International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 01 | Jan - 2025 SJIF Rating: 8.448 ISSN: 2582-3930

Speech Translation Technology In Chatting And Video Conference Platform

Nansi Jain
Prateek Maurya, Mohammad Umar, Pratik Raj
IPEC, Dept. of CSE(DS)

1. Abstract

Speech translation is an important technology when it comes to chatting and video conferencing platforms. It has proved itself to be a transformative tool to bridge language barriers, making cross cultural, locale and region communication possible. Its ability to perform in real time irrespective of the linguistic state of the input makes it highly useful. This technology requires multiple levels of processing, among which main processing is done by a language model which has multiple stage including Natural Language Processing (NLP), Speech Recognition, Machine Learning Models (MLM). These together makes the translation possible. Aside from this basic network infrastructure is required to support the transmission of chat and video over the internet. All these related technologies are being developed from a long time, over 20 years or more. But it has been just a few years that it has become accurate enough to be useful to a commercial user. The earliest studies started at Advance Telecommunications Research Institute (ATR) in Japan. It involved features like voice to text transcription, multi-language translation with support for variety of languages facilitating seamless interaction between individuals from different origins. With the increase in remote work and collaboration in international communities, this technology plays a vital role in simplifying communication and resulting in improved business outcomes and better products. This enhances user experience and removes a lot of hassle. Modern chat applications are easily compatible with this and with the technology like Web3 this translation model can be easily integrated with already existing communication infrastructure. There are yet a lot of challenges in developing something with so much vast and widespread application. Biggest challenge for this can be maintaining the contextual understanding, managing accents and ensure secure transmission to avoid any breach. We have been continuously improving on this with rapidly growing technology but as there is always room for more, this journey is going to be long. But for sure, future applications of this approach is going to have big impact on how people interact with each other without needing to think about language barriers.

2. Literature Review

Speech translation technology has revolutionized multilingual communication by combining Automatic Speech Recognition (ASR), Machine Translation (MT), and Text-to-Speech (TTS) synthesis. Early initiatives like NEC Co. Ltd. and ATR laid the groundwork, enabling real-time translation and fostering international collaboration. Recent advancements, such as neural machine translation and large datasets, have improved accuracy and natural speech output, as highlighted by Liu et al. (2022) and Kumar et al. (2022).

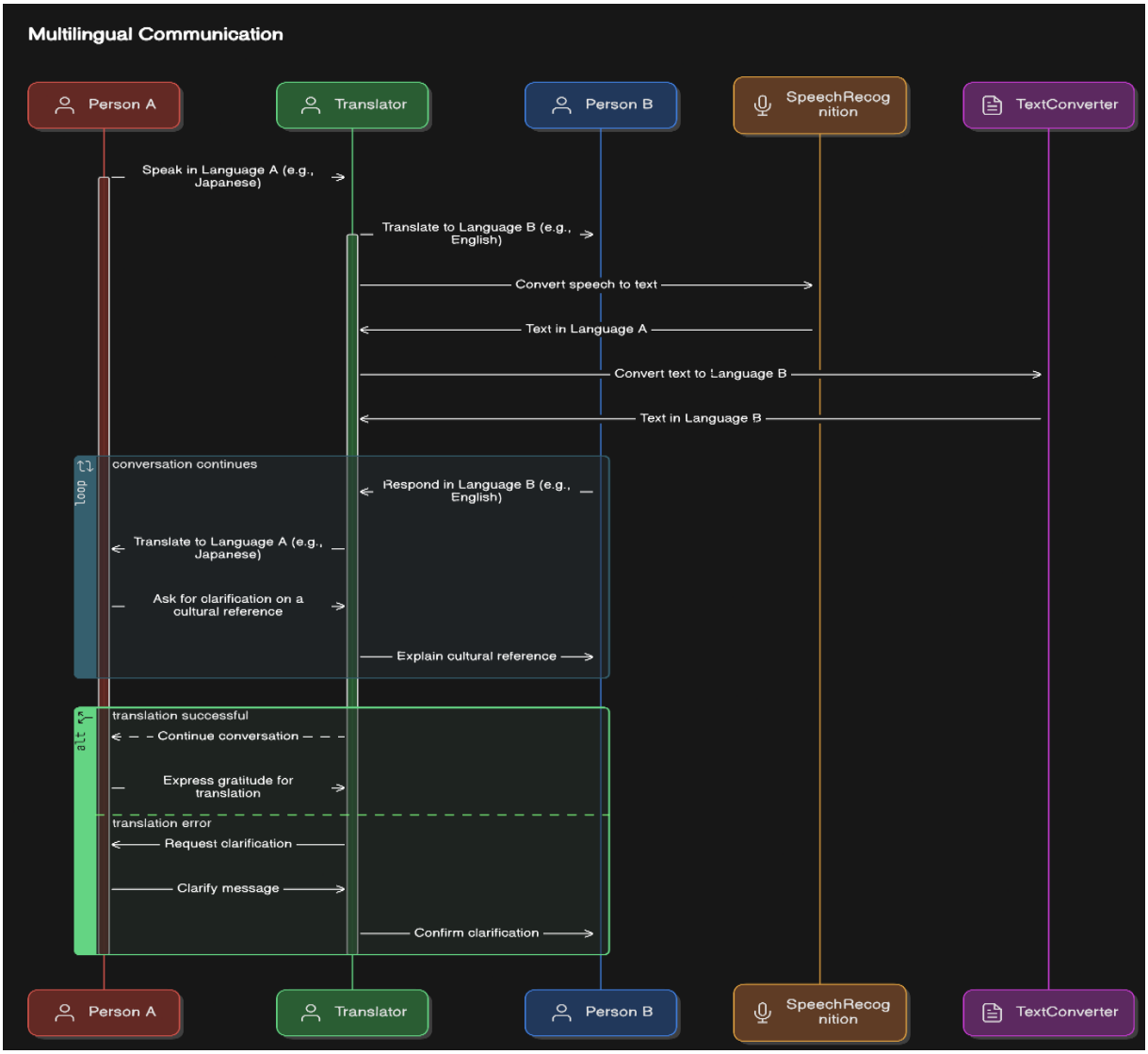
Despite these strides, challenges like maintaining context, handling accents, and ensuring security persist. Studies, including those by Doherty (2016) and Shadiev et al. (2017), emphasize addressing variability in real-world conditions. Ongoing AI and machine learning integration promise further improvements, making speech translation indispensable for global communication in business, healthcare, and beyond.

3. Keywords

Speech to speech translation (S2ST), Automatic Speech Recognition (ASR), Machine Learning Model (MLM), Web Real Time Communication (WebRTC).

3. Introduction

The rapid growth in colonization and industrialization has brought a need for inter language communication and to counter the same various solutions are being developed. There had been traditional methods like human interpretations and machine translation, but these have their own limitations where some is lacking the scalability and while other is facing accuracy. Speech Translation Technologies have been emerging to counter this situation and trying every possible use of technology to



bring the best possible solution. Technologies like real time video conferencing with multilanguage support has been one of the biggest achievements but it still lacks the level of accuracy needed for it to be reliable enough to be used commercially.

On the other hand, a different speech translation technology has been evolving in Japan for last 20 years. During the days of initial efforts, NEC Co. Ltd. Gave a demonstration of a prototype for speech translation system at Telecom'83 and ATR Interpreting Telephony Research Laboratory developed another system named ASURA. These corporations conducted international speech translation experiments, leading to improved research activities and more sophisticated processing methods. After this several speech translation projects were started and some research institutes such as CMU, SRI, AT&T, KDD, ETRI, and Korean Telecom (KT)

developed their own speech translation systems for various target pairs (e.g., JE, EJ, etc.) [2].

STT combines Automatic Speech Recognition (ASR), Machine Translation (MT), and Text-to-Speech (TTS) to create a seamless translation between languages. The technology has way more implications for global communication, enabling people to communicate across languages and cultures in real-time.

4. Overview of speech translation technology and performance

4.1. Multilingual speech translation processing architecture

Figure 1 shows the overall architecture of the speech-translation system. Figure 1 illustrates an example where a spoken Japanese utterance is recognized and converted into Japanese text; this is then translated into English text, which is synthesized into English speech. The multilingual speech recognition system compares the input speech with a phonological model which consists of a large number of speech data from multiple speakers from different origins. It then converts the voice input into a string of phonemes represented in the Japanese katakana syllabary. Further, this string is processed and converted into a string in Japanese writing achieve the specific result to distinguish every user then we need a training model for every individual. But this data can easily be created as these models don't require much training and will improve as they are used more and more. We can use "Corpus Based Speech Recognitions and Synthesis" [3] to easily generate our speech models and accurately convert the string. Once the model is ready the mapping is required to generate the voice from string. This can be the final result and every model used in the system will continue to improve making the model more and more accurate over the time.

system (mixture of kana and kanji characters), this increases the probability of the string words. In this conversion, Japanese utterance is generated as per the probability occurrence of appropriate words in the given string. This is done using an engine trained on large dataset of Japanese text. These words are then translated by conventional methods to replace each occurrence of word into English word. The English words are then rearranged to make more meaningful results. This is done by comparing the actual meaning of the input text the generated text. Every comparison brings more and more improvement until the desired accuracy is reached. During the initial period of the model, it can take long time to generate translation as it may require multiple attempts for reaching accuracy, but as it continuous to learn and grow it can reach the expected results in just a few attempts making it really useful. After all the translation is over the process of generating the speech using the translated string begins.

To start with the Natural Language Generation (NLG), a pre-trained model for language is required. This model contains data for mapping required to generate verbal results. This model is usually a common for every individual and may not be generate speech in the user's voice. If we want to

4.2. Comparative study with human speech translation capability

Recently significant improvements has been developed in the field of multilingual speech technology. This has a huge impact in speech recognition, dialog systems and speech summarization. These technologies are being heavily used in various applications including mobile devices, online systems and multiple IOT services. The mobile system exemplifies a robust multilingual speech-to-speech translation system with the ability to handle multiple dialogs in various languages. It typically involves several key components: automatic speech recognition (ASR), machine translation (MT), and text-to-

speech synthesis (TTS). These models are often combined to unified system which facilitated seamless Multilanguage translation into a unified system to facilitate seamless translation between languages. For example, the ATR system uses a corpus-based statistical machine learning framework to achieve high-quality translations between English and Asian languages.

It is theoretically a big problem to evaluate the accuracy of a translation system. If we include evaluation of the speech synthesis module, the evaluation is done by feeding multiple test strings into the system to evaluate the quality of output [3]. Human translators have innate ability to understand and translate the speech with high accuracy while maintaining the emotional understanding and deep contextual meaning hidden behind the words. This capability enables them to handle every minor details about the sentences and handle important details like nuances, idiomatic expressions and contextual meanings. These are the scenarios where a machine automated translator system usually suffers and can produce highly accurate results. Humans are also easily adaptable to various forms of same language and patterns used in different regions of the world. Humans also manage real-time translation during quick communications during a meet or speech such as conferences or meetings, where maintaining flow and coherence is very crucial.

Multilingual speech technology and human speech translation capability have their respective strengths and weaknesses. While multilingual speech technology excels in scalability and speed while maintaining fairly low costs. On the other hand humans can offer high accuracy, emotional consistency, flow of sentences, contextual meaning, etc. The choice between the two depends on the use-case, in case of critical communications human translation systems should be utilized meanwhile for non-critical communications which are usually large scale we should be using the automated system, this choice will be better in every way by providing accuracy when needed and making the models learn more by being used more and more.

5. Result

Pilot research was conducted at the Faculty of Humanities and Social Sciences, University of Zagreb among the students of information sciences or language study groups. In the research 91 students participated, among which 58.2% at undergraduate and 41.8% at graduate levels. Out of the total number of students, 87.3% of interviewed students study information sciences and language

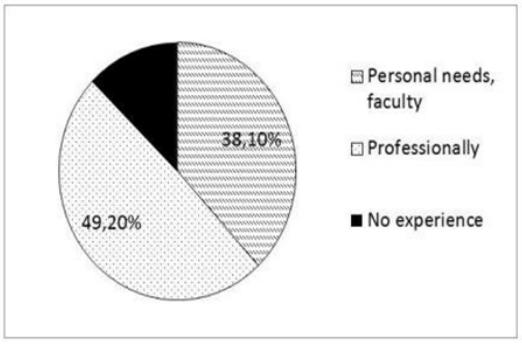


Figure 2. Previous experience in translation

group. Among the interviewed students, 77.3% of students have declared to have previous experience with translation, out of which 38.10% for personal and faculty needs, 49.2% professionally and 12.7% have no experience in translation, as presented in figure 2.

Answering the question on the type of text translated by free Internet translation tool, i.e. Google Translate, which is available for Croatian, the students gave

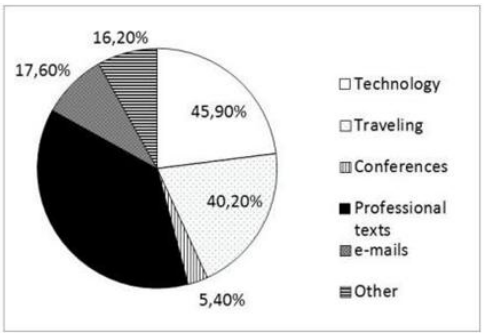


Figure 3. Types of text translated by free Internet translation tools

the following answers (MCQs): 74.3% of students use it for the translation of domain specific texts (economy, law, etc.), 45.9% technology domain, 40.0% traveling, 17.6% e-mail, Advances in speech and language research have brought speech translation close to the practical level for simple topics where there is a relatively clear value of use. At the current level, however, speech translation has only reached the stage of creating the core technologies. In order to achieve more sophisticated speech translation, 5.4% conference topics and 16.2% other (homework, games, literature or 2 do not use it) [5].

6. Conclusion

Advancements in speech and language research have brought speech to be more practically usable but still have compatibility only to simple level of translations. To make translation systems more capable and accurate results, further research and development is required.

Ongoing improvements in AI, machine learning, and natural language processing promise to make systems more accurate, faster, and bringing the capability to handle wider range of languages and contexts.

In conclusion, while human translation still plays an essential role in certain specialized contexts, speech translation technology is rapidly becoming a vital tool for breaking down language barriers in everyday communication. As this technology continues to evolve, its impact on global communication, business, and personal interactions will only grow, making it a cornerstone of digital communication in the 21st century.

Reference

[1] Hudelson, P., & Chappuis, F. (2024). Using Voice-to-Voice Machine Translation to Overcome Language Barriers in Clinical Communication: An Exploratory Study.

[2] Liu, Y., Zhang, J., & Xiong, H. (2022). Seamless multilingual communication: Recent advancements in speech translation technologies.

[3] Kumar, R., Gupta, M., & Sapra, S. R. (2022). Neural network models and datasets for enhancing speech translation accuracy.

[4] Nakamura, S., Sakti, S., & Kano, T. (2021). Integration of Speech Recognition, Machine Translation, and Speech Synthesis for Better Translation Quality.

[5] Millett, P. (2021). Continuous Improvements in Data Quality, Algorithmic Efficiency, and Real-World Testing in Speech Translation Systems.

[6] Liu, Y., Zhang, J., & Xiong, H. (2020). Specialized Translation with Automatic Speech Recognition, Machine Translation, and Text-to-Speech Integration.

[7] Al Shamsi, H., Almutairi, A., & Al Mashrafi, S. (2020). Natural-Sounding Speech Synthesis and its Impacts on Translation Practices.

[8] Liu, Y., Zhang, J., & Xiong, H. (2020). Integration of Speech Recognition, Machine Translation, and Speech Synthesis for fluent multilingual communication.

[9] Shadiev, R., Wu, T.-T., & Sun, A. (2017). Multilingual Datasets for Training Deep Learning Models in Speech Recognition.

[10] Doherty, S. (2016). Evaluating Translation Systems in Real-World Conditions: Accuracy across Accents and Variabilities.