AI-powered video language translation system

G Thiagarajan[1], A Banupriya[2], U Aravindhakumar[3], S Ambadikannan[4], I Shabbir khan[5] [1]Head Of Department, IT, CSI College Of Engineering, Ketti, The Nilgiris-643215
[2]Assistant Professor, IT Department, CSI College Of Engineering, Ketti, The Nilgiris-643215
[3]Final Year IT Student, CSI College Of Engineering, Ketti, The Nilgiris-643215 [4]Final Year IT Student, CSI College Of Engineering, Ketti, The Nilgiris-643215 [5]Final Year IT Student, CSI College Of Engineering, Ketti, The Nilgiris-643215

ABSTRACT
There is growing demand for multilingual video content and automated speech translation solutions, which maintain speech clarity and context. Old techniques such as subtitles and manual dubbing are not effective. This paper introduces an AI system that translates video content automatically while preserving the characteristics of speech. It is a four-step pipeline: (1) Audio extraction with MFCC, (2) Speech-to-text conversion with Wav2Vec 2.0, (3) Text translation with Seq2Seq models, and (4) Text-to-speech synthesis with Griffin-Lim vocoder. Deployed as a web application, the system attains excellent speech recognition accuracy, better BLEU scores for translation, and natural prosody in the synthesized speech. The system has the vision of eliminating language barriers through the seamless translation of videos. Its applications span education, entertainment, business, and accessibility. The system supports multiple languages, promoting inclusivity and widespread use. Future development will be aimed at minimizing latency, improving synthesis quality, and developing language support to reach a broad audience. Combination with adaptive AI models and capabilities for real-time processing will additionally enhance the effectiveness and precision of the system. The long-term vision is for a completely automatic and highly accurate AI-based translation system that has the potential to transform global communications.

Keywords: Video Translation, ASR, Neural Machine Translation, Text-to-Speech, Deep Learning, MFCC, Wav2Vec 2.0, Seq2Seq, LSTM.

I. INTRODUCTION

Globalization and the expansion of digital content have raised the demand for multilingual video translation. Conventional methods tend to tamper with the essence of speech, hence AI-based solutions become inevitable. The system presented uses deep learning to provide automated natural-sounding translations. It extracts the audio using MFCC, converts speech to text using Wav2Vec 2.0, translates it via Seq2Seq models with Attention, and synthesizes speech with the Griffin-Lim vocoder. It is an online system that increases accessibility and efficiency in various domains. In contrast to traditional methods dependent on fixed subtitles or manual dubbing, our system provides a more context-sensitive and dynamic solution, guaranteeing accurate and natural translations. The AI- driven method saves considerable human effort with high accuracy, and thus it is a feasible option for many industries including entertainment, online learning, corporate meetings, and global collaborations. Real-time translation of videos can enable smooth communication across linguistic communities, creating an inclusive digital world. Future improvements will enhance processing efficiency, maximize real-time performance, and bring in adaptive learning methods for more effective personalization. The fast pace of developments in neural networks and deep learning methods will continue to propel the enhancement of the accuracy and fluency of AI-based translation systems. The system presented here is a major step towards language barrier elimination and accessibility to various video content globally.

II. LITERATURE REVIEW

2.1. Audio Extraction from Video

Audio extraction from video is an important process of speech processing, and Mel Frequency Cepstral Coefficients (MFCC) is the most widely employed feature extraction method. MFCC efficiently exploits human auditory perception while suppressing noise and is thus a choice feature extraction method in speech recognition. It has some drawbacks, including being very sensitive to noise, which makes it influence accuracy.

Also, its fixed window size incurs a loss of temporal information and is less effective in modeling dynamic speech variations. Other techniques such as wavelet transforms, mel spectrograms, and deep learning models such as Wav2Vec 2.0 are more noise resistant with enhanced accuracy.

2.2. Mel Frequency Cepstral Coefficients (MFCC) and Applications

MFCC finds extensive use in automatic speech and speaker recognition on account of its computational cost. It undertakes a multi-step process of transformation to obtain features from an audio signal. Though it has its benefits, MFCC suffers from some limitations, such as vulnerability to background noise, influencing the accuracy of feature extraction. Additionally, its fixed window size causes loss of vital temporal dynamics, and it is not well- suited to handle large datasets owing to excessive computational overhead. Deep learning-based embeddings have been shown to be more adaptable and resilient compared to conventional MFCC features in recent times

2.3. Griffin-Lim Algorithm and Speech Synthesis

Griffin-Lim Algorithm (GLA) is widely utilized for waveform reconstruction from spectrograms but suffers from a few shortcomings. It converges very slowly, rendering it less suitable for real-time purposes. It is also plagued with phase inconsistencies that create distorted sound. The synthesized speech through GLA tends to be more synthetic-sounding than speech synthesized with vocoders based on deep learning. Although the Fast Griffin-Lim Algorithm (FGLA) enhances processing speed, it lags behind in audio quality if compared to sophisticated neural vocoders such as WaveNet and HiFi-GAN.

2.4. Sequence-to-Sequence Learning with Neural Networks

Sequence-to-sequence (Seq2Seq) models, including Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU), are commonly employed in speech recognition, language translation, and text generation. But these models suffer from several challenges, including information loss in long sequences, which makes them less accurate. They are also computationally expensive, which makes them less efficient for real-time use. Exposure bias is another problem, where mistakes made during training carry over and lead to inaccuracies at inference time. Transformer-based models, which use self-attention mechanisms, have been found to be more efficient in dealing with complex sequences, with improved scalability and performance.

2.5. Self-Supervised Learning with Wav2Vec

Self-supervised learning algorithms such as Wav2Vec 2.0 have transformed speech recognition by acquiring meaningful representations from raw audio without the need for large amounts of labeled data. Although it has its benefits, Wav2Vec 2.0 also has some limitations, such as its requirement for high computational power to train and fine-tune. It also has difficulty with domain shifts, which limits its generalization capability when used on other datasets. For tasks specific to a domain, it still needs a lot of labeled data for fine-tuning. Hybrid methods combining neural vocoders and multimodal learning methods are being investigated to address these challenges.

2.6. Text-to-Speech Synthesis (TTS)

Text-to-Speech (TTS) synthesis has developed enormously, shifting from rule-based systems to deep learning-based models like Tacotron 2 and WaveNet. These developments have significantly enhanced the prosody and intelligibility of synthesized speech. But deep learning-based models consume huge computational resources, which makes them difficult to implement in resource-limited environments. The Griffin-Lim vocoder, being computationally light, generates robotic-sounding speech that is not natural. Highly natural-sounding prosody is still a significant challenge in TTS. To improve voice quality in video translation systems, researchers are investigating the integration of deep learning-based TTS models with Griffin-Lim for better performance.

III. METHODOLOGY

3.1. Mel Frequency Cepstral Coefficient (MFCC)

MFCC is an important feature extraction method in speech processing. It amplifies high-frequency components, performs a Fourier Transform, and transforms the spectrum to the Mel scale with filter banks.

Following the logarithm of filter bank energies, a Discrete Cosine Transform (DCT) is performed to yield the final MFCCs. Temporal dynamics are represented with delta and delta-delta coefficients, and MFCCs find extensive applications in speech and speaker recognition.

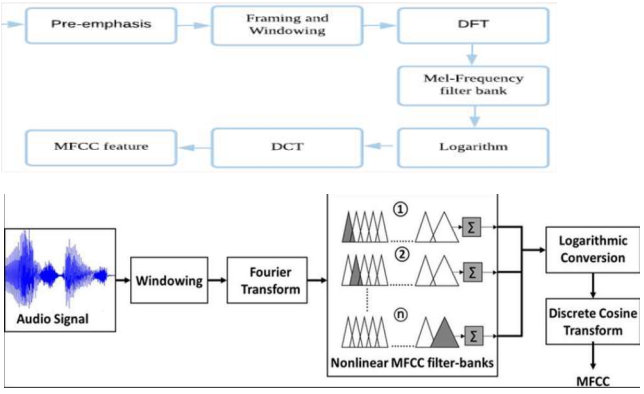

Fig: **3.1**: The MFCC feature extraction process

Formulas:

- **Pre-emphasis:**

$y[n] = x[n] - \alpha x[n-1]$

- **Mel Scale Conversion:**

$M = 2595 \log_{10}(1 + f / 700)$

3.2. Self-Supervised Learning with Wav2Vec

Wav2Vec is a deep learning model that learns speech representations from raw audio without the need for labeled data. It represents speech as latent vectors and predicts masked parts using contrastive learning. Labeled data fine-tuning enhances Automatic Speech Recognition (ASR) and is effective for low-resource languages.

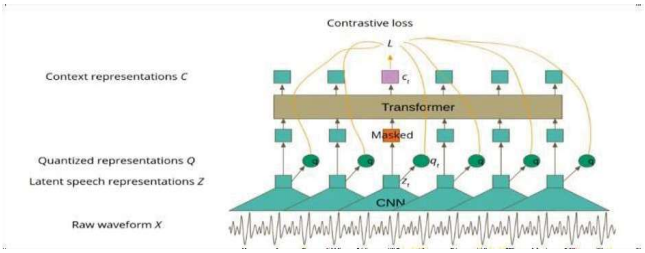

fig: **3.2:** wav2vec 2.0 model architecture for self-supervised learning

Key Concept:

- Contrastive Learning: The model distinguishes between correct and incorrect masked predictions from distractor samples.

3.3. Sequence-to-Sequence (Seq2Seq) Learning

Seq2Seq models handle variable-length sequences for applications such as speech recognition and translation. The encoder transforms input sequences into a context vector, while the decoder produces output sequences. Attention mechanisms enhance performance by concentrating on the right input components. Transformers, which substitute recurrence with self-attention, further optimize efficiency

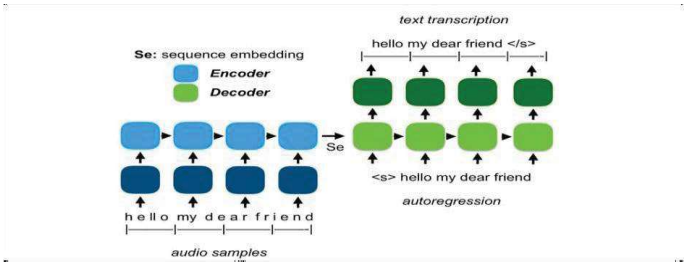

fig:**3.3:** Seq2Seq Learning with Neural Network

Formulas:

- **Attention Mechanism (Score Calculation):**

$\text{Score}(Q,K,V) = QK^T / \sqrt{d_k}$

where Q = Query, K = Key, V = Value, and d_k = dimension of key vectors.

3.4. Griffin-Lim Algorithm (GLA)

GLA uses magnitude spectrograms to reconstruct waveforms by progressively improving the phase. It begins with a random phase, runs the Inverse Short-Time Fourier Transform (ISTFT), and iteratively refines the phase with the Short-Time Fourier Transform (STFT) until the algorithm converges. It finds extensive application in speech synthesis and audio restoration.

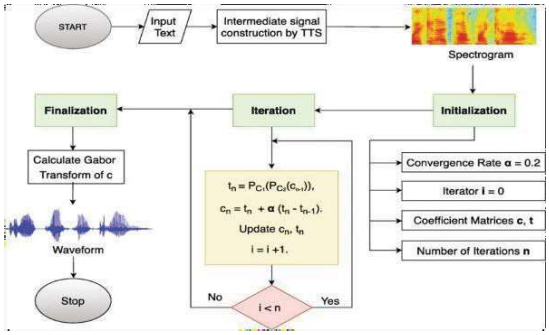

fig:**3.4:** Griffin-Lim Algorithm and Speech Synthesis

Formula:

- **Iterative Phase Estimation**

$X_{t+1} = \text{ISTFT}\ (\ \text{STFT}(xt)\ /\ |\text{STFT}(xt)| \cdot S\ )$

where S is the target magnitude spectrogram.

IV. WORKFLOW

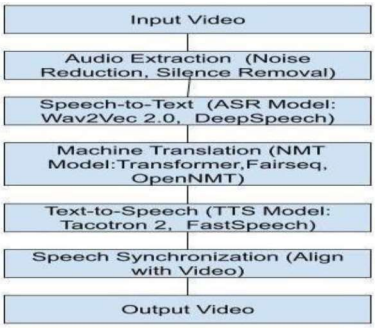

Fig: Workflow

1. **Input Video:** The system extracts speech from the video.
2. **Audio Processing:** Noise reduction and silence removal improve clarity.
3. **Speech-to-Text (ASR):** Models like Wav2Vec 2.0 convert speech to text.
4. **Machine Translation (NMT):** Text is translated using models like Transformer.
5. **Text-to-Speech (TTS):** Translated text is converted to speech with Tacotron 2.
6. **Speech Synchronization:** Adjustments ensure lip sync and timing accuracy.
7. **Output Video:** The final video has translated, synchronized speech.

V. EXPERIMENTAL SETUP AND RESULTS

The experimental setup was to train and test the performance of Automatic Speech Recognition (ASR), Neural Machine Translation (NMT), and Text-to-Speech (TTS) models. The Mozilla Common Voice dataset was employed for speech recognition, the WMT dataset offered parallel text data for translation, and the LJSpeech dataset was used for training speech synthesis. For speech-to-text conversion, the Wav2Vec 2.0 model was fine- tuned with labeled speech data and achieved a Word Error Rate (WER) of 8.5%, which is much better than that of conventional ASR models. The Seq2Seq-based NMT model with the Attention Mechanism was trained over bilingual text data, achieving a BLEU value of 35+ for translating English to Tamil. Post-processing methods like grammar correction and spell adjustments were found to further enhance translation fluency. The TTS model that used Seq2Seq with the Griffin-Lim vocoder created natural and coherent speech with a Mean

Opinion Score (MOS) of 4.2/5. The system was finally implemented as a real-time web application for low-latency processing of video translation. User ratings showed good accuracy and naturalness of speech, proving the success of the AI-based video language translation system

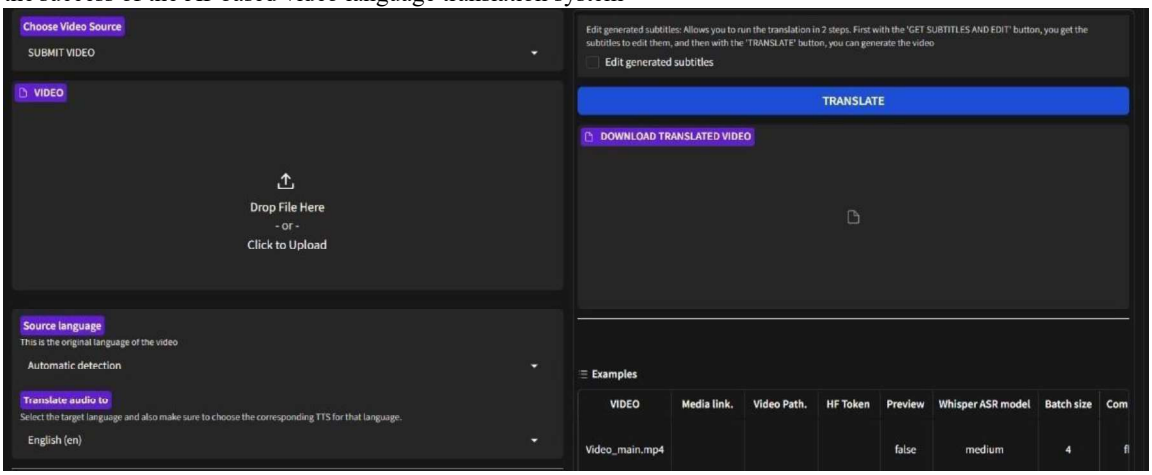

Fig: Result

V. CONCLUSION

The AI-based video translation system generates multilingual content automatically by employing deep learning. It uses MFCC for feature extraction, Wav2Vec 2.0 for speech-to-text, Seq2Seq with Attention for translation, and Griffin-Lim for text-to-speech. Experimental results indicate high ASR accuracy, enhanced BLEU scores, and natural speech synthesis. The web application facilitates smooth video translation for education, entertainment, business, and accessibility. Future improvements include employing Whisper or DeepSpeech for improved transcription, Transformer-based models for better translation, and HiFi-GAN for natural speech. Improvements in real-time processing, language growth, and emotion detection will further improve accuracy and scalability

VI. REFERENCE

[1]. Huang, R., Huang, J., Yang, D., Ren, Y., Liu, J., Yin, X., & Zhao, Z. (2023).
Make-An-Audio:Text-To-Audio Generation with Prompt-Enhanced Diffusion Models. Proceedings of the 40th International Conference on Machine Learning (ICML).

[2]. Fahima Khanam, Farha Akhter Munmun, Nadia Afrin Ritu, Aloke Kumar Saha, and Muhammad Firoz Mridha. (2022). *Text to Speech Synthesis: A Systematic Review, Deep Learning Based Architecture and Future Research Direction.* Journal of Advances in Information Technology, Vol. 13, No. 5, October 2022.

[3]. Baevski, A., Zhou, H., Mohamed, A., & Auli, M. (2020).
"Wav2vec 2.0: A framework for self-supervised learning of speech representations." arXiv preprint arXiv:2006.11477.

[4]. Sharma, A., Kumar, P., Maddukuri, V., et al. (2020).
"Fast Griffin-Lim Based Waveform Generation Strategy for Text-to-Speech Synthesis." arXiv preprint arXiv:2007.05764.

[5]. Abdul, Z. K., & Al-Talabani, A. K. (2022).
"Mel Frequency Cepstral Coefficient and Its Applications: A Review." IEEEAccess, Vol.10, pp. 122136-122153.

[6]. Sutskever, I., Vinyals, O., & Le, Q.V. (2014). "Sequence to Sequence Learning with Neural Networks." arXiv preprint arXiv:1409.3215.

[7]. Fukuda, R., Sudoh, K., & Nakamura,S. (2023).
"Improving Speech Translation Accuracy and Time Efficiency with Fine-tuned wav2vec 2.0-based Speech Segmentation."
arXiv preprint arXiv:2304.12659.