Video Dubbing with ML-driven Emotional Expression and Lip Synchronization

Akhil & Suprit Chafle & Ayush Krishna Murthi
Indian Institute of Science
Bengaluru, KA, India
{akhil2023,supritk,ayushkrishna}@iisc.ac.in

Abstract

Dubbing videos from one language to another using machine learning (ML) models presents a myriad of challenges, ranging from achieving precise lip synchronization to conveying emotional expression and preserving natural pauses between sentences. While current state-of-the-art models make strides in addressing certain facets of these challenges, they often fall short of delivering optimal results. For instance, the Hierarchical Prosody Model adeptly handles dubbing based on translated text but struggles with lip synchronization nuances. Conversely, models like Wav2Lip excel in ensuring lip synchronization but face obstacles due to output dependency issues arising from the requirement for translated speech.

In light of these limitations, our proposed solution offers a comprehensive ML-based model designed to surmount these obstacles (*illustrated in fig 1*). Our model represents a paradigm shift in the realm of video dubbing by seamlessly integrating critical components such as lip synchronization, translation imbued with emotional expression, and the integration of natural pauses. Notably, our approach distinguishes itself by its minimalistic input requirements, operating solely on the input video (currently in English) to yield the dubbed output (currently in Hindi) through the application of sophisticated ML algorithms.

Central to our model's efficacy is the adoption of a bi-directional LSTM architecture enriched with attention mechanisms. This architectural choice affords our model the capability to capture both forward and backward dependencies, thus preserving long-term contextual information essential for nuanced translation. Furthermore, our model innovatively extracts emotional cues directly from the original speech and text, obviating the need for cumbersome computer vision-based emotion extraction methodologies.

By transcending these technical barriers, our ML-based model aims to revolutionize the landscape of video dubbing, elevating the quality and efficiency of the dubbing process. In doing so, we aspire to enhance accessibility and user experience in the realm of multilingual entertainment consumption, making culturally diverse content more readily accessible and enjoyable for audiences worldwide. Through our concerted efforts, we endeavor to pave the way for a more inclusive and immersive entertainment ecosystem, wherein linguistic barriers are effortlessly transcended, and cultural diversity is celebrated. teams

1 Introduction

In our increasingly diverse global landscape, characterized by the coexistence of numerous cultures and languages, accessing entertainment content often presents a formidable challenge. India, with its staggering array of 780 languages, stands as a testament to the linguistic richness that defines our world. However, this linguistic diversity also erects barriers, particularly evident in the realm of entertainment consumption. In the contemporary era, marked by the rapid proliferation of internet-driven binge-watching, these

barriers manifest as impediments to accessing dubbed versions of films and videos across different languages. Despite the widespread availability of content in various languages, users frequently encounter difficulties in accessing accurately dubbed versions, exacerbating the challenge of language comprehension and cultural appreciation.

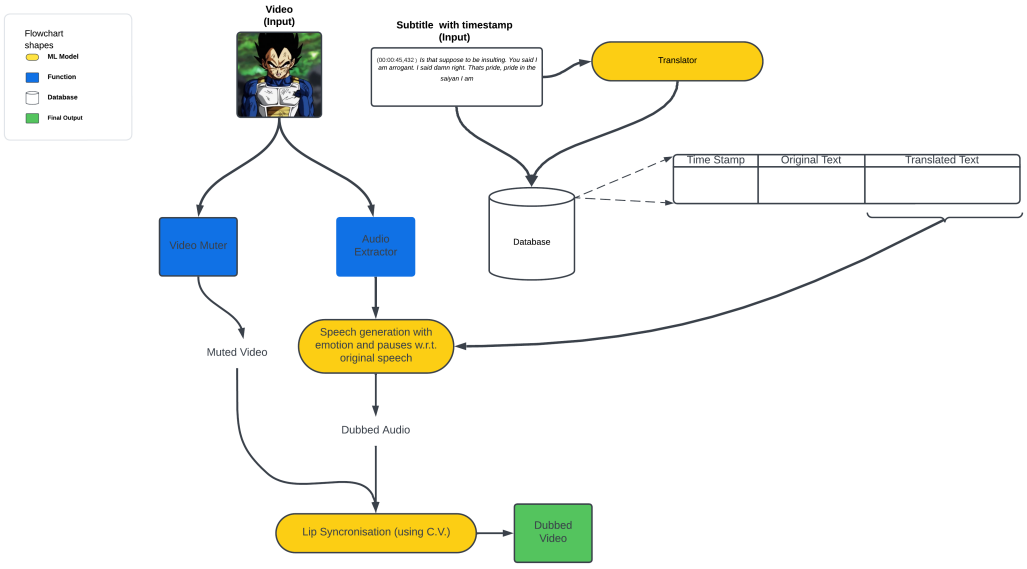

Figure 1: Flow diagram

To tackle these formidable challenges head-on, we propose a novel solution that leverages the latest advancements in artificial intelligence. Our model offers a seamless and efficient approach to dubbing videos, transcending traditional constraints such as lip synchronization and emotional expression (*illustrated in fig 1*). By amalgamating cutting-edge technologies such as the Hierarchical Prosody Model and Wav2Lip, our solution ensures not only precise lip synchronization but also authentic emotional resonance, enhancing the overall viewing experience. Furthermore, our model facilitates translation into alternate languages, with a current focus on Hindi, thereby bridging linguistic divides and expanding accessibility to diverse audiences.

Central to the efficacy of our model is the utilization of the Sequence-to-Sequence bi-directional LSTM with Attention mechanism for translation. This approach affords us distinct advantages over conventional translation models by addressing two pivotal challenges commonly encountered in translation tasks: the limitations imposed by ”no long-term dependencies” and the phenomenon of ”vanishing gradients”.

The incorporation of a solution to the ”no long-term dependencies” challenge enables our model to capture intricate long-range relationships between words, thereby facilitating the establishment of meaningful connections within the text over extensive distances. This capability enhances the contextual understanding and coherence of the translated content, resulting in more accurate and nuanced translations.

Addressing the issue of ”vanishing gradients” is equally imperative, as it directly impacts the effectiveness of the training process. In the context of gradient descent, the phenomenon of vanishing gradients occurs when the gradient approaches zero, leading to stagnation in weight updates. Within the framework of recurrent neural networks (RNNs), such as the bi-directional LSTM employed in our model, vanishing gradients can impede learning by causing gradient values to diminish exponentially across successive timestamps. This phenomenon is particularly pronounced when dealing with lengthy sequences, posing significant challenges to effective model training.

Moreover, the problem of exploding gradients, characterized by uncontrollable growth in gradient values, further complicates the training process. To mitigate this issue, our model incorporates the technique of gradient clipping, which constrains gradient values to prevent them from escalating to unsustainable levels. This strategy enhances the stability and convergence of the training process, thereby facilitating more robust and effective learning outcomes.

In essence, our proposed model represents a pioneering approach to addressing the multi-faceted challenges inherent in dubbing and translation tasks. By harnessing the power of advanced AI technologies and innovative methodologies, we aim to revolutionize the landscape of multilingual entertainment consumption, fostering greater accessibility, inclusivity, and cultural appreciation on a global scale.

2 Related Work

[1] Cong et al. (2022) Our work project relates to the windowing approach proposed in this paper. However we extend and work on improving the idea by selecting a different approach to search for better speech pacing and conversational flow and performing a systematic evaluation of the approach as well as comparing the speech translation model to other proposed models. This observation also allows for further refinement in the future where the approach is combined with more text-based emotion implementation to augment speech accordingly to further improve both convergence and performance.

[2] Prajwal et al. (2020) Lip synchronization is one key aspect of video dubbing. This paper utilizes computer vision to synchronize talking face with speech which can be more useful when combined with speech translation.

[3] Bahdanau et al. (2016): This paper builds upon the foundation of sequence-to-sequence learning and proposes an attention mechanism for NMT. The attention model allows the decoder to selectively focus on relevant parts of the encoded source sentence, leading to improved translation quality.

[4] Sutskever et al. (2014): This work introduces the concept of sequence-to-sequence learning with neural networks. They propose an encoder-decoder architecture that effectively handles variable-length input and output sequences, paving the way for speech translation applications.

3 Methodology

In the provided methodology, a comprehensive outline is presented detailing the intricate workings of the proposed model for video dubbing and translation. The elucidation commences with a succinct depiction of the model's overarching functionality, meticulously delineating its input-output schema and the sequential flow of operations.

Primarily, the model is depicted to ingest an amalgamated input comprising a video file, currently in English, alongside a corresponding subtitle file furnished with timestamps denoting the temporal occurrence of dialogues. This dual input is then channeled into the model's framework to undergo the transformative process, culminating in the generation of a dubbed video, currently in Hindi. This translation process operates line by line, with each subtitle undergoing a meticulous translation to facilitate the synchronization of translated text with the corresponding audio segments.

Illustratively, in Figure 1, the journey of the subtitle file through the model is illustrated with precision. The subtitle file traverses into the model's database, where an additional column is appended to accommodate the translated text alongside the original dialogue. This augmentation enables the establishment of a synchronized database housing both the original and translated dialogues, each annotated with their respective timestamps. Of

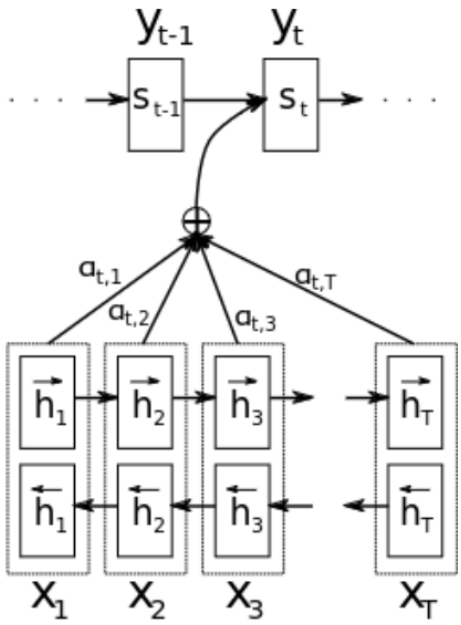

Figure 2: The graphical illustration of the proposed model trying to generate the t-th target word y_t given a source sentence $(x_1, x_2, ..., x_T)$.

particular significance is the delineation of the timestamp format, elucidating its structure wherein the segment preceding the comma denotes the onset time of the dialogue, while the subsequent numerical value represents its duration.

Having thus updated the database with the translated text, the subsequent stage involves the integration of two pivotal inputs: the translated text column and the output from the audio extractor module, which adeptly extracts audio from the original video input. This amalgamated input is subsequently channeled into the Speech Generation module, augmented with emotion and pause incorporation functionalities, tailored to replicate the emotional cadence and natural pauses inherent in the original audio.

The output from the Speech Generation module manifests as the dubbed audio, meticulously synchronized with the translated text. Concurrently, the original video undergoes processing through the Video Muter module, which effectively mutes the video component. These two streams, comprising the dubbed audio and muted video, are seamlessly integrated within the ambit of the Wav2Lip model. This pivotal module not only amalgamates the audio and video streams but also facilitates robust lip synchronization with the translated text, ensuring a coherent viewing experience.

Delving into the technical nuances, the methodology underscores the adoption of a Sequence-to-Sequence Bidirectional LSTM model with Attention mechanism for numerical machine translation. Leveraging the Hindi_English_Truncated_Corpus dataset sourced from Kaggle, the model undergoes meticulous training, aided by tokenization using byte pair encoding and the incorporation of specialized tokens such as 'SOS', 'EOS', and 'PAD'. Furthermore, the bidirectional LSTM architecture is instrumental in capturing future dependencies, essential for language modeling (*illustrated in the figure 2 below*), wherein the future words exert influence on the current word.

Integral to the translation process is the computation of attention weights, computed by alignment model which include calculation of energy function (shown in equation number 3), then softmax (shown in equation number 2) and followed by calculation of context vector

(shown in equation 1) which illuminate the interplay between input and target tokens. These attention weights serve as a conduit for capturing long-term dependencies, elucidating the semantic relationships between words. A dedicated feed-forward neural network is employed to discern the align function, with learnable parameters trained iteratively during the model's training phase.

$$c_i = \sum_{j=1}^{T_x} \alpha_{ij} h_j \tag{1}$$

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{T_x} \exp(e_{ik})} \tag{2}$$

$$e_{ij} = a(s_{i-1}, h_j) \tag{3}$$

To ensure model robustness and mitigate over-fitting, a suite of regularization techniques is employed, encompassing early stopping, dropout, and teacher forcing. Additionally, measures are instituted to combat the challenge of exploding gradients, gradient clipping is used (*illustrated in fig 3 below*) where we set one threshold value, if the the gradient is greater then the threshold value then normalise the gradient and multiply with threshold value.

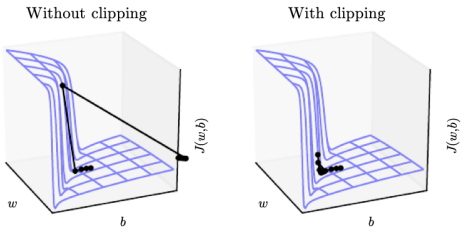

Figure 3: Example of the effect of gradient clipping in a recurrent network with two parameters w and b.

Looking ahead, the methodology charts a course towards the implementation of speech generation techniques for the translated text, accompanied by the infusion of emotion to engender an evocative auditory experience. This roadmap culminates in the seamless integration of disparate modules, orchestrated to orchestrate a dubbed video resonant with the emotional tenor and linguistic fidelity of the original. Central to this integration is the symbiotic alliance with pre-existing methodologies such as Wav2Lip, poised to augment lip synchronization and elevate the viewing experience to unprecedented heights.

4 Results

The dataset preparation phase has been completed, encompassing tasks such as data collection, preprocessing, and organization. Additionally, tokenization procedures have been executed to appropriately structure the data for subsequent processing. However, during the implementation of the trainer class, an unexpected dimensional mismatch error was encountered, hindering the smooth progression of the project. Efforts are currently underway to rectify this issue, and updates regarding the resolution of the error and subsequent results will be provided in a timely manner.

Conclusion

Up to this point, our endeavors have culminated in the successful implementation of translation functionalities utilizing the Sequence2Sequence bi-directional LSTM architecture with

attention mechanisms, alongside the creation of lip-synchronization capabilities. Moving forward, our focus shifts towards the integration of translated text-to-speech functionalities, meticulously preserving the emotional nuances and pauses inherent in the original speech. This concerted effort aims to achieve seamless alignment between the dubbed audio and the original video, effectively resolving challenges associated with lip-sync while ensuring the faithful preservation of emotional expression. By bridging these critical gaps, our model emerges as a comprehensive solution poised to enhance the quality and authenticity of dubbed content, thereby enriching the viewing experience for audiences worldwide. Additionally, our model can be adapted for multilingual applications.

Contributions

In this project, the team members, namely Akhil, Ayush, and Suprit, contributed as follows:

- Akhil: Curated datasets essential for implementing Wav2Lip, focusing on integrating emotional elements into the text-to-speech (TTS) synthesis process, and delved into understanding the Hierarchical Prosody Model (HPM).

- Ayush: Spearheaded translation efforts utilizing the Seq2Seq unidirectional LSTM model, alongside extensive exploration of the Wav2Lip model.

- Suprit: Led translation tasks leveraging the Seq2Seq bidirectional LSTM model with attention mechanisms, while also initiating exploration into Neural Machine Translation (NMT) models. Additionally, ongoing contributions involve further investigation and refinement of NMT methodologies.

References

[1] Gaoxiang Cong, Liang Li, Yuankai Qi, Zhengjun Zha, Qi Wu, Wenyu Wang, Bin Jiang, Ming-Hsuan Yang, and Qingming Huang. Learning to Dub Movies via Hierarchical Prosody Models. arXiv preprint arXiv:2212.04054, 2022.

[2] K R Prajwal, Rudrabha Mukhopadhyay, Vinay Namboodiri, and C V Jawahar. A Lip Sync Expert Is All You Need for Speech to Lip Generation In The Wild. arXiv preprint arXiv:2008.10010, 2020.

[3] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural Machine Translation by Jointly Learning to Align and Translate. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR 2016)*, pages 1–11, 2016.

[4] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to Sequence Learning with Neural Networks. arXiv preprint arXiv:1409.3215, 2014.

A Appendix

[1] A Comprehensive Guide to Neural Machine Translation using Seq2Seq Modelling using PyTorch

[2] LSTM – PyTorch 2.2 documentation