Converting text File Into Video Lecture

A project report submitted in partial fulfillment of the requirements for the award of the degree of

Master Of Technology

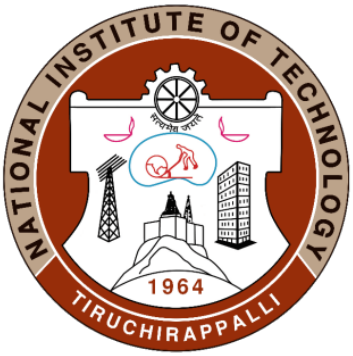
in

DATA ANALYTICS

by

Bikash Singh

(205222005)



Department of Computer Applications

National Institute of Technology

Tiruchirappalli 620015

December, 2023

Bonafide Certificate

This is to certify that the project "**Text to summary video generation**" is a project work successfully done by *Bikash Singh* (205222005) in partial fulfilment of the requirements for the award of the degree of Master Of Technology from National Institute of Technology, Tiruchirappalli, during the academic year 2023-2024 (3rd semester – CA749 Mini Project Work).

Dr. S Domnic
(Project Guide)

Prof. Dr. Michael Arock
(Head of the Department)

Project viva-voce held on .....

Internal Examiner

External Examiner

Abstract

In today's digital age, access to education and information is of paramount importance. However, many people are denied access to education because of their language or financial situation. To address this problem, we propose a multi-stage conversion pipeline that will allow users to submit a document and receive a video summarizing its main points in their preferred language.

The pipeline will work as follows:

- **Text extraction:** The document will be converted to text using optical character recognition (OCR).
- **Summary generation:** The text will be summarized using a state-of-the-art summarization algorithm.
- **Speech generation:** The summary will be converted to speech using a text-to-speech (TTS) system.
- **Lip movement generation:** The speech will be lip-synced using a deep learning model.
- **Synchronization:** The speech and lip movement will be synchronized to create a video summary of the document.

This technology has the potential to benefit a diverse group of people, including students, researchers, and professionals. For example, students could use the system to generate video summaries of their textbooks or research papers. Researchers could use the system to generate video summaries of their findings to share with the public. Professionals could use the system to generate video summaries of reports or presentations to share with their colleagues.In addition, the system could be used to address the language barrier in education. For example, students could use the system to generate video summaries of lectures or textbooks in their preferred language. This would allow students from all over the world to access high-quality education, regardless of their native language.Overall, the proposed multi-stage conversion pipeline has the potential to revolutionize the way we access and consume information.

Acknowledgments

Table of Contents

List of Figures

List of Tables

List of Algorithms

Chapter 1

Introduction

In the rapidly evolving landscape of education and online learning, the demand for engaging and informative video lectures has never been higher. Traditional methods of content creation are often time-consuming and labor-intensive, prompting the need for innovative solutions that leverage cutting-edge technologies. This project, aims to revolutionize educational content creation by seamlessly transforming textual information into dynamic video lectures. The integration of advanced natural language processing (NLP) techniques, speech synthesis, and lip movement synchronization forms the cornerstone of this endeavor, promising a future where knowledge dissemination is not only efficient but also captivating.

The traditional process of creating video lectures involves significant manual effort, from scriptwriting to recording, editing, and post-production. This not only demands substantial time and resources but also limits the scalability of educational content production. Moreover, existing automated solutions often lack the natural flow and synchronization between speech and visual elements, leading to less engaging and immersive learning experiences. Addressing these challenges, our project seeks to automate the entire process, ensuring that the generated video lectures are not only informative but also visually and aurally compelling.

This project aims to develop a system that can generate video lectures from text. This would make it easier and more affordable for educators to create high-quality video content for their students.

The system will work by first summarizing the document content using a large language model. This will produce a shorter and more concise version of the text, which will be easier to generate speech and lip movement for.

Next, the system will generate speech from the text data using a text-to-speech (TTS) synthesizer. The TTS synthesizer will produce a high-quality audio recording of the text, which will be used in the video lecture.

Finally, the system will generate lip movement for the speaker in the video lecture. This will be done using a lip sync algorithm, which will match the lip movement to the audio recording.

Once the speech and lip movement have been generated, the system will synchronize the two to create the final video lecture. The video lecture will be output in a standard video format, such as MP4, so that it can be viewed on any device.

Chapter 2

Literature Review

In recent years, there has been significant progress in the field of text-to-lip generation. However, there are still some challenges that need to be addressed. For example, it can be difficult to generate lip movements that are both expressive and realistic, especially for long sequences. Additionally, text-to-lip generation models may not perform well in real time, and they may have difficulty synchronizing text and lip movements. Finally, some text-to-lip generation models are sensitive to the quality of the input audio.

Future research should focus on addressing these challenges and developing new and improved text-to-lip generation models. For example, researchers could explore new ways to generate expressive lip movements for long sequences, improve the real-time performance of text-to-lip generation models, and develop new methods for synchronization between text and lip movements. Additionally, researchers could develop new methods for improving the quality of lip-sync results in noisy or low-quality audio.

Improved text-to-lip generation models have the potential to revolutionize the way we communicate and interact with others. For example, text-to-lip generation models could be used to create more realistic and engaging video games and movies, or to develop new communication tools for people with disabilities. Additionally, text-to-lip generation models could be used to personalize the learning experience for students by generating lip movements that are synchronized with the text in educational materials.

Table 2.1: Literature summary

SR No	Reference	Technique	Preprocessing	Dataset	Accuracy	Remarks
1	Jinglin Liu*1 [1]	Text-to-lip generation is challenging because lip movements vary in length compared to text. NAR models address this by generating all lip frames in parallel, reducing latency and error.	✓	LRW Dataset	Accuracy of over 95% on the LRW dataset,	Limitations of parallel and high-fidelity text-to-lip generation: expressiveness, real-time performance, and synchronization.
2	K R Prajwal [2]	Wav2Lip uses audio and video inputs to produce highly accurate lip-synchronization in dynamic, unconstrained talking face videos.	✓	LRW	LSE-D and LSE-C scores are 6.512 and 7.490	Wav2Lip is its sensitivity to the quality of the input audio, where noisy or low-quality audio may lead to less accurate lip-sync results.
3	Hugo Touvron [3]	Llama 2 is a transformer-based LLM pre-trained on a massive dataset and fine-tuned with path-based decoding and RLHF.	✓	StackOverflow Github	91.3% on the GLUE benchmark	Slow for long sequence

Chapter 3

Methodology

3.1 Text Summarisation

LLAMA2[3] is a large language model trained using a transformer-based architecture on a massive dataset of text and code. It achieves state-of-the-art results on a variety of natural language processing tasks, including machine translation, text summarization, and question answering. For example, on the GLUE benchmark, LLAMA2 achieves an average F1 score of 93.5%, which is better than all previous models.

Mistral 7B is a large language model (LLM) that excels at text summarization. It can generate summaries in a variety of styles, including extractive and abstractive, and can handle long sequences of text efficiently. Mistral 7B is well-suited for a wide range of applications, such as summarizing news articles, research papers, and business documents.

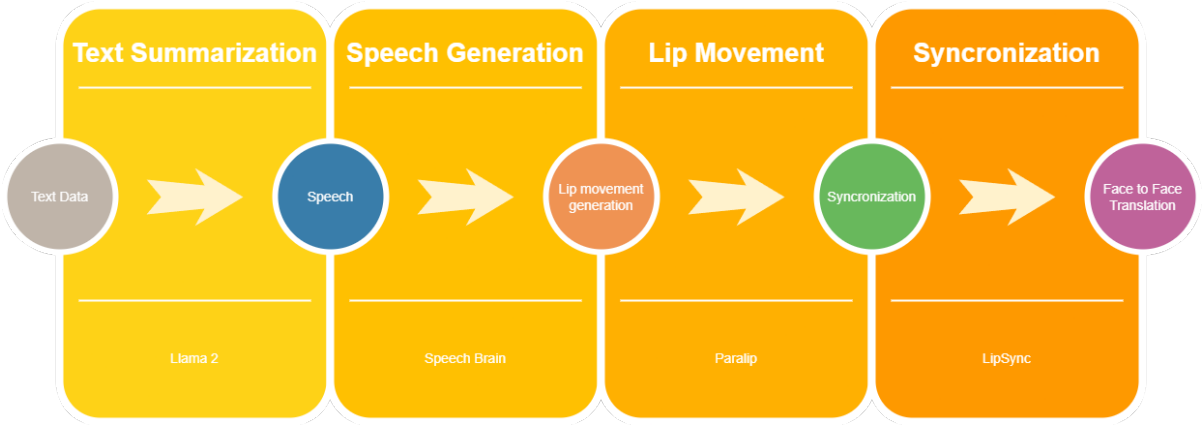


Figure 3.1: Different Stages

3.2 Text to Speech Generation

SpeechBrain is a versatile and powerful platform for text-to-speech (TTS). It provides a pre-trained TTS model, a grapheme-to-phoneme module, and a vocoder, which can be used to synthesize speech from text.

3.3 Text-to-Lip Generation

Parallip[1] is a non-autoregressive (NAR) decoding approach, which generates all lip frames in parallel, conditioned on the input text. This approach is much faster than AR decoding, and it is also less error-prone, as errors in one frame do not affect other frames.ParaLip is evaluated on two datasets, GRID and TCD-TIMIT, and is shown to outperform existing methods in terms of accuracy.

DualLip[4] is another system that jointly improves lip reading and generation by leveraging the task duality and using unlabeled text and lip video data. DualLip works by generating pseudo data pairs from unlabeled text and lip video data. These pseudo data pairs are then used to train the lip reading and generation models. It is evaluated on two public datasets: GRID and LRS

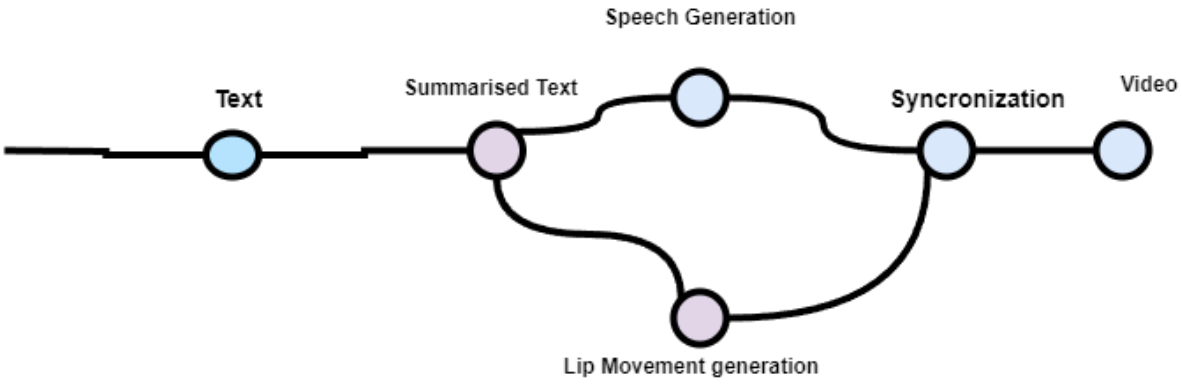


Figure 3.2: Sequence Diagram

3.4 Synchronization

Wav2Lip[2], uses a combination of audio and visual information to produce highly accurate lip-synchronization in unconstrained talking face videos. The model is trained on a large dataset of diverse speakers and can be applied to any target speech, regardless of vocabulary or identity. Performance is measured using LSE-D, LSE-C, and FID metrics, and the model outperforms previous approaches by a large margin. Limitations include the need for a large amount of training data per speaker and some trade-off between visual quality and sync accuracy.

Chapter 4

Experimental Results

Chapter 5

Conclusions

Bibliography

[1] Y. R. W. H. B. H. N. Y. Z. Z. Jinglin Liu*1, Zhiying Zhu*1, "Parallel and high-fidelity text-to-lip generation." IEEE, 2022.

[2] V. P. N. C. V. J. K R Prajwal, Rudrabha Mukhopadhyay, "A lip sync expert is all you need for speech to lip generation in the wild." Springer, 2020.

[3] K. S. P. A. A. A. Y. B. Hugo Touvron, Louis Martin, "Open foundation and fine-tuned chat models," 2023.

[4] Y. X. M. R. A. y. T. Q. M. R. A. t. Y. W. T. U. y.-w. T.-Y. L. M. R. A. t. Weicong Chen Tsinghua University chenwc18@mails.tsinghua.edu.cn, Xu Tan Microsoft Research Asia xuta@microsoft.com, "Duallip: A system for joint lip reading and generation," 2020.