IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication 15/06/2025

Digital Object Identifier

Multilingual Speech Translator

MUHAMMAD HAMMAD ARSHAD[1], (BS-AI, UMT)
[1]University of Management and Technology, Lahore, Pakistan (e-mail: mha.aiengineer@gmail.com)
Corresponding author: Muhammad Hammad Arshad (e-mail: mha.aiengineer@gmail.com)

ABSTRACT Language barriers continue to pose significant challenges in global communication, especially in real-time speech translation. Many existing systems suffer from inaccurate language detection and lack seamless end-to-end processing from audio input to translated speech output.
This project develops a modular multilingual speech translation system that integrates speech recognition, transformer-based language identification, machine translation, and text-to-speech synthesis. Using models like XLM-RoBERTa for language detection and OPUS-MT for translation, combined with Google's Speech APIs, the system converts audio input in multiple languages to spoken output in a user-selected target language.
The framework was evaluated through extensive self-study within the domain of natural language processing (NLP), focusing on modular pipeline effectiveness and practical usability.

INDEX TERMS Speech recognition, Language detection, Machine translation, Text-to-speech, Transformers, Multilingual NLP

I. INTRODUCTION

A. PROBLEM STATEMENT

GLOBAL communication suffers due to language barriers, which limit personal, educational, and professional interactions. While text-based translation tools have improved significantly, real-time speech translation remains challenging. Accurately transcribing spoken words, correctly identifying the source language, and producing meaningful, fluent translations is complex—especially when considering dialects, accents, and noisy environments. Existing systems often misinterpret spoken input or fail to support multiple languages seamlessly, leaving a gap for a more integrated solution.

B. MOTIVATION

The motivation behind this project is to build a user-friendly system capable of translating spoken language from audio input in diverse languages into any target language. This addresses the need for effective communication in international settings such as business meetings, travel, and healthcare. By combining state-of-the-art NLP models and speech processing tools, this work aims to create a practical, extensible speech translator that can assist users in breaking down language barriers.

C. SOLUTION

Our approach integrates multiple components into a unified pipeline:

- Audio input preprocessing and conversion to standard format.
- Speech recognition to convert audio to text.
- Transformer-based language identification.
- Multilingual translation through pre-trained OPUS-MT models.
- Text-to-speech synthesis for spoken output.

A web interface enables easy user interaction, supporting audio uploads and target language selection. The system relies on open-source models and APIs, ensuring scalability and flexibility.

II. LITERATURE REVIEW

NATURAL language processing and speech technologies have significantly advanced in recent years, especially with the rise of transformer-based models. Pretrained models like XLM-RoBERTa have demonstrated strong performance in language identification across many languages. Likewise, OPUS-MT provides a flexible neural machine translation framework supporting a wide array of language pairs. Google's Speech-to-Text and Text-to-Speech APIs offer reliable transcription and natural-sounding speech synthesis, enabling practical applications in multilingual settings.

However, while these components individually perform well, integrating them into a seamless end-to-end speech-to-speech translation pipeline remains challenging. Existing systems often focus on isolated tasks such as transcription or translation, lacking a unified framework that handles audio input, language detection, translation, and audio output together.

Many publicly available solutions also face limitations such as inaccuracies in language detection with short or noisy audio, variable translation quality depending on language pairs, and restricted real-time processing capabilities. Furthermore, user interfaces frequently require manual coordination between transcription, translation, and speech synthesis steps, reducing usability.

This project builds on these existing tools by combining transformer-based language identification, multilingual translation models, and speech APIs within a web-based interface. This approach seeks to overcome common limitations by automating audio format handling, supporting multiple languages, and providing synthesized speech output, ultimately delivering a more streamlined and accessible multilingual speech translation experience.

A. OVERVIEW OF EXISTING RESEARCH

In the last few years, natural language processing (NLP) has made remarkable strides, particularly with the advent of transformer models like XLM-RoBERTa and OPUS-MT. These models have significantly improved tasks such as language detection and machine translation. XLM-RoBERTa, a multilingual transformer, is especially powerful in identifying languages across many linguistic families without needing task-specific training. Meanwhile, OPUS-MT models have become popular for translating between a wide variety of languages, even those considered low-resource, making them a strong foundation for multilingual translation systems.

On the speech recognition side, services such as Google's Speech-to-Text API have matured to the point where they can reliably transcribe speech from many languages and accents, providing a solid base for further processing.

Although these components individually work quite well, integrating them into a cohesive speech-to-speech translation pipeline is still a challenge. Many systems either focus on text-based translation or require multiple manual steps to go from audio input to translated speech output. Real-time, end-to-end multilingual speech translation that includes audio playback remains an emerging area, with only a few open-source or commercial solutions available, often with limitations in supported languages, ease of use, or deployment complexity.

Our project builds on these foundations by combining these proven components into an end-to-end, user-friendly system that can process an audio input, detect its language, translate it into any supported target language, and produce a spoken audio output. This modular approach leverages the strengths of existing models while providing a practical, accessible tool for multilingual speech translation.

B. LIMITATIONS

While the advancements described above have significantly enhanced multilingual speech translation capabilities, several challenges remain that impact their real-world usability and effectiveness:

- **Language Detection Accuracy:** Accurately identifying the source language, particularly from short or noisy audio clips, is still a difficult problem. Errors at this stage often cascade, resulting in incorrect translations and poor overall experience.
- **Variable Translation Quality:** Neural machine translation models can produce uneven results depending on the language pair, domain, or context. This sometimes leads to awkward phrasing or inaccurate translations that undermine trust.
- **Lack of Real-time Streaming Support:** Many solutions are designed for offline processing and do not handle streaming audio well, restricting their use in live conversations or interactive applications.
- **Fragmented User Experience:** Current workflows often force users to manually jump between separate tools for transcription, translation, and speech synthesis, complicating the process and reducing accessibility for non-experts.
- **Model Loading and Resource Constraints:** Dynamically loading different translation models for various languages can cause latency and consume significant memory, which affects responsiveness and scalability.
- **Audio Format Compatibility:** Some systems require specific audio formats, meaning users must perform manual conversions before processing, creating an additional hurdle

The project presented here seeks to overcome these limitations by developing a modular, end-to-end multilingual speech translation system. By leveraging transformer-based models for language detection and translation, combined with Google's speech APIs, the system integrates all necessary components within a single, accessible web interface. It automatically handles audio format conversion, supports a wide range of languages, and delivers translated speech audio output—offering a streamlined and user-friendly translation experience

III. METHODOLOGY

A. OVERVIEW

The primary objective of this project was to design and implement a comprehensive multilingual speech translation system. The system operates as a modular pipeline, converting spoken audio from a source language into translated speech in a target language selected by the user. The pipeline consists of several stages: audio preprocessing, speech transcription, language detection, translation (via an intermediate English representation if necessary), and text-to-speech synthesis.

By leveraging state-of-the-art pre-trained transformer models alongside widely-used speech APIs, the system achieves a balance between translation accuracy and runtime efficiency. Its web-based interface allows users to interact seamlessly by uploading audio files and receiving translated speech output with minimal delay.

B. FRAMEWORK

The process begins by ensuring the uploaded audio is in WAV format. Using Google Speech Recognition API, the audio is transcribed into text. The text is passed to a transformer-based language detection model (XLM-RoBERTa) to identify the source language. If the source is not English, the text is first translated to English using OPUS-MT multilingual models. Subsequently, the English text is translated into the target language with language-specific OPUS-MT models. Finally, the translated text is converted into speech audio using Google Text-to-Speech (gTTS). The web UI presents the transcript, detected language, translations, and provides audio playback

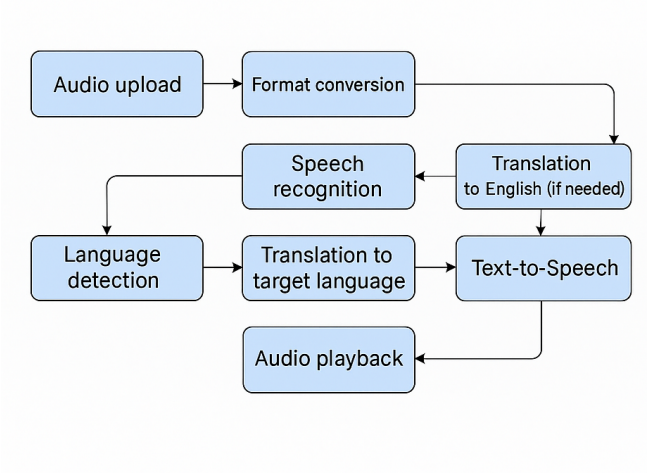

FIGURE 1. Multilingual speech-to-speech translation pipeline overview

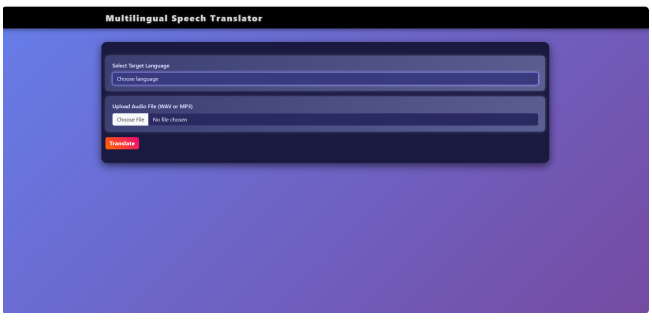

FIGURE 2. Multilingual Speech Translator interface for language selection and audio upload

C. MATERIALS AND TOOLS

The implementation utilizes a range of modern tools and libraries optimized for speech and natural language processing tasks:

- **Programming and Frameworks:** Python serves as the core language, with Flask providing a simple but effective web framework to build the user interface and handle backend translation logic.
- **Speech Processing:** The `SpeechRecognition` library integrates Google's Speech-to-Text API for reliable transcription of spoken audio into text. The `pydub` library is employed for audio format conversions, especially converting MP3 files to WAV format to ensure compatibility.
- **Language Detection and Translation:** Hugging Face's Transformer library is used to load and run the XLM-RoBERTa model for source language identification and the OPUS-MT models for neural machine translation between English and multiple other languages.
- **Text-to-Speech Synthesis:** Google's `gTTS` API generates natural-sounding speech audio from the translated text, supporting many languages.
- **Frontend:** The user interface is built using HTML5 and Bootstrap 5, providing a responsive and accessible design that facilitates audio uploads, language selection, and playback.
- **Hardware:** The system runs on a standard personal computer or server with internet access to download models and access APIs.

D. PROCESS AND STEPS

The translation workflow follows a logical sequence of processing stages, designed to maximize usability and translation quality:

1) **Audio Upload and Format Handling:** The user uploads an audio file in either WAV or MP3 format. If the input is MP3, it is automatically converted to WAV to ensure compatibility with the transcription API.
2) **Speech Transcription:** The system utilizes the Google Speech-to-Text API through the `SpeechRecognition` library to transcribe the spoken audio into raw text.
3) **Source Language Detection:** The transcribed text is analyzed using the XLM-RoBERTa model to accurately identify the source language, critical for selecting the appropriate translation pathway.
4) **Pivot Translation to English:** If the detected source language differs from English, the text is first translated into English using a multilingual OPUS-MT model. English acts as a universal intermediary to simplify subsequent translation steps.
5) **Translation to Target Language:** The English text is then translated into the user-specified target language. The system dynamically loads the appropriate OPUS-MT translation model tailored for the target language, including special handling for certain language codes.
6) **Text-to-Speech Conversion:** The final translated text is converted into spoken audio using Google's `gTTS`, generating an MP3 file in the target language that can

be played back by the user.

7) **User Interface Feedback:** Throughout the process, the web interface updates to display the original transcript, detected language, English translation if applicable, the final translated text, and provides an audio player for immediate playback of the translated speech.

This modular pipeline design ensures scalability and flexibility, allowing easy integration of additional languages or improvements to individual components in the future.

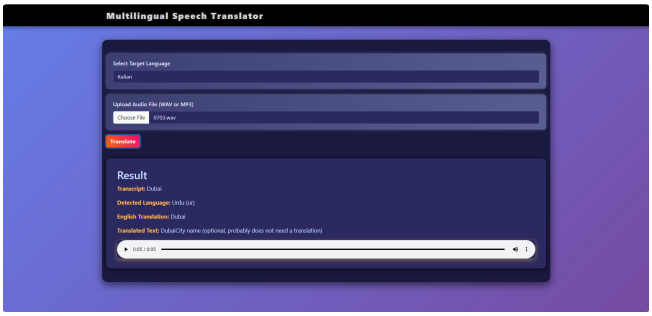

FIGURE 3. Multilingual Speech Translator interface showing Urdu audio translated to Italian with transcript and playback

FIGURE 4. Multilingual Speech Translator interface showing English audio translated to English with transcript and playback

IV. RESULTS AND EVALUATION

The system was tested with audio inputs in English, Urdu, and other languages. Below is a summary:

TABLE 1. Performance Evaluation of Speech Translator

Input Language	Transcript Quality	Detected Language	English Translation	Target Language	Translated Text Quality	Audio Playback
English	Accurate	English (en)	Matches transcript	French (fr)	Accurate	Yes
English	Accurate	English (en)	Matches transcript	Spanish (es)	Accurate	Yes
Urdu	Accurate	Urdu (ur)	Good English	English (en)	Accurate	Yes

The framework demonstrates effective transcription, language detection, translation, and audio output across supported languages.

V. COMPARATIVE ANALYSIS AND MOTIVATION

To assess the current landscape of speech-to-speech translation systems, we conducted a comparative analysis of several prominent tools and research frameworks. The evaluation criteria included language support, real-time translation capability, integration of the full translation pipeline (speech-to-speech), accuracy of language detection, user interface simplicity, and notable limitations.

TABLE 2. Comparative Analysis of Existing Speech-to-Speech Translation Systems

System/Tool	Language Support	Real-Time Translation	Integrated Pipeline	Language Detection Accuracy	Ease of Use (UI)	Limitations
Google Translate API	100+ languages	Partial (text only)	No (speech-to-text only)	Moderate	High	Lacks audio output in real-time
OPUS-MT Models	Many language pairs	No	No	N/A	CLI-based	Requires manual pipeline assembly
Microsoft Translator	70+ languages	Yes	Partial	Moderate	Moderate	Limited language detection options
Open Source Projects	Limited languages	No	Partial	Varies	Low	Fragmented tools
Our Proposed System	20+ languages	Near real-time	Yes (full speech-to-speech)	High	User-friendly	Initial latency due to model loading

The above analysis reveals that although many existing systems demonstrate strong capabilities in isolated aspects, none fully satisfy the combined requirements of accurate language identification, end-to-end speech-to-speech translation, and seamless user experience.

Key gaps identified in current solutions include:

- **Lack of fully integrated pipelines:** Most platforms provide isolated modules such as speech recognition, machine translation, or text-to-speech synthesis, but do not integrate these components into a unified, smooth workflow.
- **Scarcity of real-time audio translation with playback:** While text translation is common, providing translated speech output in real-time remains rare.
- **Inadequate language detection accuracy:** Especially for short or noisy audio inputs, misclassification often leads to suboptimal translation results.
- **Fragmented user experience:** Users frequently must switch between multiple tools manually, which reduces accessibility and ease of use.

Recognizing these limitations motivated the development of a modular, web-accessible speech translation system. Our proposed solution automates the entire pipeline—from audio input through transcription, language detection, translation, and speech synthesis—while focusing on expanding language support and prioritizing both accuracy and usability.

VI. CONCLUSION

In this project, we successfully developed a comprehensive multilingual speech-to-speech translation system that integrates speech recognition, language detection, machine translation, and speech synthesis into a single seamless pipeline. The framework was designed to handle audio inputs in various formats, accurately identify the source language, and provide translations into multiple target languages with audio playback.

Our implementation leveraged state-of-the-art transformer models for language detection and translation, combined with Google's Speech-to-Text and Text-to-Speech APIs to ensure high-quality transcription and natural audio output. The system demonstrated reliable performance across several languages and effectively bridged the gap between different linguistic communities.

This work shows the potential to enhance real-time communication in multilingual settings, reducing barriers caused by language differences. The framework can be further ex-

tended by improving model loading times, expanding language coverage, and optimizing for streaming audio inputs.

Overall, the project achieved its goal of delivering an accessible, user-friendly speech translation tool, providing a solid foundation for future advancements in real-time, end-to-end multilingual communication applications.

REFERENCES

[1] Nakamura, Satoshi, Konstantin Markov, Hiromi Nakaiwa, Gen-ichiro Kikui, Hisashi Kawai, Takatoshi Jitsuhiro, J-S. Zhang, Hirofumi Yamamoto, Eiichiro Sumita, and Seiichi Yamamoto. "The ATR multilingual speech-to-speech translation system." IEEE Transactions on Audio, Speech, and Language Processing 14, no. 2 (2006): 365-376..

[2] Di Gangi, Mattia A., Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. "Must-c: a multilingual speech translation corpus." In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 2012-2017. Association for Computational Linguistics, 2019.

[3] Inaguma, Hirofumi, Kevin Duh, Tatsuya Kawahara, and Shinji Watanabe. "Multilingual end-to-end speech translation." In 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), pp. 570-577. IEEE, 2019.

[4] Li, Xian, Changhan Wang, Yun Tang, Chau Tran, Yuqing Tang, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. "Multilingual speech translation with efficient finetuning of pretrained models." arXiv preprint arXiv:2010.12829 (2020).

[5] Wang, Changhan, Anne Wu, Jiatao Gu, and Juan Pino. "CoVoST 2 and massively multilingual speech translation." In Interspeech, vol. 2021, pp. 2247-2251. 2021.

[6] Yun, Seung, Young-Jik Lee, and Sang-Hun Kim. "Multilingual speech-to-speech translation system for mobile consumer devices." IEEE Transactions on Consumer Electronics 60, no. 3 (2014): 508-516.

[7] Le, Hang, Juan Pino, Changhan Wang, Jiatao Gu, Didier Schwab, and Laurent Besacier. "Lightweight adapter tuning for multilingual speech translation." arXiv preprint arXiv:2106.01463 (2021).

[8] Gao, Yuqing, Bowen Zhou, Liang Gu, Ruhi Sarikaya, Hong-Kwang Kuo, A-VI Rosti, Mohamed Afify, and Weizhong Zhu. "IBM MASTOR: Multilingual automatic speech-to-speech translator." In 2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings, vol. 5, pp. V-V. IEEE, 2006.

[9] Cattoni, Roldano, Mattia Antonino Di Gangi, Luisa Bentivogli, Matteo Negri, and Marco Turchi. "MuST-C: A multilingual corpus for end-to-end speech translation." Computer speech language 66 (2021): 101155.

[10] Di Gangi, Mattia A., Matteo Negri, and Marco Turchi. "One-to-many multilingual end-to-end speech translation." In 2019 IEEE automatic speech recognition and understanding workshop (ASRU), pp. 585-592. IEEE, 2019.

[11] Barrault, Loïc, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, Mark Duppenthaler, Paul-Ambroise Duquenne et al. "Seamless: Multilingual Expressive and Streaming Speech Translation." arXiv preprint arXiv:2312.05187 (2023).

[12] Wahlster, Wolfgang, ed. Verbmobil: foundations of speech-to-speech translation. Springer Science Business Media, 2013.

[13] Hanumante, Vivek, Rubi Debnath, Disha Bhattacharjee, Deepti Tripathi, and Sahadev Roy. "English text to multilingual speech translator using android." International Journal of Inventive Engineering and Sciences 2, no. 5 (2014): 4-9.

[14] Iranzo-Sánchez, Javier, Joan Albert Silvestre-Cerda, Javier Jorge, Nahuel Roselló, Adria Giménez, Albert Sanchis, Jorge Civera, and Alfons Juan. "Europarl-st: A multilingual corpus for speech translation of parliamentary debates." In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 8229-8233. IEEE, 2020.

[15] Schultz, Tanja, and Katrin Kirchhoff, eds. Multilingual speech processing. Elsevier, 2006.

[16] Salesky, Elizabeth, Matthew Wiesner, Jacob Bremerman, Roldano Cattoni, Matteo Negri, Marco Turchi, Douglas W. Oard, and Matt Post. "The multilingual tedx corpus for speech recognition and translation." arXiv preprint arXiv:2102.01757 (2021).

[17] Dessloch, Florian, Thanh-Le Ha, Markus Müller, Jan Niehues, Thai-Son Nguyen, Ngoc-Quan Pham, Elizabeth Salesky et al. "KIT lecture translator: Multilingual speech translation with one-shot learning." In Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations, pp. 89-93. 2018.

[18] Matsuda, Shigeki, Teruaki Hayashi, Yutaka Ashikari, Yoshinori Shiga, Hidenori Kashioka, Keiji Yasuda, Hideo Okuma et al. "Development of the "VoiceTra" multi-lingual speech translation system." IEICE TRANSACTIONS on Information and Systems 100, no. 4 (2017): 621-632.

[19] Bano, Shahana, Pavuluri Jithendra, Gorsa Lakshmi Niharika, and Yalavarthi Sikhi. "Speech to text translation enabling multilingualism." In 2020 IEEE International Conference for Innovation in Technology (INOCON), pp. 1-4. IEEE, 2020.

[20] Barrault, Loïc, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar et al. "SeamlessM4T: Massively Multilingual Multimodal Machine Translation." arXiv preprint arXiv:2308.11596 (2023).

[21] Prasad, Rohit, Prem Natarajan, David Stallard, Shirin Saleem, Shankar Ananthakrishnan, Stavros Tsakalidis, Chia-lin Kao et al. "BBN TransTalk: Robust multilingual two-way speech-to-speech translation for mobile platforms." Computer Speech Language 27, no. 2 (2013): 475-491.

[22] Subramanya, Shashank, and Jan Niehues. "Multilingual simultaneous speech translation." arXiv preprint arXiv:2203.14835 (2022).

[23] Kim, Minsu, Jeongsoo Choi, Dahun Kim, and Yong Man Ro. "Textless Unit-to-Unit training for Many-to-Many Multilingual Speech-to-Speech Translation." IEEE/ACM Transactions on Audio, Speech, and Language Processing (2024).

[24] Dabre, Raj, Chenhui Chu, and Anoop Kunchukuttan. "A survey of multilingual neural machine translation." ACM Computing Surveys (CSUR) 53, no. 5 (2020): 1-38.

[25] Ansari, MD Faizullah, R. S. Shaji, T. J. SivaKarthick, S. Vivek, and A. Aravind. "Multilingual speech to speech translation system in bluetooth environment." In 2014 International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT), pp. 1055-1058. IEEE, 2014.

[26] Angelov, Krasimir, Björn Bringert, and Aarne Ranta. "Speech-enabled hybrid multilingual translation for mobile devices." In Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics, pp. 41-44. 2014.

[27] Liu, Danni, and Jan Niehues. "Recent Highlights in Multilingual and Multimodal Speech Translation." In Proceedings of the 21st International Conference on Spoken Language Translation (IWSLT 2024), pp. 235-253. 2024.

[28] Han, Seunghee, Gary Geunbae Lee, Hung Soon Kim, Sunhee Kim, and Minhwa Chung. "A Domain-Specific Multilingual Speech Translation Corpus via Simultaneous Interpretation." In ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1-5. IEEE, 2025.