

# AI-Powered Digital Meeting Application with Integrated S2ST and S2SLT (VaaniMeet)

Maithili Kamble<sup>1</sup>, Garv Balgi<sup>2</sup>, Urja Mali<sup>3</sup>, Prof. Sachin Chavan<sup>4</sup>

<sup>1,2,3</sup>Student at Mahatma Gandhi Missions College of Engineering and Technology, Mumbai, Maharashtra, India.

<sup>4</sup>Professor at Mahatma Gandhi Missions College of Engineering and Technology, Mumbai, Maharashtra, India.

\*\*\*

**Abstract** - There was a great shift in the usage of digital meeting applications like Zoom and Google Meet, where the usage surged over 300% post 2020. However, it did solve long distance communication problem, the problem related to multilingual communications that hinders non-native or cross-cultural speakers and exclusion of the Deaf and Hard-to-Hear (D/HH) community due to no accessible-feature for these problems. This paper proposes VaaniMeet, a platform that integrates Speech-to-Speech Translation (S2ST) and Speech-to-Sign Language Translation (S2SLT), a solution for multilingual virtual meetings. An AI-based model for low-latency translation, VaaniMeet employs a synchronised 3D avatar to convey translated speech and sign language gestures, ensuring a smooth and seamless interaction across multiple languages. The core technical challenge- achieve a real-time synchronisation of audio translation with dynamic 3D avatar animations-is addressed through optimised edge computing and predictive rendering pipelines.

**Key Words:** *Speech-to-Speech Translation (S2ST); Speech-to-Sign Language Translation (S2SLT); Deaf and Hard-of-Hearing (D/HH), Automatic Speech Recognition (ASR), Machine Translation (MT), and Speech Synthesis (TTS), end-to-end speech (E2E).*

## 1. INTRODUCTION

During the pandemic of COVID-19 in 2020, the use of digital meetings became frequent due to lockdown, which opened up a long-distance work relief. Famous applications like Zoom Meeting and Google Meet are the most trusted applications by many organisations. Lectures, office meetings, and even family meetings were the primary and frequent reasons for use.

While this created a chance for long-distance work opportunities but made us realise a major drawback of multilingual meetings, where communication becomes hard as both parties are not comfortable with a common language. But one community that was majorly left out was Deaf and Hard-of-Hearing (D/HH). To overcome both the language barrier and provide equal and fair opportunities for everyone, regardless of disability, to prove their potential.

## 2. LITERATURE REVIEW

This paper introduces us to a second-generation direct Speech-to-Speech Translation (S2ST) model, which skips the need for intermediate text representation, which reduces the “voice leaking” effect found in first-generation models [1]. Its primary contribution in VaaniMeet is its ability to preserve the original speaker’s vocal identity (pitch and tone) in the translated output while remaining robust against noise.

This paper explores the “textless” approach to translate by utilising discrete speech units. By mapping audio directly to these units, the researchers proved that systems can translate smoothly even for languages that do not have a written script [2]. This is vital for VaaniMeet when handling vernacular or informal spoken communication, where standard text translation might fail.

The study in this paper addresses the common problems of “data paucity” in training AI. The researchers developed a method to use stupendous amounts of unlabeled data, which is in text form, to pre-train these translation models, which are responsible for improving the semantic accuracy of the system [3]. As for VaaniMeet, this research provides the logic needed to ensure the high accuracy translation even when parallel audio data are scarce for a specific language pair.

The “Unity” model focuses on escalating the translation process by training the AI with multiple Text-to-Speech (TTS) targets. This multitasking approach allows the model to generate audio much faster than the traditional system [4]. This is a core requirement of VaaniMeet to ensure that the time pause between the live audio, which is a person speaking and the translated audio, which is being heard, is minimal.

This paper is for coincident translation. It proposes a model that can translate speech in real time while the speaker is still engaged in conversation [5]. By using the “multi-tasking learning” framework, it balances the trade-off between waiting for enough content and then predicting the next and translating the half-sentence before the speaker completes the sentence and then starts working on the remaining half-sentence. Which increases the speed of translation and still tries to keep the grammar of the language to be translated.

Which allows the model to keep the conversation as natural and smooth as possible.

This research focuses on “on-device” S2ST translation. It discusses how to condense large AI models so that they can run on mobile devices like laptops or even smartphones [6]. As for VaaniMeet, this encourages the goal of data privacy, as sensitive business meetings can be translated without sending the audio data to the external cloud servers.

The “Hibiki” system focuses on high-fidelity simultaneous translation. It ensures that the generated voice combination is not robotic but maintains a human-like accent by keeping a human-like flow and emotional reverberations [7]. This is censorious for VaaniMeet’s users, making it feel less like an AI interaction and more like a real conversation.

This research paper explores the idea of how a single AI model can manage multiple languages at once, e.g., English to Hindi and English to Marathi simultaneously [8]. This research is the backbone of VaaniMeet’s ability to handle translation of one language to multiple languages in a single meeting e.g., Speaker is English and listeners are Spanish, Hindi etc. A single AI model manages to translate into multiple languages on multiple devices at single time.

This study introduces us to a more efficient processing architecture called “Latent Perceivers” [9]. Traditional AI models consume a lot of memory when processing long sentences; however, this architecture allows VaaniMeet to handle long professional lectures or meetings without slowing down or crashing the system.

This research measures and compares the use of 2D video clips versus 3D avatars for sign language [10]. The research suggests that 3D avatars provide much better clarity, which is necessary for certain signs that require hand-to-chest or hand-to-face movements. This justifies our choice of using 3D avatars in VaaniMeet for S2SLT.

This research provides the technical mapping framework for VaaniMeet’s avatar [11]. It uses “1D-CNN” to analyze the auditory features of the speaker’s voice and an “LSTM” network for predicting the sequence of hand gestures. This ensures that the avatar moments are right and logically connected, and grammatically correct in sign language.

This is one of the most crucial multi-model datasets in the world. It contains over 80 hours of high-quality video data where speech, sign language, and 3D depth data are all positioned [12]. VaaniMeet will utilize this dataset to train the avatar to recognise and recreate thousands of different complex gestures accurately.

This paper proposes a “compact” Large Vision Model design specifically for “edge” or mobile devices [13]. It ensures that

the 3D avatars’ animation is high quality and smooth, even on devices without a high-end graphics card, which is necessary for the accessibility of VaaniMeet.

This research explores how to operate an avatar’s movements using only audio input [14]. By analyzing the voice pitch and energy of the speaker’s voice, the avatar can just adjust the speed and “weight” of its signs, by making the gestural output feel more emotionally aligned with the speaker’s intentions or emotions.

This study focuses specifically on Indian Sign Language (ISL). It provides the Natural Language Processing (NLP) rules which are required to convert English or Hindi Grammar (Subject-Verb-Object) into ISL grammar (Subject-Object-Verb) [15]. This is critical for making VaaniMeet useful in the Indian Regional Context, as we are creating VaaniMeet mainly considering Indian users

This research uses “fuzzy logic” to handle the ambiguity within human language [16]. Since in human language, there are words with multiple meanings, the neuro-fuzzy converter helps the avatar to choose the most accurate sign based on the sentiment of the sentence, which reduces the translation errors.

This paper highlights a major gap in existing technology: “Non-Manual Signals” [17]. It proves that an avatar cannot fully understand unless it includes facial expressions, head tilts, and eye gaze. VaaniMeet incorporates these findings by ensuring the avatar’s face moves along with its hands.

This introduces a specific metric to measure “latency” in continuous translation. The “wait-k” logic determines that the system should wait for ‘k’ numbers of words before starting the output [18]. This is used in VaaniMeet to ensure that the avatar doesn’t start moving before it has enough data to be accurate.

This survey analyses the current meeting tools in the market. It identified that while apps like Zoom and Meet provide text captions, none of them offer an integrated 3D sign language translation [19]. This research proves VaaniMeet needs the market.

This paper looks at the future of two-way communication. It discusses using neural networks for the translation of sign language gestures back into translated audio [20]. This provides us with the future scope for VaaniMeet, where we can make a two-way communication bridge for the Deaf community. Where the user can use sign language to express, and the AI model will analyse the sign signals and translate them into the targeted language.

### 3. PROBLEM STATEMENT

Digital collaboration is frequently hindered by 2 obstacles: the linguistic barrier between global speakers and the accessibility gap for Deaf and hard to hear (D/HH) community. Standard digital meeting platforms often lack integrated, real-time tools with support multilingual dialogue or provide immediate sign language interpretation for deaf and hard-to-hear users, often requiring expensive external services or human interpreters for translation. VaaniMeet addressed these issues by unifying **S2ST** and **S2SLT** into a single seamless prototype. By converting spoken language into translated neural voice in audio form and keyword- triggered sign-language using gifs, the project provides an inclusive, real-time differences or hearing ability.

### 4. PROPOSED SYSTEM (VaaniMeet)

Imagine attending a meeting online using a digital meeting platform with foreigners and being able to speak in a language without any problem and still able to keep point. That's VaaniMeet- a platform built to make global communication efficient and smooth, and give a good experience to all participant either speaker or listeners. It's like any other digital meeting platform, but with 2 important features which translate audio to any other language for listeners and translate audio for the D/HH community into sign languages.

The meeting lobby has 2 extra buttons **S2ST**, which detects speakers language and translate it into the selected language. This makes the speaker comfortable to keep the point even if the speaker is not an expert in the listener's language, and also for the listener, as they do not get confused in understanding the accent or pronunciation, as every accent pronounces words differently and can mislead the meaning of words.

The other button is **S2SLT**, which detects the language spoken by the speaker and translates it into English and translates English language shows sign language gifs on screen in a container in the corner for participants who have issues hearing. This is especially for the D/HH community, as just audio or captions are not enough for smooth understanding, as some participants might not be good readers; hence, captions sometimes might not be helpful.

In this feature, 2 steps are common: speech-to-text and text-to-text translation. It works like this: audio is detected by the primary tool, faster-whisper, mainly known for Speech-to-Text. It includes a built-in Language Identification (LID) model. And the second step is Text-to-Text secondary tool **Googletrans** if there is any external extra noise, then it helps as a text-based backup. Then it translates based on the user's

clicked button. If s2st is clicked after TTT, Text-to-Speech is done and if s2slt is clicked, Text-to-Sign language.

### 5. SYSTEM ARCHITECTURE

A look at how VaaniMeet is built, how it passes audio and generates output in audio and sign language forms. Built for smooth and flexible communication between multiple participants without any language barrier. Parts fit like puzzle pieces, shifting only when needed. Communication happens cleanly across layers, avoiding clutter or overload. As mentioned in Fig. 1 and Fig. 2, two steps are very common which are speech-to-text and text-to-text translation. This steps are very crucial as this steps are basic steps in this features

1. Language Identification (LID):  
LID used **faster-whisper** to identify speakers language. It is mainly used to detect speaker language which from audio files. It detects the language from initial 30ms of audio input by speakers
2. Speech-To-Text (STT):  
**faster-whisper** is used again in STT for transcribing audio into raw text string this text is in speakers' language which is later pass to TTT.
3. Text-To-Text (TTT):  
TTT uses **googletrans** for translating text from source which is passed on by STT into targeted language text using all NLP rules including semantic and syntactic analysis and many more which in helps in grammar and analyze sentiments within the text.

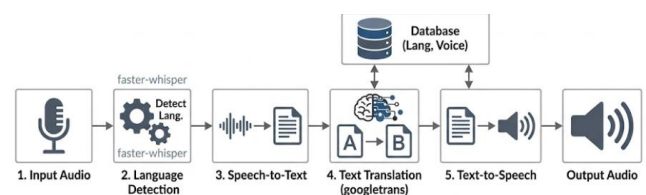


Fig.1 Block diagram of Speech-to-Speech Translation

4. Text-To-Speech:  
It used **edge-tts** which is Microsoft Edge Neural TTS used to convert translated text into high quality, human like neural voice in .mp3 file format for playback. This translated audio pickup native accent of language which makes it more like having communication with human rather than with ai. Fig.1 shows how step 1 to 4 works in flow and give audio output

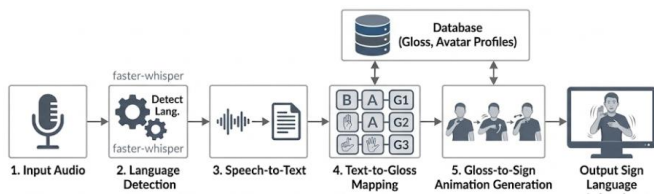


Fig.1 Block diagram of Speech-to-Sign Language Translation

5. Text-to-Gloss Mapping:  
Logic used behind this is keyword based NLP. After TTT it splits each word and finds the words in code if mentioned like in sentence-“hello, Good Morning to all”, here it splits the sentence to- ‘hello’, good-morning’, ‘to all’.
6. Gloss-to-Sign Animation Generation:  
Then it finds the keywords from code which is a predefined part in code and combine it into a long sentence.
7. Output Rendering:  
In output it shows the gifs in sequences like a sentence

6. METHODOLOGY AND IMPLEMENTATION

Fig.3 gives the layout of how features work throughout the system. A structure path is followed in this platform for both features. Using S2ST & S2SLT as main highlight the fig.3 indicates the flow of steps when user opt for speech translation or sign language translation or both. The flow is simple predictable of how when user click on buttons how it starts taking input and pass it through layers in backend and provide a flow less output.



Fig.3 Flow chart of VaaniMeet

Who uses VaaniMeet? Someone who needs translation but is not familiar with extensions or add-on services. Where this platform doesn't need extensions or add-on services as its default present.

When user login in user can join meeting using link, or unique id or even by creating a new meeting, additional option is scheduling meeting. After creating a meeting user/host can send link to other participants for joining. It has all other basic features which are very common in other platforms, like mic, camera, share screen, chat, list of participants.

Additional buttons are S2ST & S2SLT, which are for speech-to-speech translation and speech-to-sign language translation, which translates input audio into either audio in the language selected by the participant or translates into sign language in GIF form.

7. RESULTS AND DISCUSSION

A test checked whether all features within meeting lobby works and creates p2p connection within multiple participants joined in same meeting. This ensures the translations are smooth and visible and audible for listeners in the meeting lobby.

When tested, meeting was created and can be joined via unique id or link, and even schedule meetings. Details of schedule meetings are visible on home screen of platform. Multiple users join meeting using same link without lagging and p2p connect was visible when chat box was tested. And list of participants get updated as participant joins or leaves to all present participants.

The main features which are highlight of this platform which are s2st the translated audio is audible for other participants who are listeners in current meeting where as in s2slt is clicked it shows a container in corner of meeting screen which displays interpretation of audio. It translates audio into American sign language live for all participants.

One thing is clear that both translations work well and p2p connection which is the main for testing cause both speaker and listeners are important here. A shift happens when it ai replaces human, not discarding completely but lessen the cost of meeting when either speech translation or sign language translation is needed. It isn't any high end platform- its like a prototype which is meld in idea of making communication smooth regardless of obstacles.

Table 1. Lobby And Session Management

Feature	Testing Scenario	Observed Outcome
P2P	Multiple users	Stable P2P

<b>Lobby</b>	via unique ID/Link.	established; lobby remained lag-free.
<b>Scheduler</b>	Meeting creation and scheduling.	Schedule details correctly visible on home screen.
<b>UI Dynamic</b>	Participant list updates on join/leave.	Real-time tracking with zero noticeable latency.

**Table 2. Translation And Integration Performance**

Feature	Testing Scenario	Observed Outcome
<b>S2ST</b>	Real-time audio translation for listeners.	Translated audio was clear and synchronized.
<b>S2SLT</b>	Activation of visual sign container.	Live ASL rendered accurately in UI corner pane.
<b>Integration</b>	Chat and translation active simultaneously.	P2P stable; no conflict between data streams.

## 8. CONCLUSION AND FUTURE SCOPE

A platform that helps in smooth communication regardless of language barrier of disability. And overcoming language border helping people of different culture to collaborate and helping people of D/HH community to join and feel welcomed without any barrier and able to understand the speaker smoothly. Using VaaniMeet it just not translated live audio but also gives a confidence to express without feeling they might be says something wrong or misplace words instead of trying communicating in weak language experience trying VaaniMeet’s s2st and s2slt feature which also counter grammar and express correct sentiments.

One day maybe sign-to-speech translation button be visible in meeting lobby, which will translate sign language into audio giving chance with D/HH community to be speaker in meeting which will again solve one more realtime problem. And overcome the issue of opportunities for all people and cut the barrie regardless of ability of speaking language or express emotions or point to keep.

## REFERENCES

[1] T. Fukuda et al., “High-Fidelity Simultaneous Speech-To-Speech Translation (“Hibiki”) with Head-Start Decoding and Text-To-Speech Pre-Training,” arXiv preprint arXiv:2403.01836, 2024.

[2] A. N. Al-Mala'a and M. A. H. Al-Qaraawi, “SignEdgeLVM transformer model for enhanced sign language translation on edge devices,” *Neural Computing and Applications*, vol. 36, no. 1, pp. 581–600, 2024.

[3] M. H. El-kholy, M. B. E. Mohamed, and A. Mohamed, “Efficient Speech Translation with Dynamic Latent Perceivers,” in *Proc. ICASSP 2023 - 2023 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Rhodes Island, Greece, 2023, pp. 1–5.

[4] D. V. Ilyin, “Intelligent System for Automatic Bidirectional Sign Language Translation Based on Recognition and Synthesis of Audiovisual and Sign Speech,” *The Int. Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. XLVIII-2/W2-2023, pp. 325–331, 2023.

[5] A. D. Nguyen et al., “Textless Direct Speech-to-Speech Translation with Discrete Speech Representation,” in *Proc. ICASSP 2022 - 2022 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore, 2022, pp. 6427–6431.

[6] H. Inaguma, P. von Platen, K. Duh, and S. Watanabe, “Enhancing Speech-to-Speech Translation with Multiple TTS Targets,” in *Proc. ICASSP 2022 - 2022 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore, 2022, pp. 7837–7841.

[7] Y. Han, C. Wang, and J. Pino, “StreamSpeech: Simultaneous Speech-to-Speech Translation with Multi-task Learning,” in *Proc. 2022 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Seattle, WA, USA, 2022, pp. 4930–4941.

[8] P. Singh, A. Kumar, A. Singh, and S. S. Solanki, “Speech to Indian Sign Language Using Natural Language Processing,” *Future Internet*, vol. 14, no. 12, p. 370, 2022.

[9] M. Elbayad, A. Bérard, L. Besacier, and J. Niehues, “How ‘Real’ Is Your Real-Time Simultaneous Speech-to-Text Translation System?,” *Transactions of the Association for Computational Linguistics*, vol. 10, pp. 496–512, 2022.

[10] N. S. Kamble, M. K. Thounaojam, and S. R. Nirmale, “Sign language interpretation using machine learning and artificial intelligence,” in *Artificial Intelligence and Machine Learning for EDGE Computing*, Academic Press, 2022, pp. 241–260.

[11] Y. Jia et al., “Translatotron 2: High-quality direct speech-to-speech translation with voice preservation,” arXiv preprint arXiv:2107.08661, 2021.

[12] C. Liu, Q. Liu, Y. Wu, and M. Zhou, “Improving Speech-to-Speech Translation Through Unlabeled Text,” in *Proc. ICASSP 2021 - 2021 IEEE Int. Conf. on Acoustics, Speech and Signal*

Processing (ICASSP), Toronto, ON, Canada, 2021, pp. 6563–6567.

[13] M. Le, A. Chaudhary, A. Sarin, and D. Agrawal, "SimulTron: On-Device Simultaneous Speech to Speech Translation," arXiv preprint arXiv:2106.01242, 2021.

[14] B. Li, P. Koehn, R. Sennrich, B. Chen, and J. Ma, "Joint Training And Decoding for Multilingual End-to-End Simultaneous Speech Translation," in 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Cartagena, Colombia, 2021, pp. 919–926.

[15] A. Duarte, K. D. D. C. Santos, T. Braga, and S. Escalera, "Cross-modal neural sign language translation," Ph.D. dissertation, Dept. Comput. Sci., Univ. Politècnica de Catalunya, Barcelona, 2021.

[16] H. Liang, S. Wang, and Z. Liu, "Research on Speech Information to Sign Language Translation Based on 1D-GoogLeNet and LSTM," in Proc. 2020 4th Int. Conf. on Digital Signal Processing, Chengdu, China, 2020, pp. 110–114.

[17] S. Kumar, S. K. Singh, and A. K. Singh, "Speech to Sign Language Conversion Using Neuro Fuzzy Classifier," in Computational Intelligence: Theories, Applications and Future Directions - Volume I, 2019, pp. 165–177.

[18] S. A. G. Rizvi and A. R. Khan, "Efforts to Improve Avatar Technology for Sign Language Synthesis," in Proc. 3rd Int. Conf. on Information and Communication Technology for Competitive Strategies, Udaipur, India, 2018, pp. 1–6.

[19] A. Chaturvedi, V. P. Shukla, A. Tiwari, and R. Kala, "A Speech-driven Sign Language Avatar Animation System for Hearing Impaired Applications," in Proc. 24th Int. Joint Conf. on Artificial Intelligence (IJCAI 2015), Buenos Aires, Argentina, 2015, pp. 4230–4231.

[20] M. Hu, W. Li, and X. He, "Application of virtual human sign language translation based on speech recognition," in Proc. 2014 Int. Conf. on Virtual Reality and Visualization, Shenyang, China, 2014, pp. 195–199.

## BIOGRAPHIES



**MAITHILI KAMBLE**

Student, Computer Engineering MGM College of Engineering and Technology, Navi Mumbai, Maharashtra



**GARV BALGI**

Student, Computer Engineering MGM College of Engineering and Technology, Navi Mumbai, Maharashtra



**URJA MALI**

Student, Computer Engineering MGM College of Engineering and Technology, Navi Mumbai, Maharashtra



**PROF. SACHIN CHAVAN** Professor, Computer Engineering, MGM College of Engineering and Technology, Navi Mumbai, Maharashtra