opinion scores of speech samples from LSTM-RNN-based statistical parametric, HMM-driven unit selection concatenative, and proposed WaveNet-based speech synthesizers, 8-bit μ-law encoded natural speech, and 16-bit linear pulse-code modulation (PCM) natural speech. WaveNet improved the previous state of the art significantly, reducing the gap between natural speech and the best previous model by more than 50%.

2.2 Instantaneous Speech Translation Over the Internet

With the world becoming more connected than ever before, the demand for real-time multilingual communication is growing rapidly. Whether it's for business meetings that span

continents, online classrooms where students speak different languages, or friends and family connecting across borders, people want to communicate instantly and effortlessly—no matter what language they speak. This makes immediate speech translation over the internet one of the most exciting and challenging areas of modern technology research.

Historically, speech recognition and translation systems have faced a tough balancing act. They need to be fast enough to keep up with natural conversation while still being accurate enough that the meaning isn't lost or distorted. Unfortunately, traditional models have often struggled with this balance. Many speech recognition systems work well when speed isn't a concern—for example, transcribing a recorded interview or translating subtitles after a video is finished—but when it comes to translating live speech, things get tricky. These older systems can introduce delays, causing unnatural pauses or interruptions during conversations, and they sometimes sacrifice accuracy to process audio faster. This leads to misinterpretations or awkward moments where users have to repeat themselves or clarify what they meant. In real-time communication, these issues can break the flow, frustrate users, and reduce the overall usefulness of the technology.

Recent advances in machine learning and neural network architectures, however, have started to turn this situation around. One of the most promising developments comes from the hybrid transformer-CNN model known as the **Conformer**, proposed by Gulati and colleagues [4]. This model represents a leap forward in speech-to-text processing, combining the best of two powerful techniques to create a system that is both highly accurate and efficient enough for real-time use.

To understand why this is such a big deal, it helps to know a bit about how speech recognition models work. Speech is a complex, continuous signal made up of many sounds that vary over time. In order to convert speech to text, models need to understand both the local acoustic details—like the shape of a vowel or the sharpness of a consonant—and the broader context, such as how words connect to form sentences and meaning.

Traditional models might focus on one or the other. **Convolutional Neural Networks (CNNs)** are very good at detecting local features in audio signals, much like how they identify edges and patterns in images. They excel at capturing short-term acoustic patterns, such as phonemes (the

distinct units of sound). However, CNNs alone aren't designed to understand long-range dependencies—that is, how sounds and words relate to each other across an entire sentence or phrase.

On the other hand, **Transformers**, a newer neural network architecture, use a mechanism called **self-attention** that allows them to weigh the importance of different parts of the input data, regardless of where they occur. This means transformers can model relationships over long distances in the speech signal, understanding context and grammar to help predict the correct transcription. But transformers typically require large amounts of data and computational resources, and they sometimes struggle to capture the fine-grained local details that CNNs handle well.

The Conformer model brilliantly integrates CNNs and transformers, taking advantage of their complementary strengths. By doing so, it captures both the local acoustic features and the global context, making it particularly suited for speech recognition tasks. This hybrid design helps reduce the delay (latency) in processing speech, which is crucial when you want to translate conversations as they happen, without making users wait.

To put the Conformer's performance into perspective, it has achieved **state-of-the-art accuracy on the LibriSpeech benchmark**, a widely recognized dataset used to evaluate speech recognition systems. Specifically, it achieves a word error rate (WER)—a common measure of transcription errors—of about 2.1% without using an additional language model, and improves to around 1.9% with one. These numbers reflect how often the model gets the words wrong in its transcriptions, and such low error rates are a testament to the model's robustness and precision. For comparison, human transcriptionists typically have a WER around 4%, so the Conformer is approaching, or even exceeding, human-level accuracy in some cases.

What does this mean for internet-based speech translation systems? Simply put, the Conformer and models like it enable near real-time, highly accurate speech-to-text conversion with relatively low computational cost. This improvement removes one of the biggest bottlenecks in live translation: the ability to quickly and correctly transcribe spoken language. Once the speech is accurately converted to text, it can be fed into machine translation systems to be converted into another language, and then transformed back into speech for the listener.

This progress has made **uninterrupted cross-lingual communication not only more achievable but also more practical and accessible to everyday users**. Imagine joining a video conference where you don't speak the same language as the other participants but can still understand every word as if it were your own language. Or picture an online classroom where students from different countries participate fully, asking questions and engaging in discussions without language barriers. These scenarios are no longer just aspirations—they are becoming realities thanks to advances like the Conformer model.

Moreover, the improvements don't just benefit casual or professional conversations. Real-time speech translation can have far-reaching impacts in healthcare, government, tourism, customer service, and many other fields where effective communication is vital. For example, a doctor consulting with a patient who speaks a different language can get more accurate information, leading to better diagnoses and treatments. Customer support centers can serve a global client base without needing to hire multilingual staff for every language. Emergency response teams can coordinate more effectively when language is no longer a barrier. The potential applications are vast and transformative.

While the Conformer model and similar innovations have driven significant improvements, the journey is ongoing. Researchers and engineers continue to explore ways to further reduce latency, increase accuracy, and optimize computational resources. They're also working on expanding support for more languages, dialects, and accents to ensure inclusivity and fairness in speech technology.

Another exciting area of development involves integrating these speech recognition models with other cutting-edge technologies, such as **neural machine translation (NMT)** systems that have improved dramatically in recent years. By pairing high-quality speech-to-text models with advanced translation models, it's possible to create seamless end-to-end systems that not only transcribe and translate speech accurately but also capture nuance, idiomatic expressions, and cultural context.

These advancements are making communication across languages less complicated and more natural than ever before. They're helping to break down barriers that once seemed

insurmountable and opening up new opportunities for global collaboration, learning, and understanding.

In conclusion, the rising demand for real-time multilingual communication has pushed speech translation technology to new heights. Thanks to innovative hybrid architectures like the Conformer, combining CNNs and transformers, speech-to-text systems are becoming faster, more accurate, and more practical for live applications. This progress is a crucial enabler for internet-based speech translation services, making them more accessible and effective than ever.

As technology continues to evolve, the vision of effortless, real-time conversations between people of any language background moves closer to everyday reality. The world is becoming more connected, and the language barriers that once limited communication are gradually dissolving—one word, one sentence, one conversation at a time.

2.3 Text-Based Speech Translation

In the rapidly evolving field of language translation, many earlier systems focused almost exclusively on text-based solutions. What this means is that the speech input was first converted into text—a process called transcription—and then that text was translated into another language. The result was typically another block of text, which users had to read to understand the message. While this approach was a significant step forward in breaking down language barriers, it still fell short of enabling truly natural conversations. After all, when we communicate face-to-face or even over a video call, speech is much more than just words on a page—it's tone, rhythm, intonation, and emotion. Reading translated text cannot fully replicate the nuances and immediacy of spoken dialogue.

For this reason, the focus of many recent developments has shifted toward **speech-to-speech translation**—systems that take spoken language as input and produce spoken language as output in another language, effectively allowing people to talk to each other in real time without a shared tongue. However, this goal is far from trivial. Generating natural, expressive speech in another language requires not just accurate transcription and translation but also high-quality **speech synthesis** technology.

Thankfully, the last few years have witnessed tremendous breakthroughs in speech synthesis, the technology behind converting text into natural-sounding spoken words. Two landmark advancements in this area—**Tacotron** and **Deep Voice**—have dramatically raised the bar for how computer-generated speech sounds, making speech-to-speech translation more realistic and accessible than ever before.

The Rise of Neural Text-to-Speech: Tacotron and Deep Voice

Traditional speech synthesis systems were complex and often inflexible. They relied heavily on painstakingly engineered features and concatenative methods that pieced together small audio fragments recorded from real human voices. While effective to a degree, these systems were limited in their ability to produce expressive or emotionally rich speech, and creating new voices required huge amounts of data and manual effort.

Tacotron, developed by Google researchers, revolutionized this process by introducing an end-to-end neural network model that synthesizes speech directly from text. As Wang et al. [5] discuss, Tacotron's approach eliminates many of the intricate intermediate steps required by traditional systems. Instead of focusing on handcrafted features or separately tuning different components, Tacotron uses a **sequence-to-sequence model with attention mechanisms** to learn the direct mapping from a sequence of characters (letters or phonemes) to a sequence of audio features, which are then converted into speech waveforms.

This design allows Tacotron to produce speech that sounds incredibly natural and fluid, with accurate pronunciation, rhythm, and intonation—all without needing manual intervention. The attention mechanism within the model helps it understand which parts of the input text correspond to which sounds, enabling it to generate speech that aligns perfectly with the text's flow. Arik et al. [6] demonstrated how Tacotron's end-to-end nature allows it to generate human-quality speech with minimal effort and great flexibility. This has been a game-changer for text-to-speech systems, setting a new standard for naturalness and intelligibility.

Deep Voice 2: Taking Voice Synthesis to New Heights

Building on these innovations, **Deep Voice 2**, developed by Baidu researchers, introduced additional capabilities that made speech synthesis even more versatile and powerful. One of the standout features of Deep Voice 2 is its ability to handle multiple speakers using a single model. Traditional systems generally required training separate models for each voice, which was costly and time-consuming.

Deep Voice 2 uses **low-dimensional speaker embeddings**—compact representations that encode the unique characteristics of each speaker's voice. This allows the model to synthesize hundreds of different voices simply by providing it with a small set of recordings for each new speaker. As a result, it can mimic different voices with high fidelity, preserving the speaker's identity and style while maintaining excellent audio quality.

This multi-speaker feature has vast implications. For speech translation systems, it means that the voice synthesizer can preserve speaker individuality or even allow the listener to choose a preferred voice style, whether it's male or female, young or old, formal or casual. It also opens

the door for applications in entertainment, personalized virtual assistants, and accessibility tools that require varied voice outputs without having to train a new model from scratch every time.

The Impact on Speech-to-Speech Translation

By incorporating models like Tacotron and Deep Voice into speech-to-speech translation systems, the entire user experience improves dramatically. Instead of receiving dry, robotic voice outputs or just text to read, users hear translations spoken in clear, natural voices that reflect real human speech patterns and emotions. This makes conversations more engaging and easier to follow, especially in sensitive or nuanced interactions such as business negotiations, medical consultations, or personal conversations.

For example, imagine a multinational team holding a virtual meeting. With earlier text-only translations, participants might have had to wait for the text to appear and read it while others were talking—breaking the flow and making the conversation feel unnatural. With speech-to-speech systems powered by Tacotron and Deep Voice, participants can listen to translations in real time, with voices that sound warm, expressive, and understandable. This fosters smoother communication, reduces misunderstandings, and helps build trust between speakers of different languages.

How Neural Networks Changed the Game

What makes Tacotron and Deep Voice so revolutionary is their use of **neural networks** to replace traditional pipelines. Older text-to-speech systems were like factories with many specialized assembly lines—one part handled pronunciation, another controlled pitch, yet another stitched audio fragments together. Each step required extensive engineering, tuning, and expert knowledge. Any change in one component could ripple through the system and cause errors.

In contrast, Tacotron and Deep Voice are trained end-to-end on large datasets of text paired with audio recordings. During training, the neural networks automatically learn the complex relationships between text and speech, including pronunciation, prosody (the patterns of stress and intonation), and speaker identity. This makes the systems highly flexible and adaptable: they can generate speech in new voices or styles by learning from just a few examples, and they can generalize well to text they haven't seen before.

Because these models are data-driven, they continuously improve as more speech data becomes available. Developers can also fine-tune them for specific languages, dialects, or accents, ensuring that speech synthesis becomes increasingly personalized and context-aware.

Encouraging Further Development in the Field

The breakthroughs embodied by Tacotron and Deep Voice have energized the entire speech synthesis community. Researchers worldwide are building on these foundations to tackle challenges like emotional speech synthesis, expressive prosody control, and cross-lingual voice cloning. By drastically simplifying the TTS pipeline, these models free up researchers and engineers to experiment with new features rather than reinventing the basics.

This progress is essential for advancing **speech-to-speech translation** systems that are more than just functional—they must also be enjoyable and effective communication tools that people trust and want to use daily. Thanks to these neural network models, the technology is no longer limited by technical constraints but is driven by creativity, user needs, and real-world applications.

Real-World Applications and Impact

The improvements in speech synthesis don't just make technology sound better—they enable practical, life-changing applications. Here are a few examples:

- **Accessibility**: People with speech impairments or hearing difficulties can use speech synthesis to communicate more effectively, especially when combined with translation to bridge language gaps.
- **Education**: Students learning new languages can hear accurate pronunciations and intonations, improving their language skills.
- **Healthcare**: Doctors can use speech-to-speech translation to communicate with patients who speak different languages, improving diagnosis and treatment.
- **Entertainment**: Voice actors and creators can produce diverse voices without extensive recording sessions, enabling richer audio experiences in games, audiobooks, and virtual assistants.
- **Customer Service**: Automated agents can interact naturally with customers worldwide, providing personalized and multilingual support.

2.4 Voice-To-Voice Translation

In recent years, technological innovations in speech processing have taken enormous strides toward making real-time, voice-to-voice communication across languages a reality. What once seemed like science fiction—the ability to speak in your native language and be instantly understood by someone halfway across the world speaking an entirely different language—is now possible thanks to the integration of automatic speech recognition (ASR), machine translation (MT), and text-to-speech (TTS) systems.

Traditionally, these components were developed and operated separately. For instance, speech recognition would transcribe audio into text, the text would be translated by machine translation models, and the translated output would be synthesized into speech using TTS. These steps were often disjointed, leading to delays and inefficiencies. However, recent developments have led to a more unified system design, enabling direct voice-to-voice translation with minimal latency and significant improvements in quality and accuracy.

One of the most exciting advancements in this field is the development of multi-speaker TTS systems. Unlike older TTS models that relied on a single, often robotic-sounding voice, multi-speaker systems can replicate a wide range of human voices, complete with unique characteristics and speaking styles. This is especially important for communication systems, where preserving the speaker's identity and expressiveness helps maintain clarity and authenticity in conversations.

A key component of these multi-speaker systems is speaker identification and optimization. By using just a few minutes of recorded speech, modern systems can "learn" the characteristics of a new speaker and generate speech that closely mimics their voice. Deng et al. [7] report impressive results, showing that these systems can achieve high Mean Opinion Scores (MOS) for both speaker similarity (MOS: 4.64) and naturalness (MOS: 4.16). These scores reflect listeners' perceptions of how natural and speaker-like the synthetic voices are—indicating near-human quality.

The cornerstone of these achievements lies in the use of speaker embeddings. These are mathematical representations of a speaker's unique vocal traits, which are embedded in both spectral (acoustic features) and latent (underlying abstract features) spaces. Embedding in these two dimensions allows the model to normalize across different types of speech data, including out-of-domain text, which typically presents challenges due to unfamiliar sentence structures or vocabulary.

Speaker embeddings are crucial for maintaining consistency in voice synthesis, especially when the system is required to generate speech from text that doesn’t reflect typical conversational language. For example, technical documents or poetic expressions often have rhythms and structures that differ from everyday speech. A robust multi-speaker TTS system with well-constructed embeddings can still produce natural-sounding audio even in these scenarios.

The implications of this are profound. Imagine an international conference where each participant speaks in their native language but hears everyone else in their own language, spoken with the original speaker’s voice and intonation. This not only improves understanding but also preserves the emotional and contextual cues that are often lost in translation. It fosters more meaningful interactions and bridges gaps not just in language, but in culture and empathy.

Furthermore, these TTS systems can be fine-tuned to enhance specific qualities of speech, such as expressiveness, emotion, and emphasis, depending on the use case. For educational tools, this might mean slower, clearer speech with exaggerated enunciation. For entertainment applications, it could mean highly expressive voices with dynamic inflections. And in business communication, a professional, calm tone can be prioritized.

The real-time nature of modern voice-to-voice translation systems is equally critical. Delays in conversation can be disruptive and frustrating. Advanced architectures are now designed with minimal processing lag, ensuring a seamless user experience. This is made possible by optimized pipelines that tightly integrate ASR, MT, and TTS components, often leveraging cloud-based infrastructure to process and transmit data swiftly and securely.

A practical implementation of such a system includes the use of neural networks to manage each stage of processing. The ASR module uses deep learning models to recognize speech with high accuracy, even in noisy environments or across diverse accents. This recognized text is then passed to a neural machine translation (NMT) model, which translates it into the target language while preserving meaning and context. Finally, the TTS module—enhanced with multi-speaker capabilities and speaker embeddings—generates spoken output that matches the original speaker's identity.

To illustrate the effectiveness of this system, consider a real-world scenario: a tourist in Spain asks for directions in English. The system instantly recognizes the English input, translates it into Spanish, and synthesizes a natural Spanish response that the local person can understand. The local replies in Spanish, which is then translated and spoken back to the tourist in English, using a voice that sounds like the local speaker. This interaction happens in real time, with no noticeable lag, making communication smooth and intuitive.

Such technology is not limited to casual conversation. In healthcare, it can facilitate communication between doctors and patients who don't share a common language, potentially saving lives. In education, it can allow students to access lectures in their preferred language while preserving the instructor's style and tone. In global business, it can help teams collaborate more effectively, reducing misunderstandings and fostering stronger relationships.

The future of voice-to-voice translation looks even more promising as research continues to push boundaries. Innovations such as emotional TTS, which captures and conveys not just the words but the emotions behind them, and cross-lingual speaker adaptation, which allows a speaker's voice to be used in languages they've never spoken, are already under exploration.

In conclusion, the integration of ASR, MT, and advanced TTS systems—particularly those utilizing multi-speaker models and speaker embeddings—has paved the way for a new era of real-time, high-quality speech translation. These technologies do more than convert words; they

preserve identity, emotion, and intent, creating a communication experience that is as close to natural conversation as current technology allows.

The system described here, based on the innovations and findings by Deng et al. [7], represents a significant step forward in making seamless, multilingual communication a part of everyday life.

The system proposed is shown in the figure below.

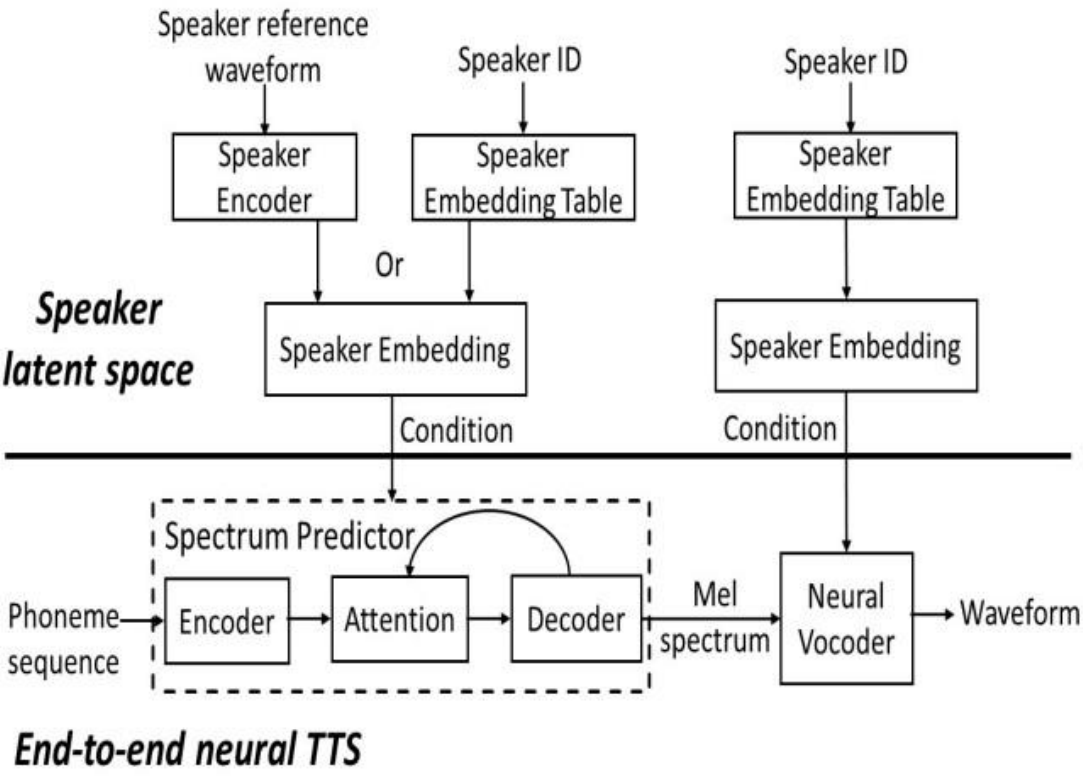


Figure 1. Neural TTS Architecture

CHAPTER 3

PROPOSED METHODOLOGY

In today's globally connected world, video conferencing platforms are more than just tools for virtual meetings—they're essential for education, business, healthcare, and international collaboration. Yet despite their rapid evolution, a major shortcoming remains: the lack of truly seamless, real-time multilingual communication. Many existing systems still fall short in delivering accurate and fluid translations during live conversations, creating barriers for users across different linguistic and cultural backgrounds.

To address these issues, we propose a next-generation, AI-powered video conferencing system that places real-time speech translation and optimization at its core. This isn't just a simple upgrade—it's a complete overhaul of how language is handled in digital conversations. At the heart of this system is a powerful multilingual engine that combines two major translation technologies: Neural Machine Translation (NMT) and Statistical Machine Translation (SMT). By using a hybrid model, the system leverages the best of both worlds: NMT's deep understanding of context and nuance, and SMT's strength in translating specialized or rare language pairs where data may be sparse.

Unlike traditional models that perform translation in disconnected steps, our approach is end-to-end and highly optimized. The system uses cutting-edge, end-to-end automatic speech recognition (ASR) models that are trained on diverse datasets. These models can transcribe speech to text with high accuracy, even in noisy environments—thanks to robust training with accents, dialects, and environmental noise included. Once the speech is transcribed, the text flows directly into a real-time translation pipeline. With streamlined architecture and minimal processing overhead, translations happen in just a few seconds—keeping conversations smooth and natural.

To tackle one of the biggest practical challenges—background noise—we've integrated AI-powered noise suppression right into the system. This is far more advanced than basic noise filters. It uses sophisticated deep learning algorithms, particularly CNNs and RNNs, to isolate human speech from environmental noise. Techniques like time-frequency masking and spectral analysis allow the system to separate voice from noise in real time, maintaining clarity even in noisy cafés or busy offices. What's more, the system continuously adapts using a feedback loop that adjusts filtering strategies based on the acoustic environment. This means the system gets better and more accurate as it listens—learning from every conversation.

But language isn’t just about words—it's about tone, intention, and interaction. That’s why we’ve also built tools to enhance engagement and responsiveness. Users can provide real-time feedback on the quality of audio and translation, which feeds directly into the system’s improvement cycle. Built-in sentiment analysis tools evaluate tone and emotional cues in conversations, giving users richer insights and enabling more emotionally intelligent interactions. These tools don’t just translate words—they help translate meaning.

Security is also a top priority. Communication often involves sensitive or confidential information, and users need to trust the platforms they use. To that end, our system incorporates multi-factor authentication (MFA) using facial recognition and voiceprint verification. This combination helps ensure that only authorized users can access secure sessions, without compromising ease of use.

Together, these components work to overcome the most pressing issues in video conferencing today. The result is a system that isn’t just more accurate—it’s more human, more adaptive, and more secure. It fosters an inclusive environment where language is no longer a barrier but a bridge.

The Evolution of Machine Translation: From SMT to Hybrid Systems

Looking at the history of translation technology, most early systems relied heavily on Statistical Machine Translation (SMT). SMT works by breaking sentences into small segments or phrases, then calculating probabilities for their translations based on large datasets of bilingual text. It’s essentially a statistical puzzle: what’s the most likely translation based on past patterns?

While SMT has strengths—such as being relatively easy to implement and adaptable to specific domains—it also has serious limitations. Since it translates fragments piece by piece, the results often lack fluency. Words may be correct, but the sentence might feel awkward or robotic. Complex sentence structures, idioms, and subtle contextual cues are difficult for SMT to handle, often resulting in stilted or inaccurate translations. Additionally, SMT struggles with speed. Since it relies on large statistical computations, it's slow to process language in real time—making it poorly suited for the demands of video conferencing.

This is where the newer hybrid approach shines. Neural Machine Translation models—particularly those built using Transformer architectures—understand the broader context of a sentence and can translate more fluently. NMT doesn’t just look at phrase-by-phrase mappings; it considers the meaning of the entire sentence, leading to more natural output. By integrating NMT with SMT, the system gains both flexibility and fluency, providing smoother translations, especially for rare or complex language pairs.

Figure 2 in the system architecture shows how SMT operates using a hierarchy of translation steps. The process begins with basic word alignments and gradually introduces phrase

reordering, duplication, and refinement. A separate language model, trained on large corpora of monolingual text, helps determine which translations sound most natural in the target language. While this approach was groundbreaking in its time, it falls short in today's fast-paced, multilingual communication environments. As discussed by Patare et al. [10], the limitations of SMT make it unsuitable as a standalone solution for real-time translation.

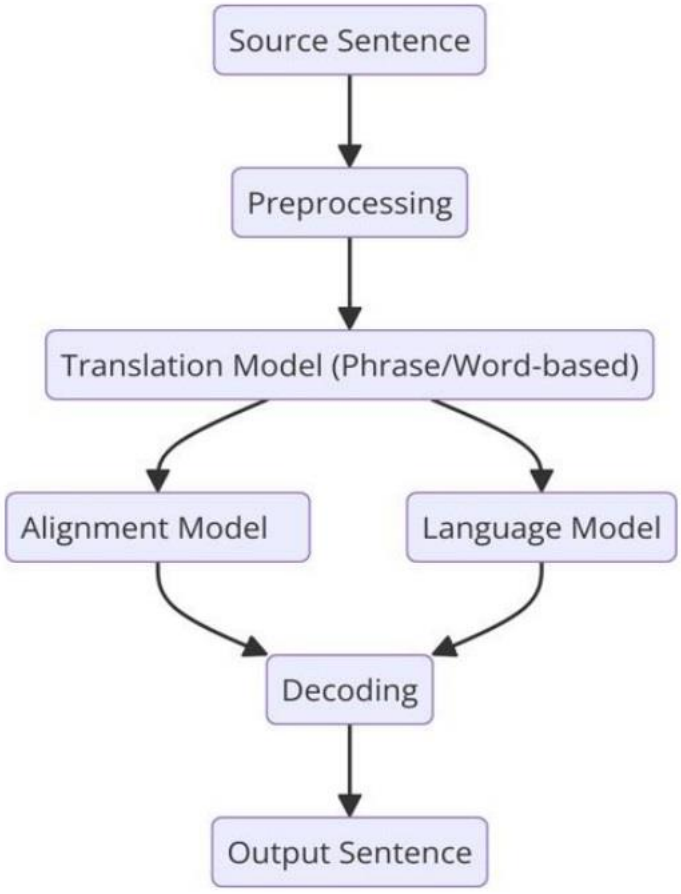


Figure 2. Statistical MT Pipeline

3.1 Algorithm to be Used (SMT vs NMT)

The limitations inherent in Statistical Machine Translation (SMT), particularly its challenges in processing complex sentence structures and context-dependent translations, necessitate a shift towards more advanced methodologies. SMT models often exhibit translation inaccuracies, suffer from high computational overhead, and struggle with real-time performance constraints due to their reliance on probabilistic phrase-based approaches. To address these inefficiencies, Neural Machine Translation (NMT) has emerged as the dominant framework for contemporary translation systems.

Neural Machine Translation (NMT)

Neural Machine Translation (NMT) represents a deep learning-based paradigm that enhances translation quality by processing entire sentences holistically, rather than fragmenting them into independent segments. This approach enables the model to capture contextual relationships, thereby producing more fluent and semantically coherent translations.

Unlike SMT, which depends on predefined linguistic rules and statistical phrase alignments, NMT utilizes neural networks to learn complex linguistic patterns directly from data. A key component of NMT is the attention mechanism, which allows the model to focus on relevant portions of the input sentence while generating translations. This mechanism significantly improves translation accuracy, particularly for longer and structurally intricate sentences.

NMT systems require large-scale multilingual datasets for training but have demonstrated substantial advancements in translation quality. Due to their ability to generalize across varied linguistic structures, they are well-suited for real-time applications and diverse language pairs. Consequently, the integration of NMT-based translation models within video conferencing systems facilitates efficient, high-fidelity, and context-aware multilingual communication, as shown by Patare et al. [10].

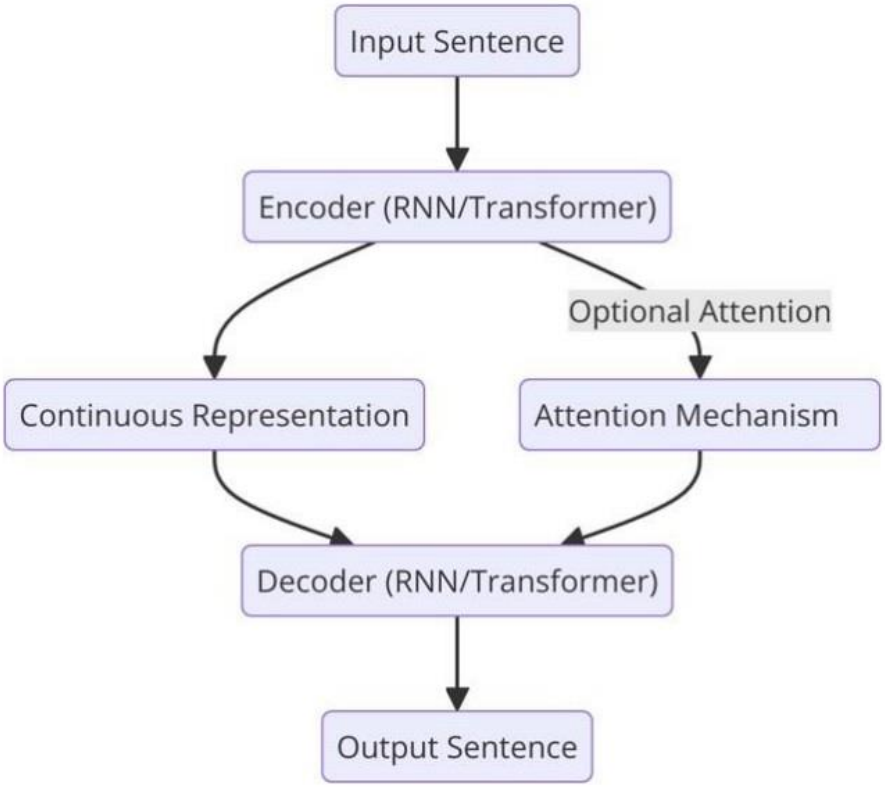


Figure 3. Neural MT with Attention

3.2 Application Areas

The proposed AI-powered multilingual video conferencing system is more than just a technological upgrade—it represents a fundamental shift in how people connect and communicate across linguistic boundaries. Its ability to enable seamless, real-time translation has transformative potential across a wide range of domains. From global corporations and freelancers to educators and content creators, this system has the power to break down language barriers, enhance collaboration, and open new opportunities for innovation and inclusion. Below, we explore its applications across five key sectors:

Startups and Multinational Corporations (MNCs): Powering a Truly Global Workforce

Seamless Global Collaboration

In today's fast-paced business world, companies—whether nimble startups or sprawling multinationals—often work with partners, clients, and teams that span multiple countries and cultures. Language differences can quickly become a roadblock to productivity and clear communication. With real-time translation built into their video conferencing tools, these companies can connect effortlessly, allowing everyone to speak in their native language and still be fully understood. The need for human interpreters or delayed post-meeting translations becomes a thing of the past.

Boosting Productivity

Meetings become more focused and productive when everyone can express themselves clearly and confidently. This system allows for richer, more inclusive discussions where all voices are heard—regardless of the speaker's language. By removing the friction caused by language gaps, businesses can speed up decision-making, solve problems faster, and unlock deeper team collaboration.

Global Recruitment Made Easy

The ability to conduct interviews, onboarding sessions, and training programs in multiple languages is a game-changer for talent acquisition. Businesses can now truly tap into a global talent pool, reaching exceptional candidates who may not be fluent in the company's primary language. The system makes the hiring and onboarding process more accessible, welcoming, and efficient for everyone involved.

Freelancers: Thriving in a Multilingual Gig Economy

Managing a Diverse Client Base

Freelancers increasingly work across borders, collaborating with clients from different countries and language backgrounds. A multilingual conferencing tool allows them to handle everything from initial project pitches to final reviews without worrying about language differences. This smooth, real-time communication builds trust and fosters lasting professional relationships.

Customizing Communication Styles

Beyond translation, the system adapts to cultural and regional nuances in language. Freelancers can better align their communication style with the expectations of each client, whether that

means using formal language, industry-specific terminology, or regional idioms. This creates a more professional and personalized client experience.

Content Creators & Live Streamers: Going Global with Ease

Wider Audience Engagement

For creators and streamers, the ability to connect with audiences in different languages is incredibly valuable. With built-in multilingual support, they can host live Q&A sessions, tutorials, or interviews that are accessible to viewers around the world. This dramatically increases reach and opens up new audience segments without needing to create separate content for each language.

Real-Time Interaction Across Languages

One of the most exciting features for creators is the ability to respond to comments and engage with viewers in real time—no matter what language they speak. Viewers can interact naturally, and creators can respond instantly with translated replies, fostering more inclusive and lively communities.

Collaborating Across Borders

Multilingual conferencing also enables creators from different countries to work together on collaborative content. Whether it's a podcast, a live event, or a co-created video, creators can now combine their talents without worrying about language barriers. This promotes cultural exchange and creative diversity.

Educational Institutions: Enabling Inclusive and Global Learning

Inclusive Classrooms for All Students

Language should never be a barrier to education. Schools and universities can use this system to support students from different linguistic backgrounds, ensuring that everyone has equal access to lessons, discussions, and group work. Multilingual support allows students to participate fully in their own language, boosting confidence and comprehension.

Creating Global Classrooms

Educators can now host truly international classrooms, where students from around the world join and collaborate as if they were sitting side by side. This global perspective enriches learning, encourages cultural exchange, and prepares students for working in international environments.

Enhancing Remote and Hybrid Learning

As remote and hybrid learning models become more common, the need for effective communication tools becomes more critical. The system supports live lectures, student presentations, and collaborative projects—all with real-time translation. This ensures that learning remains dynamic and engaging, regardless of where students are or what language they speak.

Corporate Meeting Platforms: Elevating Communication and Efficiency

Reliable, Real-Time Communication

Businesses rely on efficient and secure communication platforms for internal collaboration, client meetings, and executive discussions. This system combines HD audio and video quality, real-time multilingual translation, and intelligent noise cancellation to deliver seamless conversations across borders.

Smarter Documentation

After meetings, the platform can generate automated summaries and transcripts in multiple languages. These summaries capture key discussion points and decisions, making them accessible to everyone involved. This is especially valuable for teams operating across time zones or departments, helping keep everyone on the same page.

Feature	Existing Systems	Proposed System	Statistical Improvement
Latency (Average Delay)	10-15 seconds	Reduced to 4-5 seconds (sometimes even 3 seconds)	Up to 67% reduction in latency
Browser Compatibility	Often reliant on specific browser APIs	Works across all browsers without reliance on specific APIs	Increased accessibility across platforms
Transcription Accuracy	Often inaccurate, leading to higher error rates in translations	Utilizes NMT and SMT for state-of- the-art transcriptions	Significant improvement in transcription accuracy

Table 2. System Performance Comparison

CHAPTER 4

RESULTS AND DISCUSSION

4.1 User Interface Overview

Figure 4 illustrates the main interface of the application, featuring several key options: New Meeting, Join Meeting, Schedule Meeting, and View Recordings. The interface adopts a minimalist design philosophy, ensuring ease of navigation and a clutter-free user experience. Each option is clearly labelled with intuitive icons, enabling users to quickly initiate their desired action. The streamlined layout aligns with user-centric design principles, ensuring that even non-technical users can operate the platform with minimal guidance.

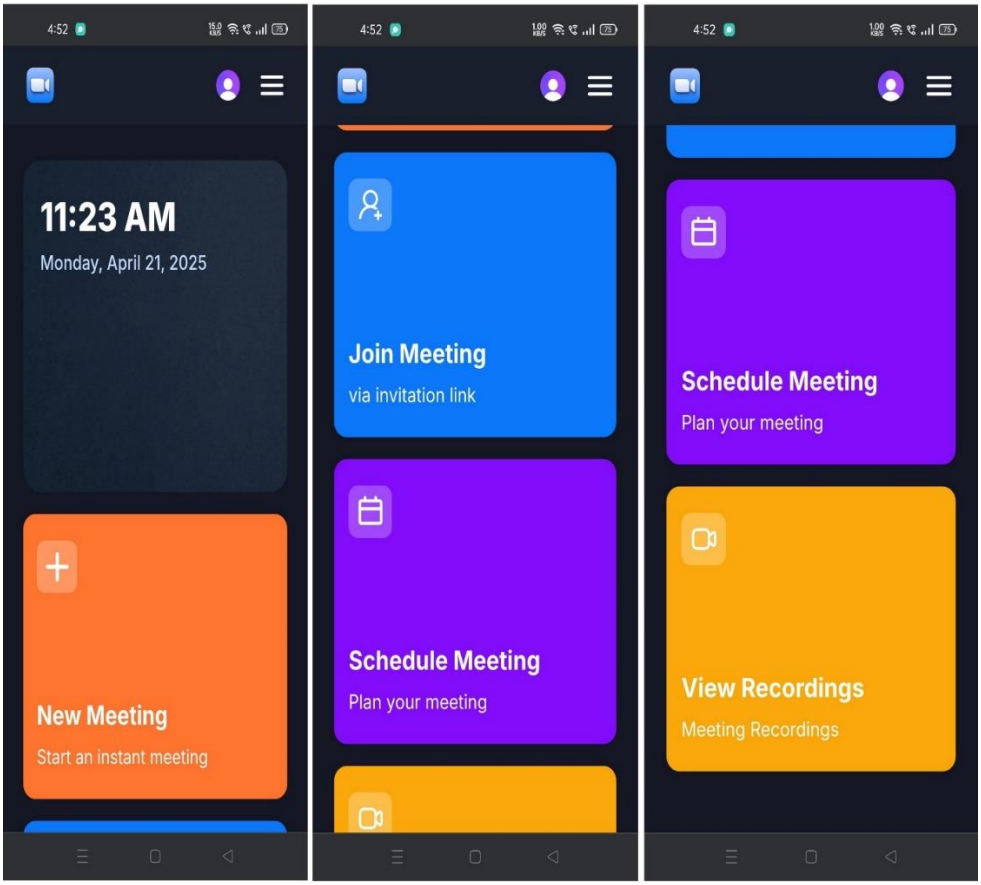


Figure 4. Working 1

4.2 Meeting Management Features

The developed platform offers a range of essential features designed to enhance usability and adaptability. Users can initiate a new meeting effortlessly, wherein a unique meeting ID and link are automatically generated. This supports spontaneous collaboration and simplifies the sharing of invitations. The Join Meeting functionality provides a straightforward way for participants to access ongoing sessions by entering a valid meeting ID, ensuring seamless integration into the conference environment.

The platform also supports the scheduling of meetings in advance. Users can select a date, time, and additional parameters, enabling organized planning and timely coordination. Complementing these capabilities, the View Recordings feature allows users to retrieve, review, and share previously conducted meetings. Recordings are systematically stored, which is particularly beneficial in business, educational, and documentation-focused scenarios.

4.3 Active Meeting Interface

Following the meeting management features, Figure 5 presents the active meeting interface where users participate in real-time video conferences. The interface displays live video feeds of all participants and includes controls for managing audio and video settings. Users can mute/unmute microphones, toggle camera feeds, and adjust basic settings dynamically, offering a flexible and responsive meeting experience.

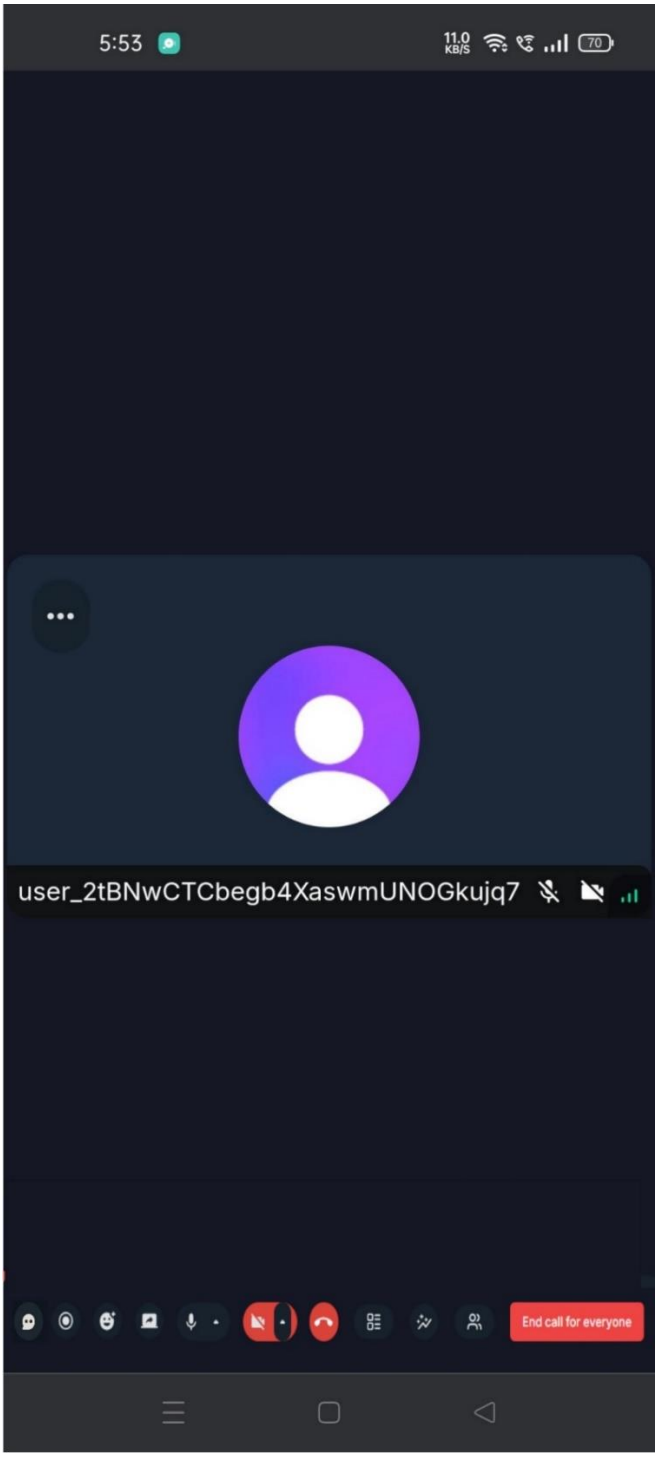


Figure 5. Working 2

4.4 Meeting Recording Capabilities

The application provides an integrated recording feature. Users can record ongoing meetings with a single tap, and recordings are automatically saved post-session for easy access. This function is crucial for users who require archival of meetings for future reference, training, or compliance purposes.

4.5 Real-Time Call Statistics

Figure 6 introduces the Call Statistics panel, which displays real-time network and performance metrics during an active meeting. Parameters such as latency, jitter, packet loss, bandwidth consumption, and frame rate are dynamically monitored. These statistics are presented through graphical and numerical displays, offering users valuable insights into their call quality and enabling quick troubleshooting of network-related issues.

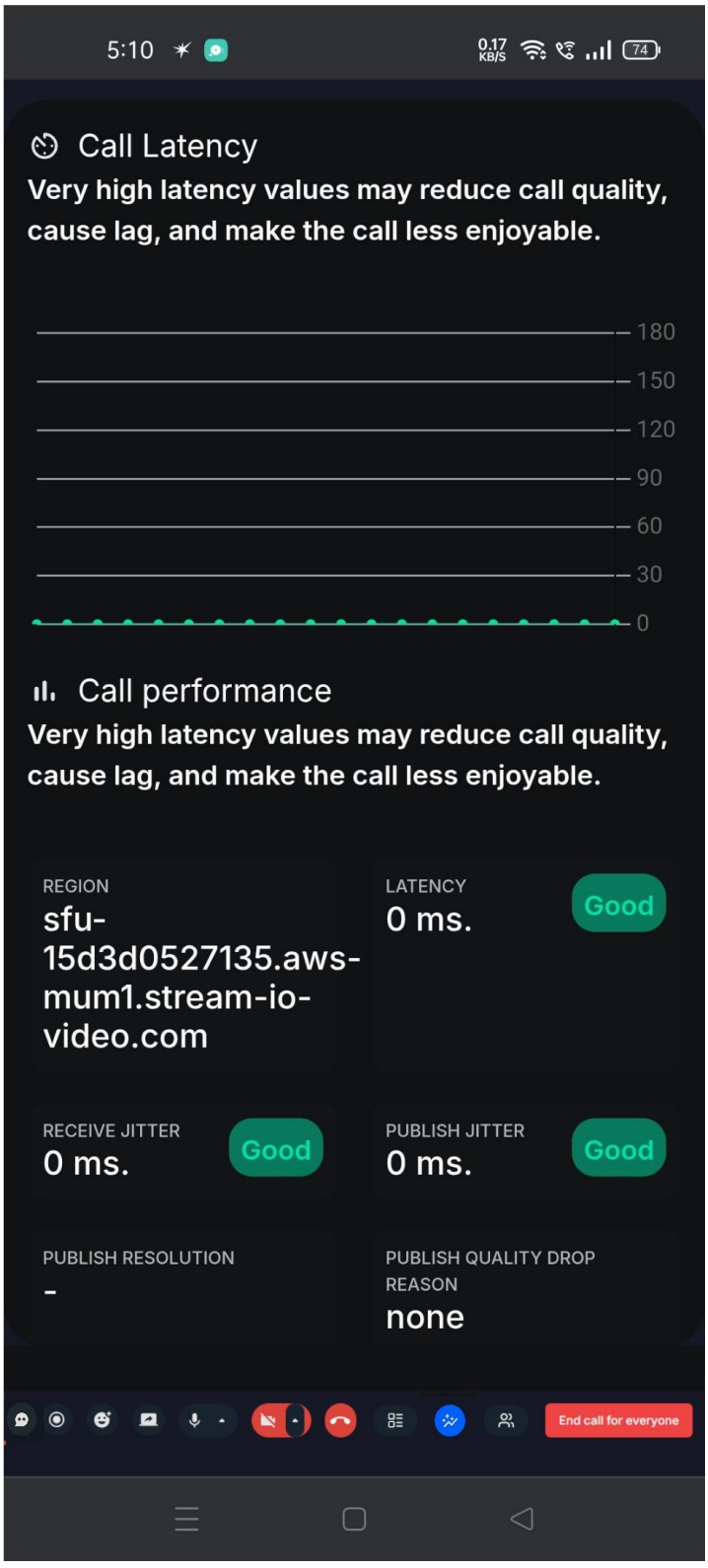


Figure 6. Working 3

4.6 Multilingual User Interface

As illustrated in Figure 7, the platform includes a Language Selection feature to support multilingual accessibility. During the initial setup or from the settings menu, users can choose their preferred language for the interface and in-app notifications. This ensures inclusivity for users from diverse linguistic backgrounds. The application dynamically adjusts all textual elements, such as buttons, labels, error messages, and instructions, according to the selected language. By supporting a wide range of regional and international languages, this feature enhances user engagement and aligns with global usability standards.

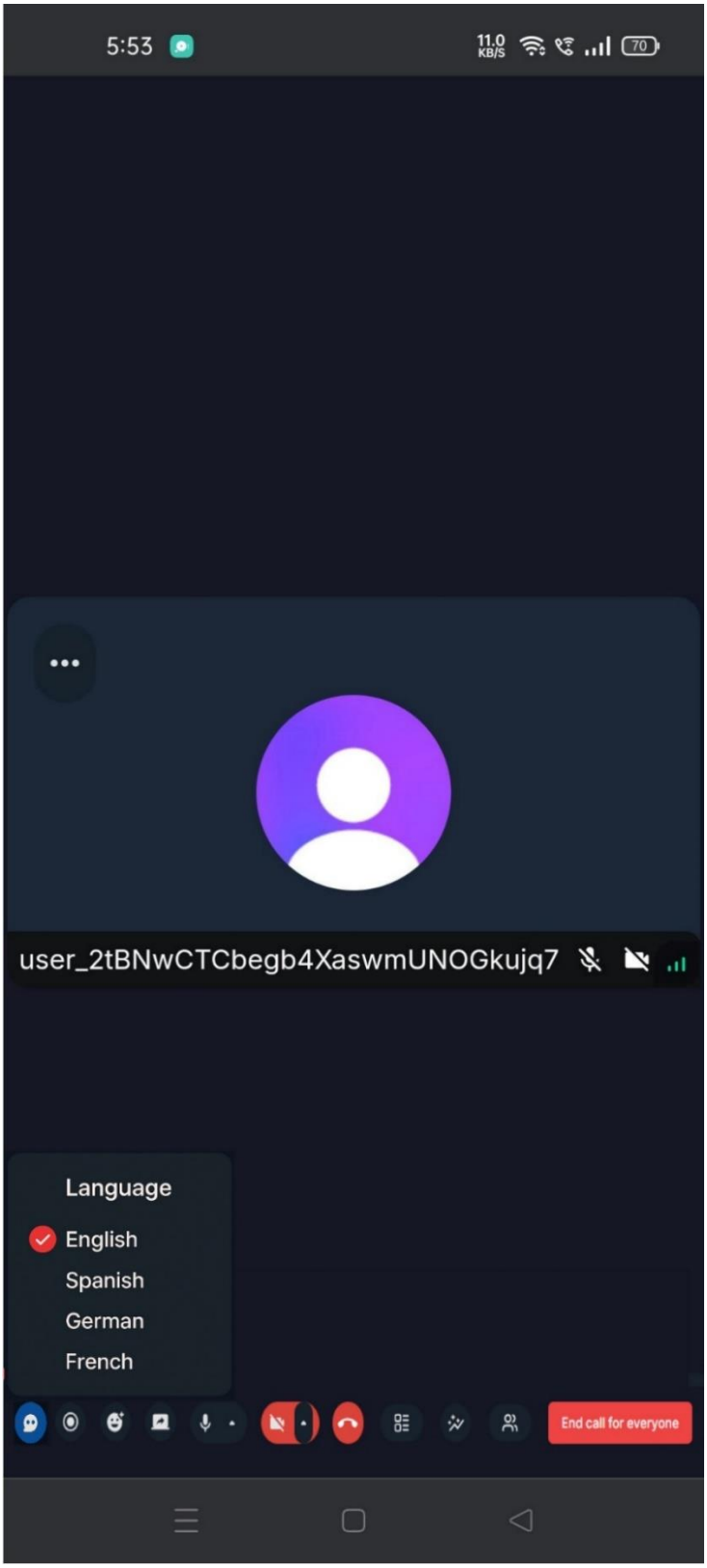
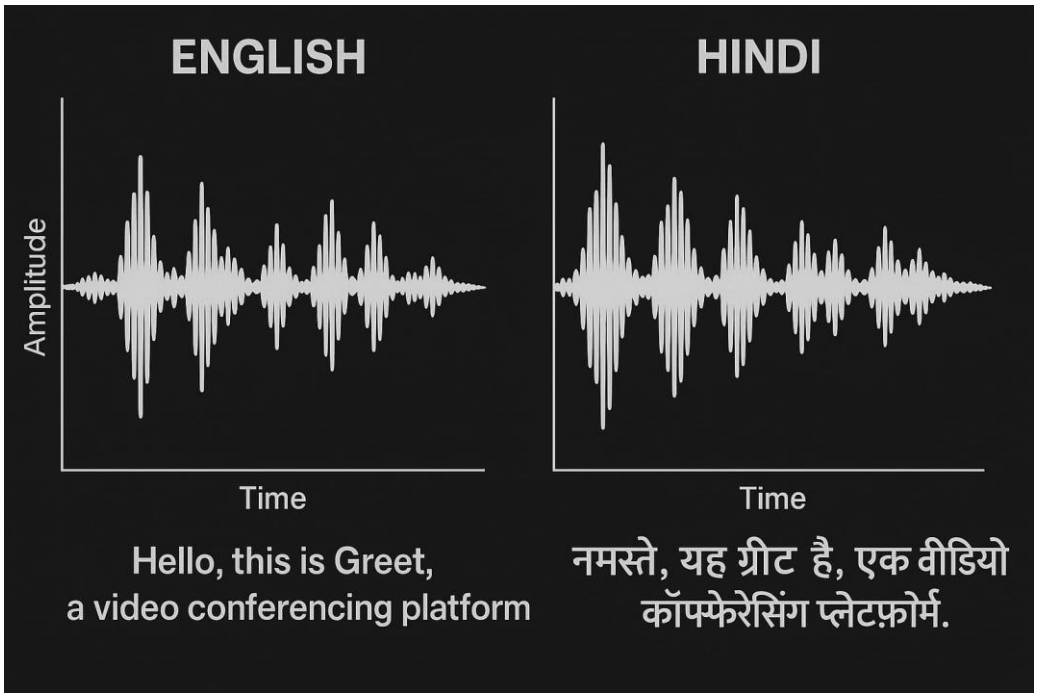

5:53
11.0 KB/S
70
user_2tBNwCTCbegb4XaswmUNOGkujq7
Language
English
Spanish
German
French
End call for everyone

Figure 7. Working 4

4.7 Audio Translation and Waveform Analysis

An important highlight of the platform is its support for multilingual communication, enhancing inclusivity and user reach. Figure 8 demonstrate this, a sample English greeting, "Hello, this is Greet, a video conferencing platform," was translated into Hindi. The corresponding audio signals of both languages were recorded and analysed. A frequency graph comparison of the two samples revealed notable differences in waveform characteristics. The English audio exhibited sharp, stress-timed patterns with distinct peaks, while the Hindi audio showed a smoother, syllable-timed flow with more evenly distributed energy.

This analysis highlights how different languages impact audio signal properties and emphasizes the platform's ability to manage these variations effectively. The system maintains high fidelity and low latency during real-time translation, ensuring that the communication remains clear and natural across different linguistic backgrounds. These results validate the robustness and versatility of the platform in bridging communication gaps and improving global accessibility.



CHAPTER 5

CONCLUSION AND FUTURE SCOPE

5.1 Conclusion

In today's rapidly evolving, interconnected world, communication remains both our greatest asset and, at times, our most stubborn obstacle. As people collaborate across continents, cultures, and time zones, one challenge continues to stand out: language. While technology has made it easier than ever to connect, the nuances and barriers created by different languages often prevent full understanding, creating friction where there should be flow. This is the precise challenge that inspired the creation of Greet—a real-time speech translation system that aims to make multilingual communication feel as seamless and natural as a face-to-face conversation.

Greet is not just a tool. It's a statement of what the future of communication can be. By integrating advanced technologies such as Google Cloud APIs, Agora SDK, and Firebase into a video conferencing platform, Greet allows people to speak freely in their native language, while others hear or read the message in their own language, almost instantly. This real-time, speech-to-speech (or speech-to-text) translation bridges linguistic divides, empowering teams, communities, and individuals to interact more fluidly and authentically.

The motivation for Greet came from observing how language—even more than distance or time zones—can hinder collaboration. Fluency in a second or third language is a remarkable skill, but even seasoned multilingual professionals may find themselves losing nuance or hesitating when switching languages. That hesitation creates a gap—not just in understanding, but in confidence, authenticity, and connection. Greet was built to close that gap, to remove that hesitation, and to make it possible for everyone to express themselves fully without worrying about whether they're being understood.

At its core, Greet captures spoken words using high-fidelity audio tools provided by the Agora SDK. The speech is then processed through Google's Speech-to-Text API, which transcribes the audio into text with high accuracy. This transcription is passed to the Google Translation API, which quickly converts it into the target language chosen by each listener. From there, the translated message can either appear as real-time subtitles or be converted back into speech through the Text-to-Speech API. Firebase supports the infrastructure by managing user preferences, synchronizing sessions, and ensuring that all data flows smoothly across devices and participants.

All of this happens in real time, usually within a fraction of a second. For the user, the experience feels magical—like having a personal interpreter who works invisibly and instantly, allowing them to focus entirely on the conversation itself rather than how to express their thoughts.

And yet, the beauty of Greet is not found only in its technological sophistication, but in the stories of the people who use it. Picture a multinational team, spread across Brazil, Germany, South Korea, and Egypt, coming together for a weekly project update. In the past, meetings would be conducted in English, requiring non-native speakers to translate in their heads while speaking carefully and often sparingly. Now, each person speaks in their native language, and everyone else hears the translation in theirs. The change is profound. Meetings become more inclusive, participation improves, and ideas flow more freely.

Or consider a university hosting an online lecture for students around the world. The professor, who speaks French, no longer needs to prepare lecture notes in English or worry about whether her complex ideas are being lost in translation. Her words are translated in real time, allowing students from India, Mexico, and Japan to understand and engage fully with the content. The focus returns to learning, not just language.

In healthcare, where clarity can literally be a matter of life and death, Greet becomes a lifeline. Doctors conducting telemedicine sessions with patients in other countries can communicate effectively and compassionately, without second-guessing whether they've been understood. Families with members speaking different languages can stay connected in a more meaningful way, sharing stories and emotions with less frustration and more joy. In moments big and small, Greet is quietly transforming how we communicate.

However, like all ambitious technologies, Greet faces its share of challenges. One of the most persistent is background noise, which can interfere with accurate speech recognition. In a noisy café, a crowded classroom, or an open-plan office, unwanted sounds can distort the input audio, leading to mistranslations or dropped words. While existing noise reduction technologies are helpful, they're not foolproof, especially in unpredictable or dynamic environments.

Another complication lies in the incredible diversity of human language. Even within the same language, dialects, regional accents, and cultural references can vary significantly. A Spanish speaker from Spain might use expressions or pronunciations unfamiliar to someone from Argentina or Mexico. Accents in English can differ drastically between regions like the U.S., the U.K., Australia, or India. Although the speech recognition and translation APIs Greet relies on are constantly improving, there's still work to be done in handling these variations accurately and gracefully.

Speed is another critical concern. Real-time translation demands fast processing at every stage—from capturing and transcribing speech to translating and rendering it for the listener. When everything works smoothly, the experience feels effortless. But under heavy load, especially in

large meetings with many participants speaking different languages simultaneously, the system can experience slight delays. Even a second or two of lag can break the natural rhythm of conversation, making exchanges feel disjointed or awkward.

Scalability is closely related. As more users adopt Greet, the infrastructure must grow with them. That means not just more servers and bandwidth, but smarter data handling, better load balancing, and possibly more localized processing using edge computing to reduce dependency on centralized cloud APIs. These are complex technical challenges that the team behind Greet is actively working to solve.

Despite these obstacles, the trajectory for Greet is undeniably promising. Future development is focused on refining performance, reducing latency, and enhancing accuracy through a combination of machine learning, user feedback, and custom-trained models. For example, domain-specific speech models can help improve accuracy in specialized fields like law, medicine, or engineering, where vocabulary can differ significantly from everyday conversation. Machine learning-based noise filters that adapt to environmental conditions in real time are also in the pipeline.

Another promising area is user-driven feedback. By allowing users to rate translation accuracy or flag errors, Greet can learn from real-world interactions and continuously improve. These feedback loops are vital for creating a system that not only performs well in ideal conditions but thrives in the messy, unpredictable reality of human conversation.

Beyond the technical roadmap, the heart of Greet's mission remains deeply human. It is about giving people the freedom to be themselves in every conversation. It is about respecting linguistic and cultural diversity while offering the tools to transcend those differences. It is about restoring confidence to those who might otherwise stay silent, and bringing new voices into conversations that matter.

Think about a young engineer, recently hired into a global team, who is brilliant in her work but hesitant in meetings conducted in English. With Greet, she speaks in Hindi and is finally heard—clearly, confidently, and in her own words. Or a refugee navigating a new healthcare system, who can describe their symptoms in Arabic and receive care without the added fear of being misunderstood. These are the quiet revolutions Greet makes possible.

As we look to the future, it's clear that the need for tools like Greet will only grow. Our world is becoming more connected, not less. Cross-cultural collaboration isn't a niche use case—it's the new normal. Whether in business, education, healthcare, or personal life, the ability to communicate across languages in real time will increasingly define the quality of our interactions.

Greet represents a step toward that future. A future where multilingual teams brainstorm without friction, where international students engage with professors directly, where families separated by distance still feel close, and where every voice—no matter the language—is heard and understood.

Of course, there's still work to be done. Language is beautifully complex, and no system will ever replace the richness of human nuance entirely. But that's not the goal. Greet doesn't seek to replace human connection—it seeks to support it, to strengthen it, to make it easier for people to be present and real with each other, regardless of the languages they speak.

In a world full of noise—literal and figurative—Greet is a tool that helps people listen, understand, and respond with clarity and empathy. It doesn't just translate words; it translates intention, emotion, and meaning. And in doing so, it reminds us that no matter where we come from or what language we speak, we all want the same things: to be heard, to be understood, and to belong.

5.2 Future Scope

Looking ahead, it's clear that Greet's journey is only just beginning. While the platform already represents a major step forward in breaking down language barriers, the next phase of development will focus on making the system even more adaptive, intelligent, and resilient in the unpredictable environments where people actually live and work. Because real life isn't a controlled lab. It's messy, it's noisy, and it's full of subtle, unspoken variables—especially when people communicate through screens from different parts of the world.

One of the most pressing challenges lies in background noise. It's an unavoidable part of modern life. Whether someone is calling in from a bustling coffee shop, a crowded home during peak hours, or a public space with constant interruptions, these ambient sounds can confuse even the most sophisticated voice recognition systems. The reality is, people can't always control their environment—but technology should be able to adapt to it. That's why one of the key areas of future work is the integration of more advanced and context-aware noise reduction techniques. These aren't just basic filters. They are smart, AI-driven algorithms that can learn to distinguish between human speech and irrelevant sounds in real time—music, traffic, chatter, the occasional barking dog—and keep the focus where it belongs: on the conversation.

But beyond noise, there's another complexity that adds richness to language—and sometimes, difficulty for machines: dialects and regional accents. Every language is a living, breathing thing. It stretches and bends depending on geography, community, and culture. The way English is spoken in Mumbai is different from how it's spoken in Manchester or Mississippi. Arabic spoken in Cairo is different from the dialects used in Morocco or the Levant. These variations carry identity, history, and emotional nuance. And yet, current translation models often struggle to interpret them with full accuracy. That's why expanding Greet's ability to handle a wider range of dialects and accents is not just a technical upgrade—it's a recognition of linguistic diversity and a commitment to inclusivity.

To do this effectively, the Greet team is exploring the use of adaptive machine learning models that can fine-tune themselves based on user input and context. Over time, with enough data and interaction, Greet could begin to understand not just formal, textbook language, but the way real people speak every day—in all their diversity. This evolution will help ensure that the system doesn't just translate correctly but does so with nuance, respecting the rhythm, style, and meaning intended by the speaker.

And while all of these enhancements sound powerful, they're only meaningful if they can be delivered consistently—everywhere, to everyone. That's where another major advancement comes into play: edge computing. Right now, much of Greet's processing happens through cloud-based APIs. While efficient, this model has its limits, especially for users in areas with slower internet connections or unstable network infrastructure. For someone attending a virtual conference from a rural village or joining a team call from a moving vehicle, delays—even slight ones—can disrupt the flow of communication.

By incorporating edge computing, Greet can bring much of the heavy lifting closer to the user's device or nearby network nodes. This means faster processing, lower latency, and greater resilience in low-bandwidth conditions. It's a major step toward making the platform not just globally available, but truly equitable. After all, real-time translation shouldn't be a luxury reserved for users with top-tier internet speeds. It should be a baseline feature—accessible to everyone, no matter where they are or what device they're using.

These future improvements are more than just enhancements to a product. They're part of a larger vision—one that's rooted in the belief that communication is a fundamental human right. And that right shouldn't be dictated by geography, language proficiency, or bandwidth. In a world that increasingly depends on virtual interaction, systems like Greet aren't just helpful—they're essential. They redefine what it means to be "present" in a digital space.

The potential applications are as broad as they are profound. In the corporate world, as businesses become more globalized, cross-border collaboration is no longer optional—it's expected. Teams made up of professionals from different continents and cultures are now the norm. Greet ensures that every team member, regardless of their native language, can contribute fully and confidently. No more defaulting to a single language that may exclude or disadvantage others. No more lost ideas due to hesitations or misunderstandings. Just clearer, faster, more authentic collaboration.

In education, Greet can reshape the learning experience entirely. Picture a virtual classroom where a professor in Madrid teaches a course attended by students from Korea, Nigeria, Brazil, and France. Instead of struggling through a second or third language, each student hears the lecture—and asks questions—in their mother tongue. The result isn't just better comprehension. It's better engagement, better discussion, and a stronger sense of community. The classroom becomes not just international, but truly inclusive.

In healthcare, the implications are equally significant. Language barriers in telemedicine are not just frustrating—they can be dangerous. A misheard symptom, a mistranslated instruction, or an unclear explanation can have serious consequences. With Greet, patients can speak freely and describe their concerns in their native language. Doctors, in turn, can respond with clarity and compassion, knowing their words will be understood. This not only improves outcomes but also builds trust—a vital ingredient in any patient-provider relationship.

Even in our personal lives, tools like Greet can make a difference. Families scattered across continents, speaking different languages, can share stories, jokes, and everyday moments without relying on one bilingual relative to play interpreter. Multilingual friendships and relationships become easier to navigate. Cultural exchange becomes more spontaneous and meaningful. Language no longer stands in the way of connection—it becomes a bridge.

And perhaps most importantly, Greet levels the playing field. It gives a voice to those who might otherwise feel left out of conversations. A young entrepreneur in Jakarta pitching to investors in San Francisco. A grassroots activist in Nairobi collaborating with a team in Geneva. A researcher in Vietnam presenting findings to a global scientific community. With real-time translation, they don't have to code-switch or self-censor. They can speak boldly, in their own voice, and still be understood.

This, ultimately, is the heart of Greet. It's not just about words—it's about empowerment. It's about honoring every person's right to express themselves fully, without compromise. It's about creating spaces where no one has to hesitate before speaking up, and where ideas rise on their merit, not their language of origin.

Of course, there's still a long way to go. Language is deeply complex. Beyond words, it includes tone, humor, emotion, cultural context. These things are hard to translate with precision, and no system—no matter how advanced—can perfectly replicate the richness of human expression. But that doesn't mean we shouldn't try. It means we must build tools that are not just technically brilliant, but emotionally intelligent. Tools that don't just process language, but respect it.

And that's what makes Greet exciting. Not just its ability to perform impressive technical feats, but its potential to reshape how we see and speak to one another. In a world that often feels divided—by politics, by borders, by misunderstanding—Greet offers a way to listen better, to connect more deeply, and to remember that at the end of the day, we all want the same thing: to be heard, to be understood, and to belong.

So as Greet evolves—with better noise handling, greater dialect sensitivity, faster performance, and broader accessibility—it's not just becoming a smarter platform. It's becoming a more human one. A system built not just for convenience, but for connection. Not just for efficiency, but for empathy.

In that sense, Greet isn't just a tool for virtual meetings. It's a small piece of a bigger vision: a world where language unites us rather than divides us, and where everyone, everywhere, has a voice that can be heard.

REFERENCES

[1] A. A. Aksenova, E. Y. Andreeva, A. A. Lugovaya, and N. Y. Tikhonova,"Challenges of Cross-Cultural Communications in the Era of the COVID-19 Pandemic," SHS Web of Conferences, vol. 103, 2021.

[2] S. Zhang, Q. Fang, S. Guo, Z. Ma, M. Zhang, and Y. Feng, "Stream-Speech: Simultaneous Speech-to-Speech Translation with Multi-task Learning," arXiv preprint arXiv:2406.03049, 2024.

[3] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "WAVENET: A Generative Model for Raw Audio," arXiv preprint arXiv:1609.03499, 2016.

[4] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution- augmented Transformer for Speech Recognition," arXiv preprint arXiv:2005.08100, 2020.

[5] Y. Wang, R. J. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, "Tacotron: Towards End-to-End Speech Synthesis," arXiv preprint arXiv:1703.10135, 2017.

[6] S. Arik, G. Diamos, A. Gibiansky, J. Miller, K. Peng, W. Ping, J. Raiman, and Y. Zhou, "Deep Voice 2: Multi-Speaker Neural Text-to- Speech," arXiv preprint arXiv:1705.08947, 2017.

[7] Y. Deng, L. He, and F. Soong, "Modeling Multi-Speaker Latent Space to Improve Neural TTS: Quick Enrolling New Speaker and Enhancing Premium Voice," arXiv preprint arXiv:1812.05253, 2019.

[8] L. Barrault, Y.-A. Chung, M. Coria Meglioli, and others, ”Seam- less: Multilingual Expressive and Streaming Speech Translation,” arXiv preprint arXiv:2312.05187, 2023.

[9] Anonymous, ”High-Fidelity Simultaneous Speech-To-Speech Transla-tion,” arXiv preprint arXiv:2502.03382, 2025.

[10] P. Patare, P. Said, A. Shaikh, and S. Shah, ”Bridging Language Barriers: The Role of AI in Real-Time Multilingual Translation for Video Con-ferencing,” International Journal of Innovative Research in Technology, vol. 11, no. 5, pp. 2001, Oct. 2024, ISSN: 2349-6002.

APPENDIX

APPENDIX A : Sample Code Snippets

1. **Agora SDK setup in Flutter**

dependencies: agora_rtc_engine: ^6.0.0 permission_handler: ^10.2.0 firebase_core: ^2.4.1 cloud_firestore: ^4.5.0 flutter_tts: ^3.5.2 http: ^0.13.5

2. **Request microphone permission & join Agora channel**

import 'package:agora_rtc_engine/agora_rtc_engine.dart'; import 'package:permission_handler/permission_handler.dart';

Future initAgora() async { await [Permission.microphone].request();

await AgoraRtcEngine.create('YOUR_AGORA_APP_ID'); await AgoraRtcEngine.enableAudio(); await AgoraRtcEngine.joinChannel('YOUR_TOKEN', 'test-channel', null, 0);

AgoraRtcEngine.onUserJoined = (int uid, int elapsed) { print('User joined: $uid'); };

AgoraRtcEngine.onUserOffline = (int uid, UserOfflineReason reason) { print('User offline: $uid'); }; }

3. **Capture microphone audio**

import 'package:flutter_sound/flutter_sound.dart';

final recorder = FlutterSoundRecorder();

Future startRecording() async { await recorder.openRecorder(); await recorder.startRecorder(toFile: 'audio.wav'); }

Future stopRecording() async { final path = await recorder.stopRecorder(); print('Recording saved at: $path'); }

4. **Send audio to Google Speech-to-Text API**

import 'dart:convert'; import 'dart:io'; import 'package:http/http.dart' as http;

Future transcribeAudio(String filePath) async { final bytes = await File(filePath).readAsBytes(); final base64Audio = base64Encode(bytes);

final url = 'https://speech.googleapis.com/v1/speech:recognize?key=YOUR_GOOGLE_API_KEY';

final requestPayload = { 'config': { 'encoding': 'LINEAR16', 'sampleRateHertz': 16000, 'languageCode': 'en-US', 'enableAutomaticPunctuation': true, }, 'audio': {'content': base64Audio}, };

final response = await http.post(Uri.parse(url), headers: {'Content-Type': 'application/json'}, body: jsonEncode(requestPayload));

final responseJson = jsonDecode(response.body);

if (responseJson['results'] != null && responseJson['results'].length > 0) { return responseJson['results'][0]['alternatives'][0]['transcript']; } else { return ''; } }

5. **Translate text using Google Translate AP**

Future translateText(String text, String targetLanguage) async { final url = 'https://translation.googleapis.com/language/translate/v2?key=YOUR_GOOGLE_API_KEY';

final requestPayload = { 'q': text, 'target': targetLanguage, };

final response = await http.post(Uri.parse(url), headers: {'Content-Type': 'application/json'}, body: jsonEncode(requestPayload));

final responseJson = jsonDecode(response.body); return responseJson['data']['translations'][0]['translatedText']; }