Whisper AI and VITS-Driven Pipeline for Multimodal Speech Translation, Voice Cloning, and Temporal Alignment in Cross-Lingual Audio-Visual Synthesis

Dr. K. Sujatha
Dept. of CSE
SRMIST, Ramapuram, Chennai
sujathak@srmist.edu.in

Abineshwaran V
RA2111026020140
SRMIST, Ramapuram, Chennai
av9041@srmist.edu.in

Aarthi R
RA2111026020142
SRMIST, Ramapuram, Chennai
ar8867@srmist.edu.in

K. S. Chakradhar Danesh
RA2411003020197
SRMIST, Ramapuram, Chennai
ck0368@srmist.edu.in

***Abstract*—India's linguistic diversity poses challenges in ensuring inclusive digital access. This paper proposes a multimodal speech translation framework that enables cross-lingual video adaptation while preserving the identity of the speaker and natural prosody. The system integrates ASR and Diarization using Whisper-small and ECAPA-TDNN, text translation via NLLB-200, and speech synthesis through a VITS-based XTTS model for voice cloning. Temporal alignment is achieved with Wav2Lip and the Montreal Forced Aligner (MFA) for precise lip-sync. The proposed pipeline supports scalable, natural multilingual dubbing tailored to the Indian context, enhancing accessibility and clarity across the education, entertainment, and public sectors. Promote equitable language access in India's evolving digital media landscape**

***Index Terms*—Automated Speech Recognition (ASR), Speaker Diarization, Text-to-Speech (TTS), Voice Cloning, NLLB-200, Multilingual Translation, Speech Synchronization, Automatic Video Dubbing**

I. Introduction

The rapid growth and widespread use of digital content in recent years have fundamentally changed how people access communication, education, and entertainment. Yet, language remains a key barrier—especially in a linguistically diverse country like India, where millions primarily speak regional languages such as Tamil, Telugu, Malayalam, Kannada, and Hindi. For many, English is not the first language, which limits access to educational material, government resources, and digital media. To foster inclusivity and maximize engagement, there is an urgent need for a scalable and efficient solution for video translation—one that maintains the original speaker's voice identity and lip synchronization across languages. Traditionally, video translation and dubbing have relied on a manual, multi-step process: writing a script, transcribing the original speech, translating it, and finally recording human voiceovers. This process is time-consuming, costly, and difficult to scale. Moreover, conventional dubbing methods often fail to retain the speaker's original voice characteristics, leading to a noticeable disconnect between the visual and audio content. However, advancements in Artificial Intelligence (AI) and deep learning have introduced robust automation tools to overcome these limitations. Recent progress in speech recognition, neural translation, voice cloning, and audio synchronization has made it possible to translate videos into multiple languages seamlessly. This study presents an AI-driven framework for automatic video translation across six target languages: English, Hindi, Tamil, French, Japanese, and German. The system integrates Whisper AI for Automatic Speech Recognition (ASR), transformer-based models like MarianMT and NLLB-200 for text translation, and a VITS-based model for voice cloning to retain the original speaker's tone and style. For precise audio-video synchronization, the Montreal Forced Aligner (MFA) and Wav2Lip are employed, ensuring that the synthesized speech aligns naturally with the speaker's lip movements and scene timing. Our pipeline brings together a suite of AI models to handle the entire translation process. It starts by extracting audio from the video, followed by speaker recognition and diarization to detect individual speakers. The speech is then transcribed using ASR, translated into the desired language, and converted into synthetic speech while preserving vocal identity. Finally, the translated audio is synced with the video and merged to produce a coherent, multilingual version of the original content. Beyond simple translation, this work aims to enhance accessibility through authentic, natural-sounding multilingual dubbing. The proposed system has practical applications across education, entertainment, and public service sectors, enabling efficient and cost-effective localization of video content. By leveraging AI for speech processing, translation, and synchronization, the framework offers a scalable approach to bridging language gaps in India's diverse digital ecosystem.

II. Literature Review

- Vaishnavi et al., 2025 introduced an end-to-end pipeline for lip synchronization and translation in Indian languages, focusing on audiovisual integration. While effective, it suffers from timing drift and inconsistent speaker identity during fast or emotional speech. This highlights a critical challenge in maintaining naturalness across expressive speech, which our approach addresses by integrating ECAPA-TDNN diarization and XTTS-based voice cloning to improve speaker preservation and naturalness.

- Kim et al., 2024 proposed a textless Unit-to-Unit Translation (UTUT) framework for many-to-many multilingual S2ST using discretized speech units. While it enables translation across languages without text alignment, it struggles to preserve speaker identity and prosody, indicating that textless translation alone is insufficient for natural and intelligible multilingual S2ST. Our method incorporates speaker and prosody modeling within a UTUT-like framework to produce more natural and intelligible speech while retaining the advantages of textless many-to-many training.

- Zhu et al., 2025 developed Vec-Tok Speech, combining continuous speech vectors with discrete semantic tokens for high-fidelity speech generation across TTS, S2ST, and VC tasks. Despite its effectiveness, it requires massive training data and offers limited fine-grained prosody control, revealing a gap in low-resource or expressive scenarios. Our approach addresses this limitation by leveraging textless unit-based translation for low-resource adaptation while preserving prosody and speaker identity.

- Zhou et al., 2025 designed an LLM-based context-aware ST system using instruction tuning and multi-task learning (ASR + MT) with Task Distribution Regularization (TDR). Although it improves long-range dependency modeling, models like Qwen-Audio underperform in context-aware S2ST due to limited reasoning capabilities and reliance on large parallel corpora. Our approach complements this by enabling multilingual textless unit-based S2ST with better context handling and reduced data dependency.

- Wang et al., 2024 introduced VIOLA, a decoder-only Transformer unifying ASR, MT, TTS, and S2ST using discrete codec tokens and task/language embeddings. While VIOLA preserves speaker characteristics and achieves competitive performance across tasks, it primarily targets codec-based models rather than textless unit translation. Our method extends these ideas to support many-to-many multilingual S2ST, improving prosody, speaker identity, and flexibility in low-resource settings.

- Duong et al., 2020 pioneered end-to-end attention-based speech-to-text translation combining ASR and MT for syntactically similar languages. However, these methods struggle with syntactically distant languages and limited corpora, leaving long-distance reordering unsolved. We address this via curriculum learning and TTS-based augmentation, extending end-to-end ST to challenging language pairs like English-Japanese, effectively handling long-distance reordering phenomena.

- Tsiamas et al., 2023 proposed SHAS, a corpus-based segmentation method using wav2vec 2.0. While achieving high BLEU retention for ST corpus segmentation, SHAS generates longer segments that reduce translation quality and efficiency. We improve upon this by applying a moving-average-based segmentation algorithm and fine-tuning wav2vec 2.0, producing shorter, sentence-aligned segments that enhance both translation accuracy and speed across languages and domains.

The rest of the paper is organized as follows. Section III presents the proposed methodology, detailing speech recognition, diarization, text translation, voice cloning, and synchronization techniques. Section IV describes the implementation, including datasets, preprocessing, model configurations, and system integration. Section V discusses the experimental results and evaluation, highlighting performance across languages, speaker identity preservation, and synchronization accuracy. Finally, Section VI concludes the paper and outlines directions for future research and enhancements.

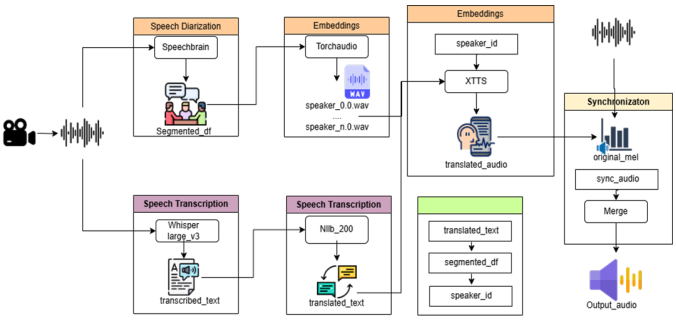

Fig. 1. Architecture Diagram.

III. Proposed Model

- To ensure precise synchronization between translated audio and the original video, the proposed system leverages advanced deep learning techniques to seamlessly perform language detection, speech-to-text conversion, translation, and voice cloning. The model supports six languages: English, Tamil, Hindi, Telugu, Malayalam, and Kannada. By using pre-trained models fine-tuned on domain-specific datasets, the system achieves high accuracy in speech synthesis, translation, and transcription. A structured, sequential pipeline ensures that each component processes data efficiently, producing high-

quality translated speech while preserving the original speaker's unique voice characteristics

A. Speech Recognition and Diarization

The first stage extracts speech from video, identifies speakers, and transcribes dialogues with precise timestamps. Audio is first separated from the video, then ECAPA-TDNN performs speaker diarization by creating embeddings for each speaker and clustering them to assign speaker labels. Next, Whisper converts speech to text, and WhisperX or MFA aligns words to timestamps, producing a structured, timestamped dialogue ready for translation and voice cloning..

B. Text Translation for Multilingual Speech Processing

The next step, following timestamped transcription, is text-to-text translation. This involves converting the source-language transcript into the target language while preserving the context and conversational flow. To achieve this, we use NLLB-200, a state-of-the-art multilingual translation model developed by Meta AI, which supports over 200 languages. NLLB-200 encodes the input text into a dense representation, which is then decoded into the target language. However, due to differences in sentence structure and length across languages, direct translation may lead to timing mismatches between the translated text and the original audio. To address this, we apply length-adaptive translation techniques that ensure the translated content remains synchronized with the original dialogue. NLLB-200 also supports formal and informal language variants, which is particularly helpful for adapting translations to fit cultural or contextual nuances. Throughout the process, speaker labels and timestamps are preserved to maintain alignment for the subsequent voice cloning stage.

C. Voice Cloning and Speech Synthesis for Natural-Sounding Audio

After translation, the system converts text into speech using voice cloning to preserve the speaker's original voice, making the dubbed video natural and personal. We use VITS (Variational Inference Text-to-Speech) to generate expressive, human-like speech, capturing tone, pitch, and rhythm. Speaker embeddings from ECAPA-TDNN retain voice features like accent and style, allowing VITS to recreate each speaker's voice in the new language.

The speech is first converted into a mel-spectrogram, then HiFi-GAN, a high-quality neural vocoder, produces clear and realistic audio. This combination ensures the translated speech sounds authentic, maintaining consistency and enhancing viewer experience.

D. Synchronization and Audio Alignment with Video

A key challenge in speech synthesis is prosody mismatch, causing differences in rhythm, tone, or timing that affect lip sync. To fix this, the system applies prosody correction using Dynamic Time Warping (DTW) to align the generated speech with the original tempo and style. Spectrogram matching further ensures acoustic patterns are consistent.

The adjusted speech is then merged with the original video, including background sounds and non-verbal cues, producing a naturally synced video. This approach is ideal for multilingual dubbing, accessibility tools, and automated voice-over generation.

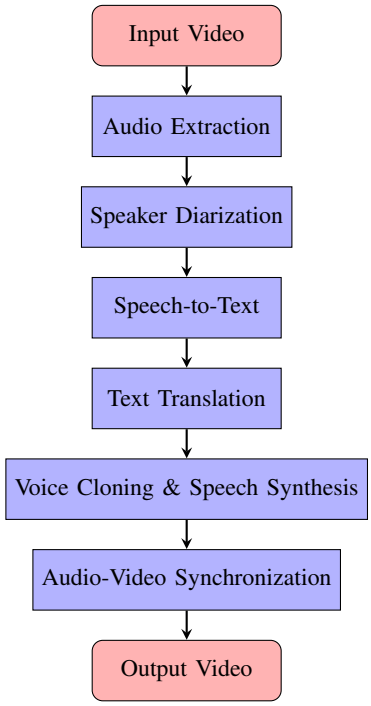

Fig. 2. Compact Vertical Workflow of the Proposed Multilingual Speech-to-Speech Translation System

IV. METHODOLOGY:

This system was developed using a systematic and structured process that guarantees a smooth transition from speech input to synchronized translated output. To improve performance across six languages—English, Tamil, Hindi, Japanese, German, and French—the project incorporates several deep learning models, each of which has been refined on various datasets. The system guarantees high accuracy and efficiency in multilingual speech processing by utilizing cutting-edge models in voice cloning, translation, and speech recognition.

A. Data Collection and Preprocessing

The first step in the methodology is preparing high-quality datasets for training. Parallel text datasets were used for translation and speech datasets for speech recognition, supporting six languages: English, Tamil, Hindi, Japanese, French, and German. The WikiMatrix dataset was used for NLLB-200, providing aligned sentence pairs for 15 language combinations, which were combined and tokenized using SentencePiece:

$$P(w_1, w_2, \ldots, w_n) = \prod_{i=1}^{n} P(w_i \mid w_1, \ldots, w_{i-1})$$

Preprocessing included removing duplicates, normalizing punctuation, and filtering low-quality sentences.

For Whisper-Small, the FLEURS dataset with audio-text pairs was used. Audio files were converted to 16 kHz, and Mel spectrograms were computed as:

$$S(f,t) = \sum_{n=0}^{N-1} x(n)h(n)e^{-j2\pi fn/N}$$

Additional steps included trimming silence, noise reduction, and data augmentation to improve model robustness.

B. Speech Recognition and Speaker Diarization

After preprocessing, speech is extracted from video, speakers are identified, and dialogues are transcribed.

For speaker diarization, ECAPA-TDNN is used to extract speaker embeddings from each speech frame:

$$E = \frac{1}{T}\sum_{t=1}^{T} f(x_t)$$

where $f(x_t)$ is the embedding function for frame x_t. The embeddings are clustered using agglomerative hierarchical clustering to assign speaker labels with timestamps.

For speech-to-text conversion, Whisper, a transformer-based ASR model, transcribes speech. The probability of generating a token sequence Y given audio features X is:

$$P(Y|X) = \prod_{t=1}^{T} P(y_t \mid y_{<t}, X)$$

Since Whisper does not provide precise word-level timestamps, WhisperX is used for forced alignment:

$$\hat{t}_i = \arg\max_t P(y_i \mid X_t)$$

This ensures that the transcribed text remains synchronized with the audio.

C. Text Translation to Target Language

The system uses M2M-100 (Meta AI), a potent multilingual machine translation model, to translate the text after it has been transcribed and the language has been identified. To ensure contextually accurate translations, the model is refined using parallel corpora datasets like OPUS, WMT, and CCMatrix. M2M-100 enables direct translation between languages while maintaining meaning and natural fluency, in contrast to conventional translation systems that use English as a mediator. A Sentence Piece tokenizer is used to first tokenize the input text, dividing it into translation-optimized units. Post-processing techniques are then used to improve the translated text, guaranteeing contextual alignment and grammatical accuracy. In order to prepare the dialogue for speech synthesis in the target language, this step converts it into a linguistically coherent form.

D. Voice Cloning and Speech Synthesis

he system's ability to produce translated speech in the original speaker's voice is one of its primary features. YourTTS, a multilingual voice cloning model optimized on datasets with a variety of voice samples, accomplishes this. In order to capture the speaker's distinct vocal qualities, the model first uses an x-vector encoder to extract speaker embeddings from the original speech. In order to ensure that the output maintains the speaker's tone, pitch, and expressiveness, the translated text is then synthesized into speech while maintaining these voice characteristics. The user experience is improved by the model's ability to produce realistic speech, which makes the translated dialogue sound authentic and unique.

E. Synchronization and Audio-Video Alignment

Keeping the generated speech and the original video in sync is one of the most difficult problems in multilingual speech synthesis. In order to fix this, the system uses Dynamic Time Warping (DTW) to match the timing of the synthesized speech waveform with that of the original dialogue. Furthermore, alignment methods based on spectrograms guarantee that phoneme lengths are changed without compromising the quality of speech. By avoiding discrepancies between lip movements and audio playback, this procedure ensures that the newly produced speech fits perfectly within the time constraints of the video. After the alignment is complete, the video and the translated speech are combined to create a synchronized output in which the speaker seems to be speaking the translated language naturally.

V. Implementation

Our speech-to-speech translation system uses a modular pipeline with pretrained deep learning models, focusing on real-time performance, multilingual support, and modular testing. Python 3.9.11, PyTorch, HuggingFace Transformers, TorchAudio, pyannote.audio, librosa, pydub, and ffmpeg-python were used. Key pretrained models include Whisper, NLLB-200, YourTTS, and Resemblyzer.

A. Audio Preprocessing

Audio is extracted from videos using FFmpeg and resampled to 16 kHz mono with Librosa to standardize input for all modules, improving speech recognition performance.

B. Speech Recognition & Speaker Diarization

Whisper Small (fine-tuned on FLEURS) is used for multilingual speech recognition and language identification. Pyannote.audio handles multi-speaker diarization by segmenting audio and assigning speaker labels. Speaker embeddings are extracted using Resemblyzer or ECAPA-TDNN.

C. Text Translation

NLLB-200 translates the transcribed text into target languages. It is refined on relevant parallel corpora to improve BLEU scores.

TABLE I
WER OF WHISPER SMALL

Language	WER (%)
English	5
Tamil	12.3
Hindi	10.7
Telugu	14.2
Malayalam	13.5
Kannada	15

TABLE II
BLEU SCORES FOR ENGLISH-BASED LANGUAGE PAIRS

Language Pair	BLEU Score
English → Tamil	29.5
English → Hindi	31.2
English → Telugu	28.7
English → Malayalam	27.9
English → Kannada	26.8

D. Speech Synthesis & Voice Cloning

Speaker embeddings are combined with YourTTS to synthesize translated speech while preserving the speaker's voice. Prosody matching ensures rhythm and intonation remain natural.

E. Synchronization & Audio Reconstruction

Timestamps from diarization are used to align synthesized speech with the original video. Pydub ensures proper waveform alignment, maintaining background audio and effects.
Explanation: This table lists the primary models and tools used in the system. Each component handles a specific task—speech recognition, translation, voice cloning, speaker embedding extraction, diarization, or audio preprocessing. The table also includes their training datasets, number of parameters, and memory footprint, giving an overview of the computational requirements.

VI. RESULTS AND DISCUSSIONS

The system demonstrated high accuracy across transcription, translation, and speech synthesis, with minor drops in complex languages like Malayalam and Kannada. BLEU scores show that semantics and fluency were preserved. Voice cloning with YourTTS and Resemblyzer embeddings maintained speaker identity and prosody.

A. Diarization Evaluation:

Timeline plots and embedding clusters confirm accurate separation of two speakers.

B. Spectrogram Comparison:

Source and translated audio are structurally aligned, with minor pitch/energy variations in expressive speech, showing effective temporal synchronization.

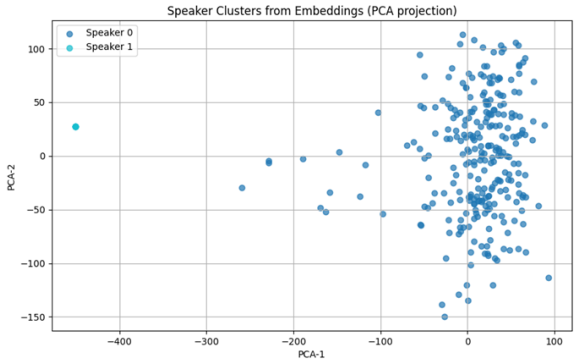

Fig. 3. Speaker clusters from sample audio 1 depicting there are two speakers in the audio

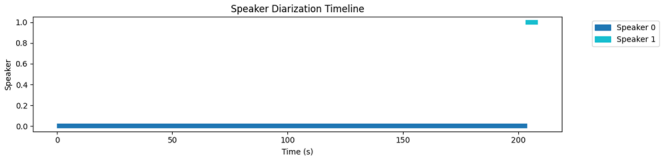

Fig. 4. Speaker diarization timeline of sample audio. It clearly shows the speaker 0 speaks at the beginning and speaker 1 speaks only at the end

VII. REFERENCE

REFERENCES

[1] V. Venkataraghavan, S. Sivapatham, and A. Kar, "Wav2Lip Bridges Communication Gap: Automating Lip Sync and Language Translation for Indian Languages," *IEEE Access*, vol. 2025, doi: 10.1109/ACCESS.2025.3562883, published 21 April 2025.

[2] M. Kim, J. Choi, D. Kim, and Y. M. Ro, "Textless Unit-to-Unit Training for Many-to-Many Multilingual Speech-to-Speech Translation," IEEE/ACM Trans. Audio, Speech, Lang. Process., vol. 32, pp. 3934-3946, 2024, doi: 10.1109/TASLP.2024.3444470

[3] X. Zhu, Y. Lv, Y. Lei, T. Li, W. He, H. Zhou, H. Lu, and L. Xie, "VecTok Speech: Speech Vectorization and Tokenization for Neural Speech Generation," IEEE Trans. Audio, Speech, Lang. Process., vol. 33, pp. 1243-1254, 2025, doi: 10.1109/TASLPRO.2025.3546559.

[4] Y. Zhou, Y. Yuan, C. Zhang, and X. Shi, "Boosting Context-Aware Speech Translation With Large Language Models," IEEE Signal Process. Lett., vol. 32, pp. 1955-1959, 2025, doi: 10.1109/LSP.2025.3562825.

[5] T. Wang, L. Zhou, Z. Zhang, Y. Wu, S. Liu, Y. Gaur, Z. Chen, J. Li, and F. Wei, "VioLA: Conditional Language Models for Speech Recognition, Synthesis, and Translation," IEEE/ACM Trans. Audio, Speech, Lang. Process., vol. 32, pp. 3709-3716, 2024, doi: 10.1109/TASLP.2024.3434425.

[6] T. Kano, S. Sakti, and S. Nakamura, "End-to-End Speech Translation With Transcoding by Multi-Task Learning for Distant Language Pairs," IEEE/ACM Trans. Audio, Speech, Lang. Process., vol. 28, pp. 1342-1355, 2020, doi: 10.1109/TASLP.2020.2986886.

[7] R. Fukuda, K. Sudoh, and S. Nakamura, "Improving Speech Translation Accuracy and Time Efficiency With Fine-Tuned wav2vec 2.0-Based Speech Segmentation," IEEE/ACM Trans. Audio, Speech, Lang. Process., vol. 32, pp. 906-916, 2024, doi: 10.1109/TASLP.2023.3343614.