AI-Powered Real-Time Speech-to-Speech Translation for Virtual Meetings Using Machine Learning Models

Karunya S
Department of Computer Science and Engineering
Sri Sairam Engineering College
Chennai, India
karunyasrinivasan2003@gmail.com

Jalakandeshwaran M
Department of Computer Science and Engineering
Sri Sairam Engineering College
Chennai, India
eshwarjalacoc@gmail.com

Thanuja Babu
Department of Computer Science and Engineering
Sri Sairam Engineering College
Chennai, India
tanujaaa1326@gmail.com

Uma R
Department of Computer Science and Engineering
Sri Sairam Engineering College
Chennai, India
uma.cse@sairam.edu.in

2023 Intelligent Computing and Control for Engineering and Business Systems (ICCEBS) | 979-8-3503-9458-0/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICCEBS58601.2023.10448600

***Abstract*—In our interconnected world, language diversity poses communication challenges, particularly in virtual meetings. Our solution, a Real-Time Speech-to-Speech Translation system for Virtual Meetings, bridges these gaps. It captures speech in one language, providing clear and understandable translations in real-time during virtual meetings. By seamlessly integrating Automatic Speech Recognition (ASR), Machine Translation (MT), and Text-to-Speech (TTS) components, this system transcends language barriers, enabling participants to engage effortlessly and effectively in multilingual virtual interactions. It's more than text; it fosters spoken interaction, revolutionizing cross-lingual communication in virtual meetings. Applications abound, from enhancing global business negotiations to aiding virtual travelers and connecting educators to broader international audiences of diverse languages in virtual educational platforms. In an era where virtual communication is paramount, our project empowers meaningful connections, proving technology's remarkable ability to unite people and transcend language barriers in virtual settings worldwide.**

Keywords – Language Barriers, Speech Recognition, Translation Technology, Machine Learning models, ASR, MT, TTS, Cross-Lingual Communication, Virtual Meetings.

I. INTRODUCTION

In an era of unprecedented digital connectivity and global interactions, the significance of effective communication transcends geographical boundaries. However, within this vast tapestry of interconnectedness, language diversity often poses intricate barriers to seamless dialogue, particularly in the realm of virtual meetings. The inability to effortlessly converse across linguistic divides can significantly impede the productivity and inclusivity of virtual meetings, hindering progress in both professional and personal contexts.

To confront this contemporary challenge head-on, we embark on a journey to introduce a pioneering solution - the Real-Time Speech-to-Speech Translation system for Virtual Meetings. This transformative project signifies a paradigm shift in communication technology, heralding a new era where linguistic disparities no longer impede the free flow of ideas and collaboration within virtual meeting spaces. Our innovative system is meticulously designed to transcend the limitations of traditional text-based translation by facilitating fluid and real-time spoken conversations during virtual meetings. It achieves this by harnessing cutting-edge speech recognition and machine translation technologies, and evaluation metrics that are precisely tailored for the virtual communication landscape. Through this venture, we envision a world where the boundaries of language no longer constrain virtual meetings, fostering a global community of collaboration and understanding.

In the pages that follow, we will delve into the intricate details of our Real-Time Speech-to-Speech Translation system for Virtual Meetings, exploring its development, applications, and the profound impact it promises to have on the way we communicate within the dynamic and interconnected sphere of virtual meetings.

II. EXISTING SYSTEM

Current speech-to-speech translation systems primarily rely on machine translation services and mobile applications. These systems enable users to speak in one language and receive real-time translations as shown in Fig. 1. However, they often face challenges in terms of accuracy, context-awareness, and seamless conversation flow, especially in virtual meetings. They may also require a constant internet connection for cloud-based translation services during virtual meetings.

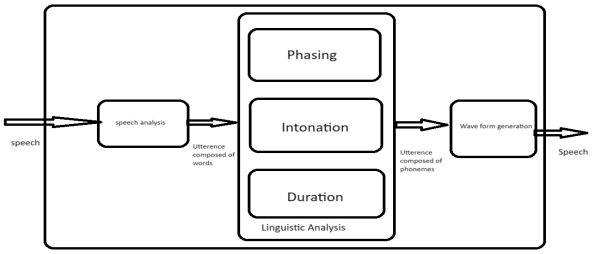

Fig. 1. Flowchart of Existing System

Our project seeks to build upon these existing systems by offering a more advanced and precise solution tailored explicitly for virtual meetings. We aim to enhance the accuracy and context-awareness of translations, enabling seamless cross-lingual conversations even in virtual settings with limited internet connectivity.

III. PROPOSED SYSTEM

Our proposed Real-Time Speech-to-Speech Translation system for Virtual Meetings will build upon the latest advancements in speech recognition and machine translation technologies, specifically designed for virtual communication. Here's an overview of its key components:

A. User-friendly Interface for Virtual Meetings

The system will feature a user-friendly interface accessible to individuals participating in virtual meetings with varying levels of technological expertise.

B. Secure Authentication for Virtual Meetings

To ensure authorized access during virtual meetings, the system will employ a robust authentication system, requiring users to have unique login credentials for virtual sessions.

C. Real-time Speech Recognition in Virtual Meetings

The core of our system will encompass state-of-the-art speech recognition capabilities tailored for virtual meetings. It will capture spoken input from users in one language during virtual meetings.

D. Precise Machine Translation for Virtual Meetings

We will integrate advanced machine translation algorithms that accurately convert the spoken input into the desired target language during virtual meetings. The system's strength lies in its precision and fluency, crucial for effective virtual communication.

E. Natural Speech Output in Virtual Meetings

Unlike traditional translation systems, ours will excel in delivering translated content as clear and natural speech during virtual meetings. This feature enables users to engage in real-time conversations with ease during virtual meetings.

F. Cross-platform Accessibility for Virtual Meetings

The system will be accessible across various virtual meeting platforms, ensuring convenience and widespread usability.

G. Offline Mode for Virtual Meetings

Recognizing the importance of accessibility, our system will include an offline mode for virtual meetings. Users can continue utilizing its translation capabilities even without a stable internet connection, making it valuable in remote or low-connectivity virtual meeting environments.

H. Customization Options for Virtual Meetings

Users will have the flexibility to customize the system to specific domains or industries, tailoring it to their unique virtual meeting needs.

I. Applications in Virtual Meetings

The proposed system's applications are diverse, including facilitating international business communication in virtual meetings, aiding virtual tourists, enhancing virtual educational outreach to global audiences, and fostering cross-lingual cultural exchange in virtual settings.

Our proposed Real-Time Speech-to-Speech Translation system for Virtual Meetings is poised to revolutionize cross-lingual communication in the context of virtual meetings. It represents a significant step toward a world where language barriers no longer impede meaningful interactions in virtual communication environments.

IV. LITERATURE SURVEY

All referenced papers primarily focus on speech-to-speech translation across various languages. Hence, in this study, we explored the convergence of speech-to-speech translation and virtual meeting platforms, enabling seamless multilingual communication.

[1] Prior studies have established the efficacy of word embeddings in enhancing ASR and ST models, providing valuable contextual and semantic information from textual data. Our research extends this by integrating advanced deep learning techniques like transformer-based architectures (e.g., BERT and GPT) to further optimize ASR and ST models. This approach aims to bridge the gap between spoken and textual languages, resulting in reduced word error rates in ASR, improved translation metrics in ST, and ultimately more accurate and efficient spoken-to-textual language conversion.

[2] End-to-end speech translation remains a challenge for syntactically distant language pairs due to long-distance reordering complexities. This study pioneers an attention-based encoder-decoder model for English-Japanese language pairs with differing word orders (SVO vs. SOV). To address the lack of parallel speech-text data, text-to-speech synthesis (TTS) is employed for data augmentation. The proposed model incorporates transcoding and curriculum learning (CL) strategies to guide the model, starting with ASR or MT tasks and gradually transitioning to end-to-end speech translation. Results indicate significant performance improvements compared to conventional cascade models, particularly for distant language pairs.

[3] Unsupervised Neural Machine Translation (UNMT) has achieved remarkable results, particularly for language pairs like French-English and German-English, through methods like unsupervised bilingual word embedding (UBWE) and cross-lingual masked language model (CMLM) pre-training. This paper empirically explores the relationships between UNMT and UBWE/CMLM, revealing that the quality of UBWE and CMLM significantly influences UNMT performance. To address this, the paper introduces a novel UNMT structure with cross-lingual language representation agreement, offering two approaches: UBWE agreement and CMLM agreement. These methods, including regularization and adversarial training, ensure the preservation of UBWE and CMLM quality during UNMT training. Experimental results across several language pairs demonstrate substantial improvements over conventional UNMT.

[4] Inspired by the limitations of existing neural machine translation (NMT) models in capturing alignment between input and output, our project introduces a valuable add-on to NMT technology. We propose an innovative approach that incorporates explicit phrase alignment into NMT models. This enhancement significantly improves NMT's interpretability, addressing issues related to transparency and model understanding. Moreover, our approach empowers NMT systems to effectively handle lexical and structural constraints, expanding their applicability to a wider range of translation tasks. Through our project, we contribute to advancing NMT technology, making it more versatile and interpretable, ultimately enhancing the quality of translations across various language pairs and domains.

[5] The paper explores training multilingual and multi-speaker text-to-speech (TTS) systems based on language families for Indian languages, addressing the challenges of linguistic diversity and data scarcity. However, it primarily focuses on training TTS systems and adaptation within language families. In our project, we aim to extend this approach to real-time speech-to-speech translation in virtual meetings, utilizing language family-based TTS models for natural and contextually relevant speech synthesis. Additionally, we will incorporate real-time translation capabilities to bridge language barriers in virtual meetings, creating a comprehensive communication solution. This holistic approach distinguishes our project from existing research and enhances the practicality of virtual meetings for diverse language users.

V. IMPLEMENTATION

Our Real-Time Speech-to-Speech Translation system for Virtual Meetings aims to revolutionize cross-lingual communication during virtual meetings. It integrates advanced technologies to capture, translate, and produce speech, ensuring a seamless flow of conversations across language barriers.

The system's initiation revolves around a user-friendly interface developed with HTML5, CSS3, and JavaScript. It's responsive, ensuring compatibility with various devices and screen sizes. User interface (UI) and user experience (UX) principles are followed for intuitive navigation and accessibility features for a seamless user experience.Prioritizing the utmost privacy and security, the authentication process verifies users' identities, ensuring that only authorized individuals can participate in virtual meetings.

We harnessed the GigaSpeech dataset as a foundational resource for training our system. Given the dataset's extensive audio recordings, we initiated a data preprocessing pipeline to optimize it for further training our models. This pipeline involved segmenting the lengthy audio recordings into shorter, coherent fragments, typically spanning a few seconds to a minute each. This segmentation was vital to ensure that our training models could efficiently handle real-time processing, a necessity for seamless virtual meetings. In addition to segmentation, we addressed transcription peculiarities within the

GigaSpeech dataset in order to enhance the reliability of our models. This included the removal of non-speech elements such as laughter, disfluencies, and background noise annotations. By eliminating these extraneous elements, we crafted transcriptions that portrayed clean and coherent speech. Furthermore, we applied rigorous text standardization techniques, effectively managing variations in punctuation, capitalization, and formatting. This standardization fostered consistency across the dataset, facilitating robust model training.

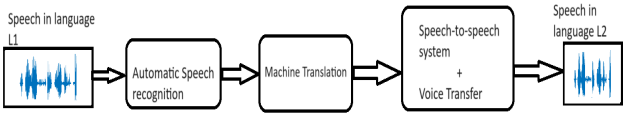

Fig. 2. Components of our translation system

Following the data processing phase, our system proceeds to the model training stage. We used Automatic Speech Recognition (ASR)[1], Machine translation (MT)[2], and Text-to-Speech (TTS)[2] models as seen in Fig. 2. In this step, we utilize the refined and segmented GigaSpeech dataset to train our ASR, MT and TTS models as seen in Fig. 3. The Automatic Speech Recognition (ASR) serves as the initial component of our system. It is trained to utilize speech recognition technology models to convert spoken language into textual transcripts. ASR takes the spoken input and transcribes it into text in the source language. These transcriptions serve as the foundation for the subsequent translation process. This model is designed to be highly accurate, capturing not only the words spoken but also nuances, accents, and variations in speech. We employ recurrent neural networks (RNNs)[1] and fine-tune them using the segmented audio data.

Once ASR transcribes the spoken content into text, the Machine translation (MT) model is used to translate the source input to text in the desired output language. This model receives the transcribed texts generated by the Automatic Speech Recognition (ASR) component, which converts spoken language into text, with a strong emphasis on precision and fluency. This MT model goes beyond mere word-by-word translation, considering contextual nuances, idiomatic expressions, and language flow for precise and fluent translations. It understands the relationships between words, ensuring that the translated content retains clarity and sounds natural to native speakers. This nuanced approach is crucial for effective cross-lingual communication, especially in the dynamic context of virtual meetings. To train our MT model, we rely on parallel data, consisting of source language transcripts and their corresponding translations in the targeted output language. This data forms the basis for the model to learn the intricate patterns and language structures necessary for accurate translation. The training process involves iteratively optimizing the model's parameters to minimize translation errors and maximize fluency.

In order to generate final vocal output in target language, we integrated a Text-to-Speech (TTS)[5] model into our system. The TTS model is responsible for converting the translated text generated by the MT model into natural and coherent speech in the target language. We employ deep neural networks and generative adversarial networks, to train our TTS model. This training process involves using the translated text from the MT model as input and generating corresponding speech waveforms as output. Fine-tuning and optimization are performed to ensure that the synthesized speech is clear, natural, and maintains appropriate intonation.

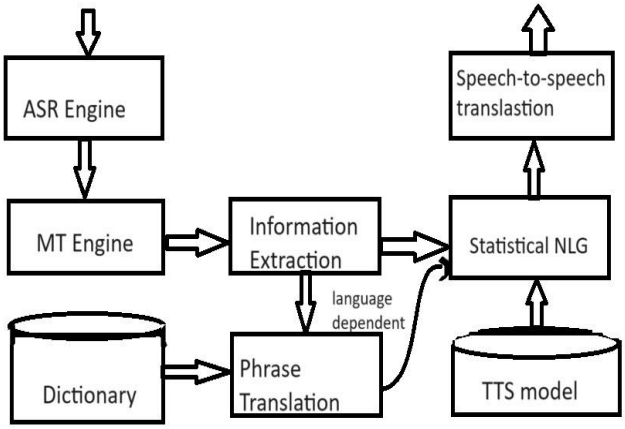

Fig. 3. Mechanism of Speech-to-Speech Translation

To optimize the performance of ASR, MT, and TTS models, extensive training iterations are conducted, fine-tuning model parameters and adjusting hyperparameters. We employ evaluation metrics, such as Word Error Rate (WER)[1] for ASR and Bilingual Evaluation Study (BLEU)[6] score for MT, to assess and enhance the model's accuracy and fluency. For TTS, we make use of Mean Opinion Score (MOS)[7] and Naturalness in order to assess the naturalness to ensure the quality of the synthesized speech. This iterative training approach ensures that our system becomes proficient in accurately transcribing and translating spoken language, thus achieving high-quality, human-like speech synthesis.

We develop RESTful APIs using the Flask framework in Python to enable communication between our translation system and the virtual meeting platform. We design API endpoints to accept audio data from the virtual meeting's microphone feed, translate spoken language during the virtual meeting, and then the API endpoints send back the translated audio to the virtual meeting platform in real time. The data exchanged between the virtual meeting platform and our system is typically in a structured JSON format, allowing easy parsing and interpretation. To ensure that translated speech aligns perfectly with the ongoing conversation, we

synchronize the audio streams. This involves timestamping audio data, so the translated content is delivered at the right moment, maintaining a natural conversation flow. Our system is designed to support multiple languages. Through language detection algorithms, it identifies the source language of the speaker and translates it into the chosen target language. We use language codes and recognition models to achieve accurate language detection. We integrate TTS engines that generate natural-sounding speech in the target language. Parameters such as pitch, speed, and voice type can be customized.

The virtual meeting platform's user interface is extended to include our translation features. Participants can select target languages, enable or disable translation, adjust volume levels, and view closed captions of translated content. To maintain a low-latency experience, we optimize our system for real-time processing. This includes efficient data transmission, minimal processing delays, and robust error handling. We implement an offline mode to handle situations with limited internet connectivity. Participants can still use our system with pre-loaded translation models and TTS voices.

Through this implementation, our system ensures that language diversity no longer hinders effective virtual communication. By bridging linguistic gaps, it empowers users to engage confidently in meaningful interactions and collaborations within the virtual meeting space.

VI. RESULTS

Upon seamless integration into virtual meetings, our real-time speech-to-speech translation system yields a range of highly valuable outputs that significantly enhance the virtual meeting experience.

Our system facilitates secure participation within the virtual realm, offering real-time translation of spoken content. As the participants converse in their native languages, the system diligently transcribes, translates, and articulates their words in the chosen target languages. This effortless cross-lingual communication ensures that language differences do not hinder the effectiveness of conference meetings.

Participants are ensured that they can engage in fluid and comprehensible conversations by maintaining a high degree of naturalness and fluency in the translated content. Unlike conventional text-based translations, our system generates spoken translations that sound clear and coherent, mimicking human speech patterns.

Users are granted a spectrum of customization options to tailor the translation process to their preferences. These include language selection, adjustment of speech speed, pitch modulation, voice type, and the ability to activate or deactivate translation features. Such flexibility empowers users to adapt their virtual meeting experience according to their individual needs and desires.

In addition, our system generates textual closed captions within the virtual meeting interface. This feature is particularly advantageous for participants who prefer reading translations or those with hearing impairments, ensuring inclusivity and accessibility. The translated content seamlessly aligns with the ongoing discourse, reducing interruptions and preserving the natural flow of conversation.This synchronization ensures the fluidity of conversations.

Offline mode feature enables participants to continue benefiting from translation capabilities in environments with limited or no internet connectivity, relying on pre-loaded translation models and TTS voices.

Ultimately, our system's outcomes culminate in enriched virtual meeting interactions and collaborations. Language diversity ceases to be a barrier, empowering participants to engage confidently and proficiently, transcending linguistic boundaries.

VII. CONCLUSION

In the realm of virtual meetings, our Real-Time Speech-to-Speech Translation system marks a groundbreaking stride in breaking down language barriers. It empowers seamless multilingual conversations, uniting participants worldwide. This technology transcends borders, enabling effective global communication for business, education, and cultural exchange. Our project underscores the pivotal role of technology in fostering connections and meaningful interactions. In a world where virtual meetings dominate, it ensures that language never hinders the exchange of ideas and collaboration among diverse participants.

VIII. FUTURE WORKS

Future work in Real-Time Speech-to-Speech Translation for Virtual Meetings includes enhancing translation accuracy, expanding language support, and optimizing resource usage. Efforts should target adaptability to diverse accents, incorporating more languages, and supporting real-time voice recognition for multiple speakers. Integrating sentiment analysis and real-time subtitles for accessibility are promising additions. Collaboration with virtual meeting platforms can make this technology widely accessible, transforming virtual meetings into inclusive global forums, promising an even more seamless and inclusive virtual communication experience.

References

[1] S.P.Chuang, A.H.Liu, T.W.Sung, and H.y.Lee, "Improving Automatic Speech Recognition and Speech Translation via Word Embedding Prediction," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 29, 2021.

[2] T.Kano, S.Sakti, and S.Nakamura,"End-to-End Speech Translation With Transcoding by Multi-Task Learning for Distant Language Pairs," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 28, 2020.

[3] H.Sun, R.Wang, K.Chen, M.Utiyama, E.Sumita, and T.Zhao,"Unsupervised Neural Machine Translation With Cross-Lingual Language Representation Agreement," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 28, 2020.

[4] J.Zhang, H.Luan, M.Sun,F.Zhai, J.Xu, and Y.Liu,"Neural Machine Translation With Explicit Phrase Alignment," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 29, 2021, pp. 1001-1011.

[5] A.Prakash and H.A.Murthy,"Exploring the Role of Language Families for Building Indic Speech Synthesisers," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 31, pp. 734-747, 19 December 2022.

[6] Hajar Chatoui and Oğuz Ata, " Automated Evaluation of the Virtual Assistant in Bleu and Rouge Scores" , in 2021 3rd International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA).

[7] Babak Naderi and Sebastian Möller , " Transformation of Mean Opinion Scores to Avoid Misleading of Ranked Based Statistical Techniques" in 2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX).