VISVESVARAYA TECHNOLOGICAL UNIVERSITY

BELAGAVI – 590018, Karnataka

INTERNSHIP REPORT

ON

"Lip To Speech Synthesis"

Submitted in partial fulfilment for the award of degree(18CSI85)

BACHELOR OF ENGINEERING IN
YOUR BRANCH

Submitted by:

NAME:- JAYAKEERTHANA S

USN:-1GG20CS010

Conducted At

VARCONS TECHNOLOGIES
213, 2st Floor,
18 M G Road, Ulsoor,Bangalore-560001

GOVERNMENT ENGINEERING COLLEGE

B.M.Road,Ramanagara-562 159
2023-24

GOVERNMENT ENGINEERING COLLEGE

B.M.Road,Ramanagara-562 159
Affiliated to VTU, Belagavi, APPROVED by AICTE New Delhi

CERTIFICATE

This is to certify that the Internship titled **"Lip to Speech Synthesis""** carried out by **Ms. Jayakeerthana S,** a bonafide student of Government Engineering College Ramanagara, in partial fulfillment for the award of **Bachelor of Engineering**, in **Computer Science and Engineering branch** under Visvesvaraya Technological University,Belagavi, during the year 2023-2024. It is certified that all corrections/suggestions indicated have been incorporated in the report.
The project report has been approved as it satisfies the academic requirements in respect of Internship prescribed for the course Internship / Professional Practice (18CSI85)

Signature of Guide	**Signature of HOD**	**Signature of Principal**

External Viva:

Name of the Examiner	Signature
1. ……..	1. …….
2. ……..	2. …….

D E C L A R A T I O N

I, **Jayakeerthana S**, final year student of student of Computer Science and Engineering Branch, Government Engineering College Ramanagara - 560 082, declare that the Internship has been successfully completed, in **VARCONS TECHNOLOGIES**. This report is submitted in partial fulfillment of the requirements for award of Bachelor Degree in Computer Science and Engineering branch, during the academic year 2023-2024.

Date : 18/04/2024 :

Place : Mandya

USN : 1GG20CS010

NAME : JAYAKEERTHANA S

OFFER LETTER

INTERNSHIP OFFER LETTER

Date: 12th March, 2024

Name: **JAYAKEERTHANA S**
USN: **1GG20CS010**

Dear Student,

We would like to congratulate you on being selected for the **Machine Learning With Python (Research Based)** Internship position with **Varcons Technologies**, effective Start Date **12th March, 2024**. All of us are excited about this opportunity provided to you!

This internship is viewed as being an educational opportunity for you, rather than a part-time job. As such, your internship will include training/orientation and focus primarily on learning and developing new skills and gaining a deeper understanding of concepts of **Machine Learning With Python (Research Based)** through hands-on application of the knowledge you learn while you train with the senior developers. You will be bound to follow the rules and regulations of the company during your internship duration.

Again, congratulations and we look forward to working with you!.

Sincerely,

Spoorthi H C
Director
VARCONS TECHNOLOGIES
213, 2st Floor,
18 M G Road, Ulsoor,
Bangalore-560001

A C K N O W L E D G E M E N T

This Internship is a result of accumulated guidance, direction and support of several important persons. We take this opportunity to express our gratitude to all who have helped us to complete the Internship.

We express our sincere thanks to our Principal, for providing usadequate facilities to undertake this Internship.

We would like to thank our Head of Dept – branch code, for providing us an opportunity to carry out Internship and for his valuable guidance and support.

We would like to thank our (Lab assistant name) Software Services for guiding us during the period of internship.

We express our deep and profound gratitude to our guide, Guide name, Assistant/Associate Prof, for her keen interest and encouragement at every step in completing the Internship.

We would like to thank all the faculty members of our department for the support extended during the course of Internship.

We would like to thank the non-teaching members of our dept, forhelping us during the Internship.

Last but not the least, we would like to thank our parents and friends without whose constant help, the completion of Internship would have not been possible.

NAME:
JAYAKEERTHA S
USN:1GG20CS010

ABSTRACT

In this paper, we propose a novel lip-to-speech generative adversarial network, Visual Context Attention GAN (VCA-GAN), which can jointly model local and global lip movements during speech synthesis. Specifically, the proposed VCAGAN synthesizes the speech from local lip visual features by finding a mapping function of vise me-to-phone me, while global visual context is embedded into the intermediate layers of the generator to clarify the ambiguity in the mapping induced by homophone. To achieve this, a visual context attention module is proposed where it encodes global representations from the local visual features, and provides the desired global visual context corresponding to the given coarse speech representation to the generator through audiovisual attention. In addition to the explicit modeling of local and global visual representations, synchronization learning is introduced as a form of contrastive learning that guides the generator to synthesize a speech in sync with the given input lip movements. Extensive experiments demonstrate that the proposed VCA-GAN outperforms existing state-of-the-art and is able to effectively synthesize the speech from multi-speaker that has been barely handled in the previous works.

Table of Contents

CHAPTER 1

COMPANY PROFILE

1. COMPANY PROFILE

A Brief History of Compsoft Technologies

Compsoft Technologies, was incorporated with a goal ”To provide high quality and optimal Technological Solutions to business requirements of our clients”. Every business is a different and has a unique business model and so are the technological requirements. They understand this and hence the solutions provided to these requirements are different as well. They focus on clients requirements and provide them with tailor made technological solutions. They also understand that Reach of their Product to its targeted market or the automation of the existing process into e-client and simple process are the key features that our clients desire from Technological Solution they are looking for and these are the features that we focus on while designing the solutions for their clients.

Sarvamoola Software Services. is a Technology Organization providing solutions for all web design and development, MYSQL, PYTHON Programming, HTML, CSS, ASP.NET and LINQ. Meeting the ever increasing automation requirements, Sarvamoola Software Services. specialize in ERP, Connectivity, SEO Services, Conference Management, effective web promotion and tailor-made software products, designing solutions best suiting clients requirements.

Compsoft Technologies, strive to be the front runner in creativity and innovation in software development through their well-researched expertise and establish it as an out of the box software development company in Bangalore, India. As a software development company, they translate this software development expertise into value for their customers through their professional solutions.

They understand that the best desired output can be achieved only by understanding the clients demand better. Compsoft Technologies work with their clients and help them todefiine their exact solution requirement. Sometimes even they wonder that they have completely redefined their solution or new application requirement during the brainstormingsession, and here they position themselves as an IT solutions consulting group comprising of high caliber consultants.

They believe that Technology when used properly can help any business to scale and achieve new heights of success. It helps Improve its efficiency, profitability, reliability; to put itin one sentence ” Technology helps you to Delight your Customers” and that is what we wantto achieve.

CHAPTER 2

ABOUT THE COMPANY

ABOUT THE COMPANY

Compsoft Technologies is a Technology Organization providing solutions for all web design and development, MYSQL, PYTHON Programming, HTML, CSS, ASP.NET and LINQ. Meeting the ever increasing automation requirements, Compsoft Technologies specialize in ERP, Connectivity, SEO Services, Conference Management, effective webpromotion and tailor-made software products, designing solutions best suiting clients requirements. The organization where they have a right mix of professionals as a stakeholders to help us serve our clients with best of our capability and with at par industry standards.They have young, enthusiastic, passionate and creative Professionals to develop technological innovations in the field of Mobile technologies, Web applications as well as Business and Enterprise solution. Motto of our organization is to "Collaborate with our clients to provide them with best Technological solution hence creating Good Present and Better Future for our client which will bring a cascading a positive effect in their business shape as well". Providing a Complete suite of technical solutions is not just our tag line, it is Our Vision for Our Clients and for Us, We strive hard to achieve it.

Products of Compsoft Technologies.

Android Apps

It is the process by which new applications are created for devices running the Android operating system. Applications are usually developed in Java (and/or Kotlin; or other such option) programming language using the Android software development kit (SDK), but other development environments are also available, some such as Kotlin support the exact same Android APIs (and bytecode), while others such as Go have restricted API access.

The Android software development kit includes a comprehensive set of development tools. These include a debugger, libraries, a handset emulator based on QEMU, documentation, sample code, and zutorials. Currently supported development platforms include computers running Linux (any modern desktop Linux distribution), Mac OS X 10.5.8 or later, and Windows 7 or later. As of March 2015, the SDK is not available on Android itself, but softwaredevelopment is possible by using specialized Android applications.

Web Application

It is a client–server computer program in which the client (including the user interface and client- side logic) runs in a web browser. Common web applications include web mail, online

retail sales, online auctions, wikis, instant messaging services and many other functions. web applications use web documents written in a standard format such as HTML and JavaScript,which are supported by a variety of web browsers. Web applications can be considered as a specifific variant of client–server software where the client software is downloaded to the client machine when visiting the relevant web page, using standard procedures such as HTTP. The Client web software updates may happen each time the web page is visited. During the session, the web browser interprets and displays the pages, and acts as the universal client for any web application. The use of web application frameworks can often reduce the number of errors in a program, both by making the code simpler, and by allowing one team to concentrate on the framework while another focuses on a specifified use case. In applications which are exposed to constant hacking attempts on the Internet, security-related problems can be caused by errors in the program.

Frameworks can also promote the use of best practices such as GET after POST. There are some who view a web application as a two-tier architecture. This can be a “smart” client that performs all the work and queries a “dumb” server, or a “dumb” client that relies on a “smart” server. The client would handle the presentation tier, the server would have the database (storage tier), and the business logic (application tier) would be on one of them or on both. While this increases the scalability of the applications and separates the display and the database, it still doesn‟t allow for true specialization of layers, so most applications will outgrow this model. An emerging strategy for application software companies is to provide web access to software previously distributed as local applications. Depending on the type of application, it may require the development of an entirely different browser-based interface, or merely adapting an existing application to use different presentation technology. These programs allow the user to pay a monthly or yearly fee for use of a software application without having to install it on a local hard drive. A company which follows this strategy is known as an application service provider (ASP), and ASPs are currently receiving much attention in the software industry.

Security breaches on these kinds of applications are a major concern because it can involve both enterprise information and private customer data. Protecting these assets is an important part of any web application and there are some key operational areas that must be included in the development process. This includes processes for authentication, authorization, asset handling, input, and logging and auditing. Building security into the applications from the beginning can be more effective and less disruptive in the long run.

Web design

It is encompasses many different skills and disciplines in the production and maintenance of websites. The different areas of web design include web graphic design; interface design; authoring, including standardized code and proprietary software; user experience design; and

search engine optimization. The term web design is normally used to describe the design process relating to the front-end (client side) design of a website including writing mark up. Web design partially overlaps web engineering in the broader scope of web development. Web designers are expected to have an awareness of usability and if their role involves creating mark up then they are also expected to be up to date with web accessibility guidelines. Web design partially overlaps web engineering in the broader scope of web development.

Departments and services offered

Compsoft Technologies plays an essential role as an institute, the level of education, development of student's skills are based on their trainers. If you do not have a good mentor then you may lag in many things from others and that is why we at Compsoft Technologies gives you the facility of skilled employees so that you do not feel unsecured about the academics. Personality development and academic status are some of those things which lie on mentor's hands. If you are trained well then you can do well in your future and knowing its importance of Compsoft Technologies always tries to give you the best.

They have a great team of skilled mentors who are always ready to direct their trainees in the best possible way they can and to ensure the skills of mentors we held many skill development programs as well so that each and every mentor can develop their own skills with the demands of the companies so that they can prepare a complete packaged trainee.

Services provided by Compsoft Technologies.

- Core Java and Advanced Java
- Web services and development
- Dot Net Framework
- Python
- Selenium Testing
- Conference / Event Management Service
- Academic Project Guidance
- On The Job Training
- Software Training

CHAPTER 3

INTRODUCTION

2. INTRODUCTION

Introduction to ML

Machine learning is a subfield of artificial intelligence that focuses on developing algorithms and models that enable computers to learn from data and make predictions or decisions without being explicitly programmed. It involves the use of statistical techniques to enable computers to improve their performance on a specific task through experience**.**

An innovative branch of artificial intelligence (AI) called machine learning has emerged as a game-changing innovation that enables computers to learn and make judgements without explicit programming. The healthcare, financial, entertainment, and transportation sectors have all seen radical change as a result of this shift in the computer paradigm. Fundamentally, machine learning enables computers to analyse data, spot patterns, and make predictions or judgements in a manner similar to how people learn through trial and error.

The concept of machine learning stems from the idea that computers can be designed not just to perform predefined tasks, but to adapt and improve their performance based on data. This ability to learn from data enables machines to tackle complex problems that were once considered beyond their reach. Whether it's recognizing faces in images, recommending personalized content on streaming platforms, predicting stock prices, or diagnosing medical conditions, machine learning algorithms have demonstrated their efficacy in diverse domains.

This introduction explores the main ideas, types, and applications of machine learning in an effort to provide readers a basic grasp of the subject. We will explore the underlying ideas and vocabulary that underlie machine learning, as well as the wide variety of algorithms at its disposal and the broad array of practical applications that make use of this game-changing technology, as we go deeper into this fascinating area. This investigation will be a helpful beginning place for your research of the machine learning world, whether you are a newbie excited by the possibilities of machine learning or a seasoned practitioner seeking deeper insights.

Problem Statement

Lip To Speech Synthesis

with key design choices to achieve accurate lip to speech synthesis in unconstrained scenarios

Goal: Understand the working of Lip To Speech Analysis and Improve the accuracy

proposed to generate text from lip movements, most of them are limited to small vocabulary space**.** forecasting, but they want to explore machine learning techniques to enhance accuracy.

The problem statement could be: "Develop a machine learning model to predict understanding speech by visually interpreting the movements of the lips. The goal is to improve sales forecasting accuracy by at least 10% compared to the current methods."

CHAPTER 4

SYSTEM ANALYSIS

4. SYSTEM ANALYSIS

1. Existing System

To the best of our knowledge, this is the first attempt to use 3D convolutional neural networks for audio-visual matching in which a bridge between spatio-temporal features has been established to build a common feature space between audiovisual modalities. Our source code1 has been released online as an open source project The audio-video synchronization process, which calls on audio-visual matching skills, is one of the most difficult applications of audio- visual recognition. In order to determine how well the two modalities coincide, various audio-visual identification tasks have been used in the research for this work. Various methods have been used to address the audio-visual matching issue. Canonical Correlation Analysis (CCA) and Co-Inertia Analysis are two methods that are based on data-driven ways to calculate the off-sync time (CoIA).

2.Proposed System

Since an audio speech and lip movements in a single video are supposed to be aligned in time, the speech can be synthesized to have the same duration as the input silent video. Let x RT ×H×W ×C be a lip video with T frames, height of H, width of W, and channel size of C. Then, our objective is to find a generative model that synthesizes a speech Y, RF ×4T, where y is a target Mel-spectrogram with F Mel-spectral dimension and frame length of 4T. The frame length of Mel-spectrogram is designed to be 4 times longer than that of video by adjusting the hop length during Short-Time Fourier Transform (STFT). To generate elaborate speech representations, the proposed VCA-GAN refines the viseme-to-phoneme mapping with the global visual context obtained from a visual context attention module, and learns to produce a synchronized speech with given lip movements. Please note that we treat the Mel-spectrogram as an image and train the model with 2D GAN

3.Objective of the System

Parallel to the development of Lip2Speech, Visual Speech Recognition (VSR) have achieved a great advancement. Slightly different from the Lip2Speech, VSR identifies spoken speech into text by watching a silent talking face video. Several works have recently showed state-of-the-art performances in word- and sentence-level classifications. proposed a largescale audio-visual

dataset and set a baseline model for word-level VSR. Stafylakis et al. proposed an architecture that is combined of residual network and LSTM, which became a popular.

Architecture for word-level lip reading. Martinez et al. replaced the RNN-based backend with Temporal Convolutional Network (TCN). Proposed to utilize audio modal knowledge through memory network without audio inputs during inference for lip reading. Achieved end-to-end sentence-level lip reading network by adopting the CTC loss. Different from the VSR methods, the Lip2Speech task does not require human annotations, thus is drawing big attention with its practical aspects.

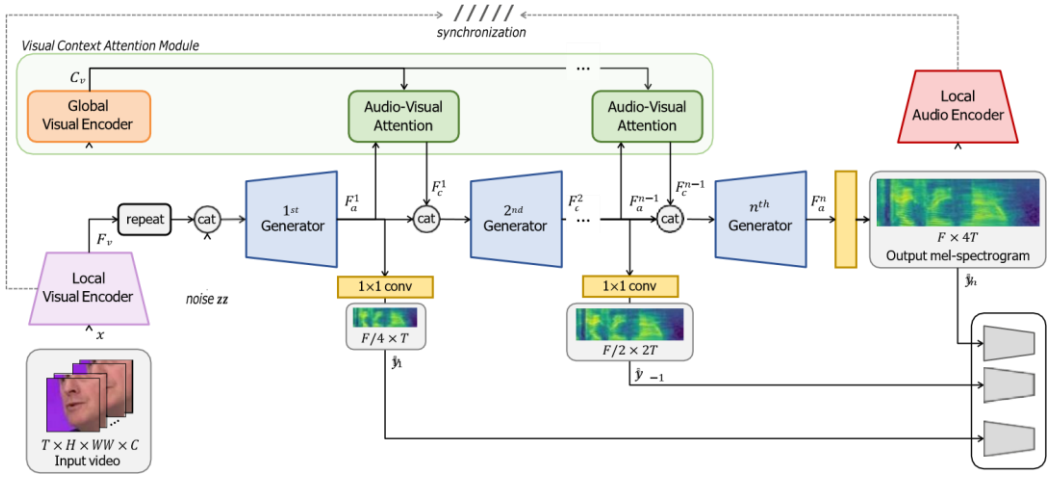


Figure: Overview of the VCA-GAN.

CHAPTER 5

REQUIREMENT ANALYSIS

5. REQUIREMENT ANALYSIS

Hardware Requirement Specification

- CPU: Intel Core i5 or equivalent.
- RAM: 4 GB RAM is recommended for processing small to moderate amounts of data. However, if you plan to process large datasets or run multiple instances concurrently, consider 8 GB or more.
- Storage: You'll need enough storage space for the code, dependencies, and data
- Network Connectivity: A stable internet connection is necessary for collecting data from online sources, such as Twitter and external sentiment analysis services.

Software Requirement Specification

A] Functional Requirements

1. Data Retrieval:

- The system must fetch real-time stock price data from Yahoo Finance for a specified stock symbol.
- It should make periodic HTTP requests to the Yahoo Finance API to retrieve the latest data.

2. Data Processing:

- The system must process the raw data to extract relevant information, including the stock's last price, date and time of the update, price change, high and low prices, and trading volume.
- Data processing should handle scenarios where data may be missing or null.

3.Data Storage:

- The system must store the processed stock price data in an Elasticsearch index.
- It should create an index with predefined mappings for efficient data storage and retrieval.

4.Configuration:

The system should allow users to configure various parameters through a command-line interface (CLI), including the Elasticsearch index name, whether to delete an existing index, the stock symbol to monitor, and the update frequency.

5.Logging:

- The system should provide detailed logging of its activities, including data retrieval, processing, and storage.

- Logging should be configurable based on the verbosity level specified by the user.

6.Error Handling:

- The system must handle errors gracefully, log exceptions, and continue operation even in the presence of errors.
- It should handle exceptions related to data retrieval, processing, Elasticsearch interaction, and other potential issues.

7.Index Management:

- The system should allow users to choose whether to delete an existing Elasticsearch index before creating a new one.
- It should create the index with appropriate mappings if it doesn't already exist.

8.Data Integrity:

- The system should ensure that data integrity is maintained by checking for and handling missing or null values in the retrieved data.
- It should validate that essential data fields (e.g., last price, date) are present and not empty before storing data in Elasticsearch.

9.Continuous Operation:

- The system should run continuously, periodically fetching and storing stock price data based on the specified update frequency.
- It should handle keyboard interrupts (Ctrl-C) gracefully to allow users to stop the operation.

10.Elasticsearch Connection:

- The system should establish a connection to an Elasticsearch server using the provided credentials (host, port, user, and password).
- It should handle authentication and establish secure communication with Elasticsearch.

B] Non-Functional Requirements

1. **Performance:**
   - **Response Time:** The system should have low latency in fetching and storing data to ensure real-time or near-real-time data updates.

- **Throughput:** It should handle a reasonable volume of HTTP requests and data storage operations per unit of time.

2. **Scalability:**
   The code should be able to scale horizontally to handle a growing number of stock symbols or increased data volume.

3. **Reliability:**
   - The system should operate continuously without frequent failures, ensuring reliability in data collection and storage.
   - It should recover gracefully from transient errors, such as network interruptions.

4. **Availability:**
   - The system should have a high level of availability, minimizing downtime for data collection.
   - Measures should be in place to ensure system availability, even in the face of server failures or maintenance activities.

5. **Security:**
   - **Data Security:** The code should ensure the security of sensitive data, including Elasticsearch credentials, during transmission and storage.
   - **Access Control:** Access to the Elasticsearch server and the code itself should be restricted to authorized users only.

6. **Maintainability:**
   - The code should be well-documented, making it easy for developers to understand, modify, and maintain.
7. **Logging and Monitoring:**
   - The system should provide comprehensive logging for monitoring and debugging purposes.
   - It should integrate with monitoring and alerting tools to detect and respond to issues promptly.

8. **Compatibility:**

The code should be compatible with various operating systems (e.g., Windows, macOS, Linux) to accommodate diverse deployment environments.

9. **Resource Efficiency:**
    - The system should use system resources efficiently to minimize CPU and memory utilization.
    - It should release resources when not in use to prevent resource exhaustion.

10. **Usability:**
    - The command-line interface (CLI) should be user-friendly, with clear and concise options and help messages.
    - Error messages should be informative and actionable for users and administrators.

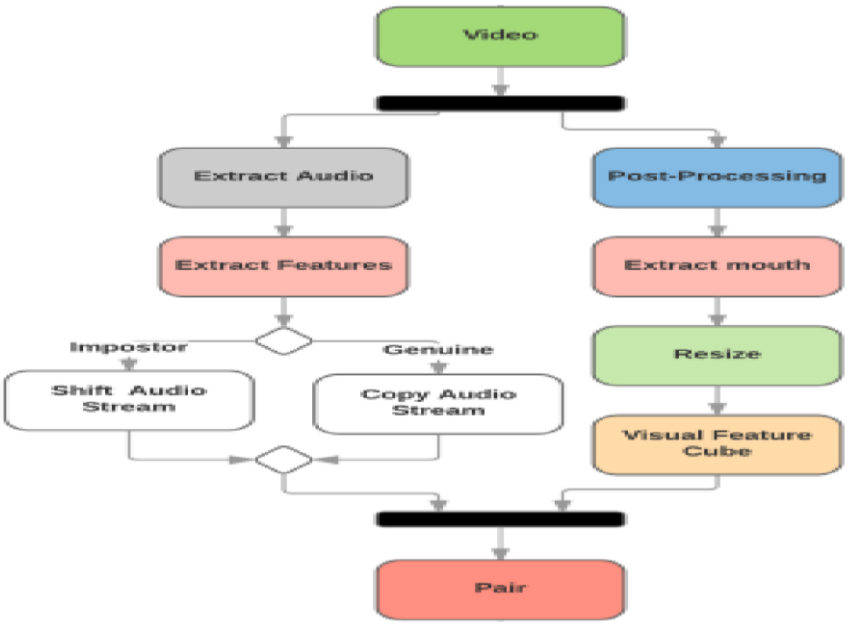
11. **Portability:**
    - The code should be portable and not rely on specific hardware or software dependencies beyond the required Python libraries.
    - It should be deployable across different environments without major modifications.

12. **Version Control and Configuration Management:**
    - The code should be managed using version control systems (e.g., Git) to track changes and facilitate collaboration

CHAPTER 6

DESIGN ANALYSIS

6. DESIGN & ANALYSIS



The proposed architecture utilizes two non-identical Convents which uses a pair of speech and video streams. The network 0.3 second of a video clip. The main task is to determine if a stream of audio corresponds with a lip motion clip within the desired stream duration. In the two next sub-sections, we are going to explain the inputs for speech and visual streams.

The architecture is a **coupled 3D convolutional neural network** in which *two different networks with different sets of weights must be trained*. For the visual network, the lip motions spatial information alongside the temporal information are incorporated jointly and will be fused for exploiting the temporal correlation. For the audio network, the extracted energy features are considered as a spatial dimension, and the stacked audio frames form the temporal dimension.

The speech features have been extracted using [SpeechPy] package. The frame rate of each video clip used in this effort is 30 f/s. Consequently, 9 successive image frames form the 0.3 second visual stream. The input of the visual stream of the network is a cube of size 9x60x100, where 9 is the number of frames that represent the temporal information. Each channel is a 60x100 gray-scale image of mouth region.

CHAPTER 7

IMPLEMENTATION

7. IMPLEMENTATION

Implementation is the stage where the theoretical design is turned into a working system. The most crucial stage in achieving a new successful system and in giving confidence on the new system for the users that it will work efficiently and effectively.

The system can be implemented only after thorough testing is done and if it is found to work according to the specification. It involves careful planning, investigation of the current system and it constraints on implementation, design of methods to achieve the change over and an evaluation of change over methods a part from planning.

Two major tasks of preparing the implementation are education and training of the users and testing of the system. The more complex the system being implemented, the more involved will be the system analysis and design effort required just for implementation.

The implementation phase comprises of several activities. The required hardware and software acquisition is carried out. The system may require some software to be developed. For this, programs are written and tested. The user then changes over to his new fully tested system and the old system is discontinued.

TESTING

The testing phase is an important part of software development. It is the Information zed system will help in automate process of finding errors and missing operations and also a complete verification to determine whether the objectives are met and the user requirements are satisfied. Software testing is carried out in three steps:

1. The first includes unit testing, where in each module is tested to provide its correctness, validity and also determine any missing operations and to verify whether theobjectives have been met. Errors are noted down and corrected immediately.
2. Unit testing is the important and major part of the project. So errors are rectified easily in particular module and program clarity is increased. In this project entire system is dividedinto several modules and is developed individually. So unit testing is conducted to individual modules.
3. The second step includes Integration testing. It need not be the case, the software whose modules when run individually and showing perfect results, will also show perfect results when run as a whole.

CHAPTER 8

SNAPSHOTS

8. SNAPSHOTS

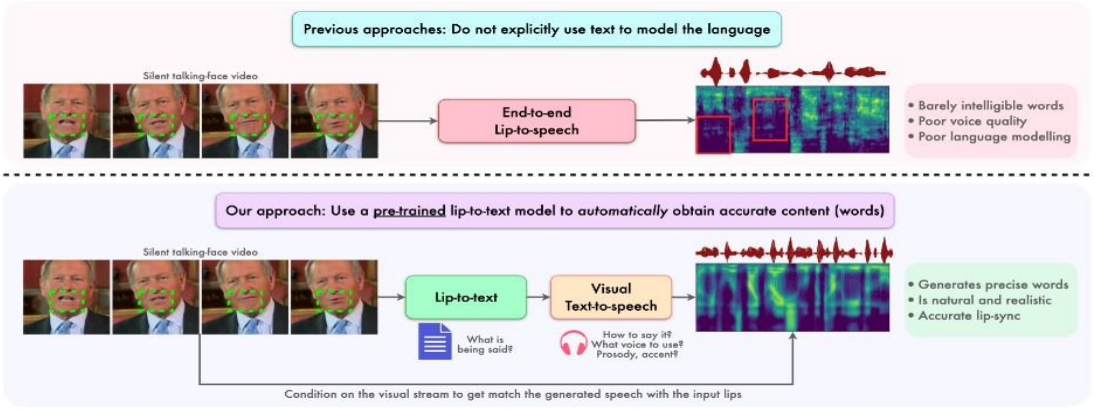
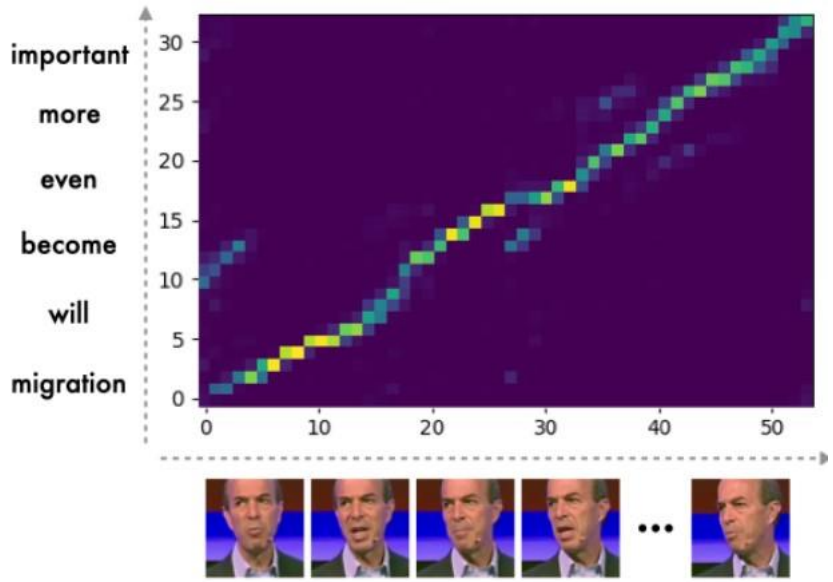


Table 1 contains the results on the TCD-TIMIT dataset. We observe that our approach achieves comparable results to previous methods

Dataset	Method	PESQ ↑	STOI ↑	ESTOI ↑	LSE-C ↑	LSE-D ↓
TCD-TIMIT (Harte and Gillen, 2015)	GAN-based (Vougioukas et al., 2019)	1.22	0.51	0.32	-	-
	Lip2Wav (Prajwal et al., 2020a)	1.35	0.56	0.36	6.610	7.915
	VCA-GAN (Kim et al., 2022)	1.43	0.58	0.40	-	-
	VAE-GAN (Hegde et al., 2022)	1.35	0.55	0.35	-	-
	Ours	1.34	**0.61**	**0.42**	**6.623**	**6.901**
LRW (Chung and Zisserman, 2016a)	GAN-based (Vougioukas et al., 2019)	0.72	0.10	0.02	1.983	9.426
	Lip2Wav (Prajwal et al., 2020a)	1.19	0.54	0.34	2.526	8.286
	VAE-GAN (Hegde et al., 2022)	0.78	0.15	0.03	2.538	8.173
	VCA-GAN (Kim et al., 2022)	1.33	0.56	0.36	-	-
	SVTS (Mira et al., 2022)	1.49	0.64	0.48	-	-
	Multi-task L2S (Kim et al., 2023)	1.56	0.64	0.47	4.876	8.102
	Lip-to-Text + TTS baseline	0.69	0.10	0.01	1.993	12.872
	Ours	**1.61**	**0.71**	**0.56**	**6.812**	**6.974**
LRS2 (Chung et al., 2017)	Lip2Wav (Prajwal et al., 2020a)	0.58	0.28	0.11	1.874	11.48
	VAE-GAN (Vougioukas et al., 2019)	0.60	0.34	0.17	2.507	8.155
	VCA-GAN (Kim et al., 2022)	1.24	0.40	0.13	4.016	7.914
	SVTS (Mira et al., 2022)	1.34	0.49	0.29	-	-
	Multi-task L2S (Kim et al., 2023)	1.36	0.52	0.34	4.001	8.192

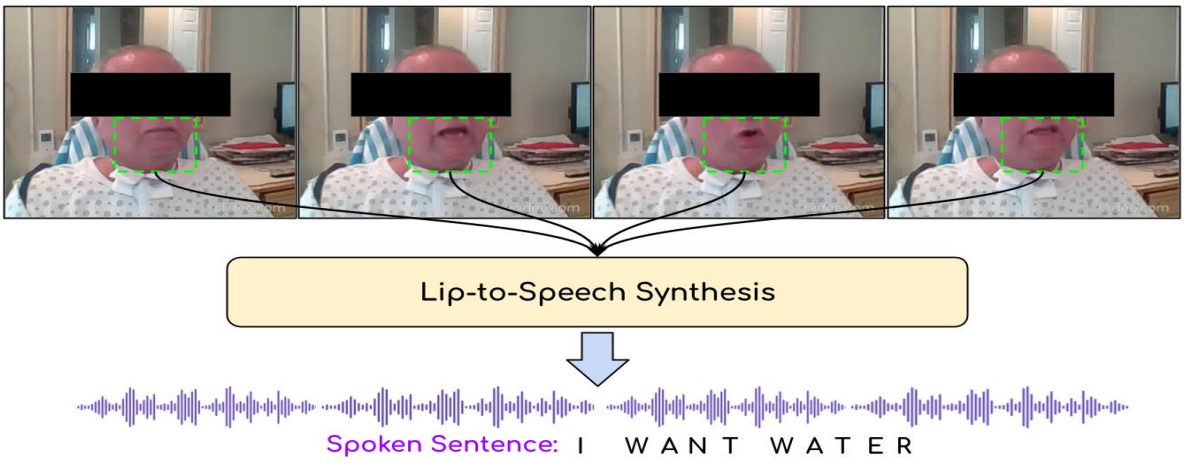
we outperform the previous methods in all the speech quality metrics, indicating the robustness and superiority of our approach. we depict how our model temporally aligns video and text sequences in the process of generating speech



(A) Intelligibility, (B) Content clarity, (C) Sync Accuracy, (D) Overall perceptual quality. Our model produces natural and realistic speech outputs that is largely preferred by the users in comparison to other approaches.

Method	(A)	(B)	(C)	(D)
GAN-based (Vougioukas et al., 2019)	2.05	1.87	1.99	2.12
Lip2Wav (Prajwal et al., 2020a)	1.01	1.03	1.34	1.01
VAE-GAN (Hegde et al., 2022)	1.07	1.33	2.18	2.57
VCA-GAN (Kim et al., 2022)	2.18	1.88	2.97	2.54
Multi-task L2S (Kim et al., 2023)	2.19	1.85	3.01	2.64
Lip-to-Text + TTS baseline	**3.61**	2.87	1.01	2.96
Ours	3.49	**3.52**	**3.82**	**3.31**

Further, we also test our model on the other deaf speakers studied in (Sen et al., 2021) and observe accurate performance.



Comparison of using generated text from different lip-to-text network in our pipeline. We also report the WER of the lip reading model (L2T-WER) on the LRS2 test set as a reference.

Method	L2T-WER	PESQ ↑	STOI ↑	ESTOI ↑	LSE-C ↑	LSE-D ↓
Deep Lip Reading (Afouras et al., 2018a)	51.3	1.17	0.40	0.22	7.847	6.904
AV-HuBERT (Shi et al., 2022)	46.1	1.27	0.53	0.40	7.960	7.003
VTP (Ours)	22.6	1.47	0.65	0.47	8.083	6.586
GT text	-	1.51	0.69	0.50	8.781	6.106

We present the effect of using different visual representations for training the Visual TTS module

Method	PESQ ↑	STOI ↑	ESTOI ↑	LSE-C ↑	LSE-D ↓
Face crops	1.17	0.40	0.22	7.847	6.904
VTP (Ours)	1.47	0.65	0.47	8.083	6.586

Code Implementation

The input pipeline must be provided by the user. The rest of the implementation consider the dataset which contains the utterance-based extracted features. **Lip Tracking**
For lip tracking, the desired video must be fed as the input. At first, cd to the corresponding directory:

```
python VisualizeLip.py --input input_video_file_name.ext --output
output_video_file_name.ext
```

The run the dedicated python file as below:

Running the aforementioned script extracts the lip motions by saving the mouth area of each frame and creates the output video with a rectangular around the mouth area for better visualization.

The required arguments are defined by the following python script which have been defined in the VisualizeLip.py file:

```
ap = argparse.ArgumentParser()
ap.add_argument("-i", "--input",
```

```
required=True, help="path to input video
file")
ap.add_argument("-o", "--output", required=True,
            help="path to output video file")
ap.add_argument("-f", "--fps", type=int, default=30,
            help="FPS of output video")
ap.add_argument("-c", "--codec", type=str, default="MJPG",
            help="codec of output video")
```

Some of the defined arguments have their default values and no further action is required by them.

Training / Evaluation

At first, clone the repository. Then, cd to the dedicated directory:

```
cd code/training_evaluation
```

```
python train.py
```

Finally, the train.py file must be executed:

For evaluation phase, a similar script must be executed:

```
python test.py
```

CHAPTER 9

CONCLUSION

CONCLUSION

The package was designed in such a way that future modifications can be done easily. The following conclusions can be deduced from the development of the project:

- Automation of the entire system improves the efficiency
- It provides a friendly graphical user interface which proves to be better when compared to the existing system.
- It gives appropriate access to the authorized users depending on their permissions.
- It effectively overcomes the delay in communications.
- Updating of information becomes so easier
- System security, data security and reliability are the striking features.
- The System has adequate scope for modification in future if it is necessary.

9. REFERENCE

Ariel Ephrat and Shmuel Peleg. Vid2speech: speech reconstruction from silent video. In 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) [2] Ariel Ephrat, Tavi Halperin, and Shmuel Peleg. Improved speech reconstruction from silent video. In Proceedings of the IEEE International Conference on Computer Vision Workshops,

[3] Hassan Akbari, Himani Arora, Liangliang Cao, and Nima Mesgarani. Lip2audspec: Speech reconstruction from silent lip movements video. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)

[4] KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar. Learning individual speaking styles for accurate lip to speech synthesis. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition

[5] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks.

[6] Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder-decoder approaches.

[7] Martin Cooke, Jon Barker, Stuart Cunningham, and Xu Shao. An audio-visual corpus for speech perception and automatic speech recognition.

[8] Naomi Harte and Eoin Gillen. Tcd-timit: An audio-visual corpus of continuous speech. [9] Joon Son Chung and Andrew Zisserman. Lip reading in the wild. In Asian Conference on Computer Vision

[10] Konstantinos Vougioukas, Pingchuan Ma, Stavros Petridis, and Maja Pantic. Videodriven speech recon-struction using generative adversarial networks.

[11] Ravindra Yadav, Ashish Sardana, Vinay P Namboodiri, and Rajesh M Hegde. Speech prediction in silentvideos using variational autoencoders.

[12] Daniel Michelsanti, Olga Slizovskaia, Gloria Haro, Emilia Gómez, Zheng-Hua Tan, and Jesper Jensen. Vocoder-based speech synthesis from silent videos. 34 | P a g e [13] Masanori Morise, Fumiya Yokomori, and Kenji Ozawa. World: a vocoder-based hig