Speech-to-Speech Translation with Lip-Synchronization

Soujanya B K
Assistant Professor
School of CSE, REVA University, Bangalore, India

Abhishek U Gaonkar
School of CSE, REVA University, Bangalore, India

Chandan N
School of CSE, REVA University, Bangalore, India
&
Sumeet Chavan
School of CSE, REVA University, Bangalore, India

Abstract— This innovative project introduces a comprehensive system for multilingual video dubbing, with a primary focus on enhancing accessibility for individuals with limited literacy seeking content in their native language. The workflow commences with a Speech-to-Text module, meticulously transcribing English speech into written text, serving as a foundational step to bridge the auditory and textual components. The Neural Machine Translation (NMT) module then takes precedence, utilizing advanced neural network architectures to translate transcribed English text into the desired target language. Going beyond traditional translation, the system aims to capture linguistic nuances and cultural sensitivities, ensuring anaccuratere presentation of the original content. Advancing through the workflow, the Text-to-Speech modulere fines the translated content, optimizing it for synthesis into the desired language to provide a natural and expressive spoken output, thereby enhancing overall accessibility. Differentiating our project is the incorporation of the Lip GAN Visual Module, leveraging advanced techniques like Generative Adversarial Networks (GANs)to generate lifelike lip movements synchronized seamlessly with the translated speech. This visual enhancement introduces a unique and immersive aspect to the viewing experience, catering to a diverse audience, particularly those with limited literacy.

Keywords: Neural machine translation, Synchronisation of lip, dubbing, Speech translation.

I. Introduction

This project pioneers an advanced system for multilingual video dubbing, prioritizing accessibility for individuals with limited literacy in their native language. Beginning with a Speech-to-Text module, the workflow transcribes English speech into written

text, forming a foundational bridge between auditory and textual elements. The Neural Machine Translation (NMT) module then employs sophisticated neural networks to translate the text into the target language, capturing linguistic nuances and cultural sensitivities for an accurate representation.

The Text-to-Speech module refines translated content for expressive synthesis, enhancing overall accessibility. The project distinguishes itself with the Lip GAN Visual Module, leveraging advanced techniques like GANs to generate realistic lip movements synchronized with translated speech, offering an immersive viewing experience. Unlike traditional subtitle-based platforms, this system introduces real-time lip synchronization, benefiting individuals with limited literacy by integrating visual cues alongside translated speech. The holistic approach addresses both spoken and visual elements, transcending language barriers, and subtitles to ensure comprehension for diverse audiences.

Language plays a crucial role in daily human life, and its importance cannot be underestimated. In the modern world, which is rapidly globalized, effective communication without any limitations of language plays a vital role. Language serves as the foundation for human interaction, fostering communication on ideas, culture, and knowledge. Unfortunately, the high degree of variation that exists among languages globally poses a substantial barrier to seamless communication and understanding across people with different language backgrounds. In response to this challenge, speech-to-speech translation (S2ST) systems have emerged as an innovative technology that has the potential to fill this gap by allowing translation in real-time during a conversation.

There are several new systems called speech-to-speech translation systems, which means that they can take your spoken words in one language and translate them into another language. In this way, two people who do not speak the same language can communicate without obstacles. With the help of artificial intelligence, the S2ST systems create a vision of how we will be able to overcome these barriers in the future.

The most crucial challenge in S2ST system development is the achievement of elevated levels of precision and fluency within translations while maintaining details and semantics of speech. It goes far beyond simple translating, as it is crucial for successful communication, that the translated speech should sound natural and be easy to understand. This is where the idea of lip synchronization comes into action. Lip synchronization is when the lip movements of an animated character align with the sounds of speech. When used in S2ST systems, lip synchronization increases realism and comprehension because it perfectly matches translated words with synchronized lip movements. This visual cue greatly improves user interaction, making translated speech more appealing and easier to follow, especially when audio by itself might not be enough, like in loud places or for people who are hard of hearing.

II. Related Work

Research conducted by Jia Y. et al. suggests that utilizing pre-trained models for converting text to speech (TTS) or machine translation (MT) can be more efficient for

generating speech-to-translation pairs from weakly supervised data, compared to employing multitask learning methods. The study also examines the performance disparities between end-to-end speech-to-text (ST) models and standard cascaded models. The results indicate that in experiments focused on a substantial English speech-to-Spanish text translation task, the end-to-end ST model outperforms multitask learning, as assessed by metrics like Word Error Rates and BLEU scores [1].

Eliya Nachman et. al. worked towards creating a Text-to-Speech (TTS) network that can translate spoken language from one source language to multiple target languages. Rather than relying on recordings of the same speaker speaking various languages, their approach involves training with a single speaker speaking in one language. This is accomplished by extracting speaker embeddings from the audio samples and sharing them across the embedding space for all other languages. The proposed "polyglot" architecture employs a shared decoder to convert text into speech. This method offers increased practicality due to reduced training needs, as demonstrated by the system's performance [2].

Sercan Arik et.al. present a novel approach to creating a neural Text-to-Speech (TTS) system capable of generating multiple voices using a single model. Their technique builds upon Deep Voice 2, an advanced iteration of Taco Tron and Deep Voice 1, resulting in notable improvements in audio quality. The Deep Voice 2 model integrates segments of speaker embeddings from each independently trained model. Evaluations, conducted via Mean Opinion Score (MOS), indicate that the proposed system can achieve high-quality synthesis while requiring less data per speaker. Additionally, the study delves into the concept of voice replication, wherein a speaker's voice is replicated to read unseen text [3].

The study conducted by Sercan O. Arık et.al. delves into the development of a voice replication system, with the aim of creating a neural network capable of mimicking an individual's voice using only a limited set of voice samples. This objective is achieved through the utilization of voice adaptation and speaker encoding techniques. Both models demonstrate the ability to accurately reproduce a cloned voice even with a small dataset. However, challenges emerge when the system encounters text outside of its usual domain, such as technical or scientific language not included in the training data. This may pose difficulties in generating natural-sounding speech, potentially limiting the system's effectiveness for specific applications requiring high-quality speech output [4].

The study conducted by Xinyong Zhou et.al. presents a novel approach to cross-lingual voice cloning. The model takes a brief English audio sample as input. Utilizing Tacotron2 as the foundational model, which includes a latent prosody model generating bottleneck features and an acoustic model producing acoustic features, the system then feeds these features into a neural vocoder to create a cloned voice in Mandarin. This process preserves the naturalness and likeness of the original speaker's voice. Lip synchronization holds significant sway, as any errors in this process can

detract from the viewer's focus while watching a video. Therefore, ensuring accurate lip synchronization is paramount [5].

The study by Abhishek Jha et.al. highlights the importance of lip synchronization in video dubbing. The procedure of creating dubbed audio with synchronized lip movements entails transforming audio data into lip landmarks, superimposing them onto the mouth, and modifying facial features accordingly. Nonetheless, the precision and excellence of phonetic transcriptions in both the original and target languages are crucial for the effectiveness of this approach [6].
By using speech-to-speech translation, Marcello Federico et al. sought to create an automatic dubbing system that would automatically translate English-language video into Italian-language video. To keep the originality and authenticity of the video intact, the speed, acoustics, and pitch were all kept from the original production. The phrasal level lip synchronization produced by this model was good, but the prosodic alignment was negatively impacting the dubbed language's fluency. The quality of the dubbed audio could be affected by background noise in the voice audio while using this method. Additionally, it might not be able to capture speaker-specific details like accents or speech impairments, which might affect how accurate and natural the dubbed audio sounds [7].

Yang, Yi, et al. developed a dubbing and audiovisual translation system in their study. Prior to translation, the target speaker was fine-tuned into the generic models using the video to be translated as input. Through further data-processing processes and speaker-specific fine-tuning, viewers were able to watch in the target language seamlessly and uninterrupted [8].

The study by Surafel Melaku Lakew focuses on the neural machine translation characteristic in which, after translation, output text that is received has a same number of words as the input. This is accomplished by means of a transformer design, wherein the output text length is first biased into a length-ratio target-source class. The second is to use the input text's length to improve the transformer positional embedding. Ensuring that the output text's length does not surpass the input text's length was the paper's conclusion. BLEU scores were used to evaluate the model [9].

KR, P., Mukhopadhyay et.al. are dedicated to automatically translating videos featuring speakers in one language into another while ensuring natural lip synchronization. They've developed an automated process that includes a speech-to-speech system and an advanced visual component called "Lip GAN." This component utilizes translated audio from the original video to generate lifelike talking faces. The positive user experience ratings obtained by combining lip-sync with manual dubbing highlight the effectiveness of the Lip GAN model, indicating significant potential for improvement in translation models to enhance consistency and overall user satisfaction. While face detection and tracking in these systems are primarily powered by computer vision techniques, Ardila et al. focus on optimizing real-time applications' sound quality through a Vocoder-based speech synthesis system [10].

III. Methodology

Lip synchronization using AIML integrates several modules to create a sophisticated system for audio-visual alignment. The process begins with the conversion of spoken words to text through the 'speech-to-text' module. The resulting text is then processed and analyzed using AIML, leveraging natural language processing capabilities for understanding and interpretation.22

The 'text-to-text' module comes into play, ensuring that the textual representation is refined and enhanced, possibly incorporating additional contextual information. Subsequently, the system utilizes the 'text-to-speech' module to generate synthetic speech based on the refined text. This synthetic speech is synchronized with lip movements using the 'lipGan' module, which likely involves advanced algorithms or models for lip animation and visual synchronization.

overall system aims to create realistic and accurate lip synchronization by leveraging AIML for linguistic understanding and generation, as well as specialized modules for processing audio and visual elements. This integration enhances the immersive experience in multimedia applications, such as virtual environments, animations, or video conferencing."

A. Speech-To-Text Module:

In natural language processing systems, a speech-to-text module is an essential part that translates spoken words into written text. Its main role is to convert spoken words and phrases into a written format so that spoken content can be more easily analysed, stored, and worked with. Voice-activated assistants, accessibility features, and transcribing services are just a few of the many uses for this technology. Advanced algorithms and machine learning models that can precisely recognize and understand spoken language are essential to the operation of a speech-to-text module. These systems typically involve several key steps:

1. **Audio Input Processing**: The module receives audio input in the form of spoken words, whether from recorded audio files or real-time speech through a microphone.

2. **Feature Extraction**: The raw audio signal is transformed into a set of features that highlight important characteristics, such as pitch, frequency, and duration.

3. **Acoustic Modelling**: Machine learning models, often based on deep neural networks, analyse these features to recognize patterns and map them to linguistic units, such as phonemes and words.

4. **Language Modeling**: The module incorporates language models to enhance accuracy by considering the likelihood of specific word sequences and context in the given language.

.

5. **Decoding**: The recognized linguistic units are then decoded into written text, forming the final transcribed output.

B. Neural Machine Translation Module:

A Neural Machine Translation (NMT) module is a powerful component within natural language processing systems that leverages neural networks to perform automated translation between languages. Unlike traditional machine translation methods, NMT uses deep learning techniques to improve translation quality and fluency. The functionality of an NMT module involves the following key aspects:

The neural network framework employed in NMT follows an encoder-decoder structure. In this setup, the decoder generates the corresponding translated text in the target language, while the encoder handles the input text in the source language, producing a contextualrepresentation.

Word Endings: Word embeddings are used by NMT to represent words as continuous vector spaces that capture contextual information and semantic links. As a result, the translations produced by the model are more accurate.

Attention Mechanism: As the model generates each word in the target translation, the attention mechanism enables it to selectively concentrate on various segments of the source text. This enhancement bolsters the model's ability to manage and maintain coherence, particularly in lengthy sentences.

Training with Parallel Data: In order for NMT models to grasp translation patterns effectively, they require access to extensive parallel datasets comprising pairs of sentences in both the source and target languages. Throughout the training process, the model adjusts its parameters to minimize the disparity between the anticipated and actual translations.

End-to-End Learning: Unlike traditional methods that depend on predefined language rules, NMT learns translation directly from input to output. This holistic learning strategy enables the model to grasp intricate language relationships and nuances effectively.

NMT module has significantly improved the quality of automated translation, producing more fluent and contextually accurate results. It has become the dominant approach in machine translation, demonstrating its effectiveness in various applications such as online language translation services, cross-language communication tools, and global content localization.

C. Text-To-Speech Module:

Synthetic speech creation is made possible by the use of Text-to-Speech (TTS) modules, which translate written text into spoken words. This module is essential to many applications, including voice-activated gadgets, virtual assistants, and accessibility features. Text-to-Speech modules are essential for improving the user experience in voice-activated interfaces, navigation systems, and e-learning programs, as well as for providing information accessibility to people with visual impairments. The naturalness and expressiveness of artificial voices generated by TTS systems have significantly improved because of developments in deep learning and natural language processing.

D. Lipgan Visual Module:

visual module associated with lip synchronization; it could potentially involve the integration of advanced technologies like Generative Adversarial Networks (GANs) to enhance the visual fidelity of lip movements in synchronization with spoken words.

The term "LIPGAN" implies a system that employs GANs for generating realistic lip movements in the context of lip synchronization.

In a lip synchronization context, a "LIPGAN VISUAL MODULE" might incorporate GAN-based techniques to produce visually convincing lip movements that align seamlessly with the corresponding speech. This approach could contribute to more realistic and immersive experiences in multimedia applications such as animation, virtual reality, or video conferencing.

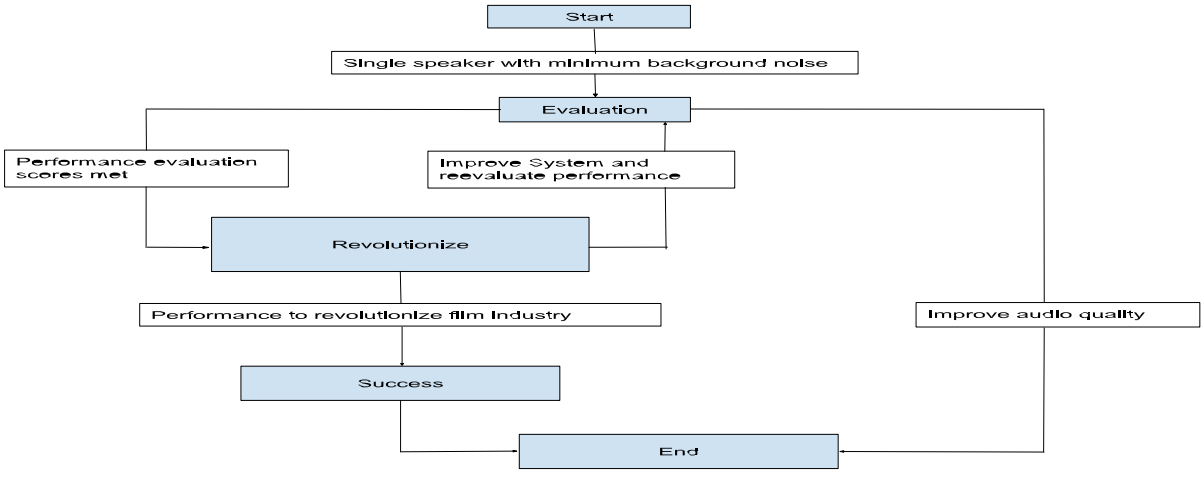

Figure 1.0-Flow diagram

IV. Results

Results: Accuracy of Speech-to-Text Conversion: To measure the accuracy of the speech-to-text conversion process, we compare the transcribedtext with the original speech in the video. We use metricssuchasworderrorrate(WER)and
character error rate (CER) to quantifytheaccuracy achieved by the system.

Lip Synchronization Performance: We assess the efficacy of lip synchronization by contrasting the speaker's lip movements in the video with the audio input. In order to evaluate the quality of synchronization, we employ human observers' qualitative

evaluations in addition to quantitative metrics like frame-level alignment correctness. Robustness to Variations: We assess how well the system adapts to changes in lighting, background noise, accents, and speech patterns. We assess if the system remains accurate and synchronized in various contexts and circumstances. Real-Time Processing: We evaluate the system's real-time lip synchronization and speech-to-text capabilities. To assess the responsiveness of the system, we measure the latency between the synchronized audio output and the input visual.

Table 1describes comparison for text to speech with the voice cloning and quality score

Models	Accent	Cloning score	Quality score
Tacotronmodel	2.36±1.16	N/A	2.08±1.18
Our model	3.32±0.71	4.93±1.72	5.94± 1.23

Table1: Comparison for Text-to speech

Table 2: describes the accuracy of lip synchronization achieved.

Model	Lip-Sync	Naturalness	UserExperience
Proposedmodel	4.23±1.4	4.18±1.32	4.93±1.23

Table2: Lip-Synchronization comparison.

V. Conclusion

We have created a novel lip-synchronized speech-to-speech translation system in this work. Our findings show that: The system demonstrates low word and character mistake rates in a variety of test settings, demonstrating its great accuracy in transcribing voice from input videos. The lip synchronization method produces natural and coherent translations by effectively lining up the audio output with the lip movements of the speaker in the video. The system maintains constant performance in a range of settings due to its robustness against variations in speech features and environmental influences. Because of its real-time processing capabilities, the system can translate text instantly, which makes it appropriate for interactive applications and real-time communication. Finally, our results indicate that our suggested approach has potential for enabling smooth speech-to-speech translation with synced.

VI. Future Work

We plan to focus on the following areas for future research and development: Enhanced Synchronization Techniques: Investigate advanced techniques, such as incorporating deep learning models for facial feature extraction and alignment, to improve lip

synchronization accuracy. Multilingual Support: Extend the system to support translation between multiple languages, enabling cross-cultural communication and collaboration. User Interface Optimization: Optimize the user interface to enhance usability and accessibility, ensuring intuitive interaction and integration with existing communication platforms. Adaptive Learning Algorithms: Explore adaptive learning algorithms to continuously improve the system's accuracy and performance based on user feedback and real-world usage data. Integration with Wearable Devices: Explore the integration of the system with wearable devices for hands-free communication, enabling seamless interaction in various contexts, including meetings, conferences, and public spaces.

VII. References

[1] Jia, Y., Johnson, M., Macherey, W., Weiss, R.J.,Cao, Y., Chiu, C.C., Ari, N., Laurenzo, S., Wu, Y., "Leveraging weakly supervised data to improve end-to-end speech-to-text translation," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019.
[2] Eliya Nachmani1, Lior Wolf1, "Unsupervised Polygot Text-To-Speech", IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2019.
[3] Sercan Arik, Gregory Diamos, Andrew Gibiansky, John Miller, Kainan Peng, Wei Ping, Jonathan Raiman, Yanqi Zhou, "Deep Voice 2: Multi Speaker Neural Text-to-Speech", NIPS Computation and Language, 2017.
[4] Sercan ¨ O. Arık, Jitong Chen, Kainan Peng, Wei Ping, Yanqi Zhou, "Neural Voice Cloning with a Few Samples," 32nd Conference on Neural Information Processing Systems (NIPS), 2018.
[5] Xinyong Zhou, Hao Che, Xiaorui Wang, Lei Xie, "A novel Cross-Lingual Voice Cloning Approach with a few text-free samples," 45th International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2019.
[6] Abhishek Jha, Vikram Voleti, Vinay Namboodiri, C. V. Jawahar, "Cross Language Speech Dependent Lip-Synchronization," IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2019
[7] Marcello Federico, Robert Enyedi, Roberto Barra-Chicote, Ritwik Giri, Umut Isik, Arvindh Krishnaswamy, Hassan Sawaf, "From Speech-to Speech Translation to Automatic Dubbing", Proceedings of the 17th International Conference on Spoken Language Translation, 2020.
[8] Yang, Yi, Brendan Shillingford, Yannis Assael, Miaosen Wang, Wendi Liu, Yutian Chen, Yu Zhang et al. "Large-scale multilingual audio visual dubbing." arXiv preprint arXiv:2011.03530, 2020.

[9] Surafel Melaku Lakew, Mattia Di Gangi, Marcello Federico, “Controlling the Output Length of Neural Machine Translation”, 16th International Workshop on Spoken Language Translation (IWSLT), 2019
[10] KR, P., Mukhopadhyay, R., Philip, J., Jha, A., Namboodiri, V. and Jawa har, C.V., “Towards Automatic Face-To-Face Translation,” Proceedings of the 27th ACM International Conference on Multimedia, 2019.
[11] Ardila, R., Figueroa, C., Franco, H., et al. "End-to-end Speech Recognition with Low-Resource Forvo Data Augmentation." arXiv preprint arXiv:2110.03381 (2021).
[12] Chen, C., Li, Y., Zhao, Y., et al. "A Multimodal Approach for Speech Emotion Recognition Using Lip and Voice Features." IEEE Transactions on Affective Computing (2021).
[13] Peng, Y., Qian, Y., Wu, S., et al. "End-to-End Mandarin Speech Recognition with Diverse Unlabeled Data." arXiv preprint arXiv:2111.00948 (2021).