VISVESVARAYA TECHNOLOGICAL UNIVERSITY

BELAGAVI–590018, Karnataka

INTERNSHIP REPORT

ON

"Lip to speech synthesis"

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

BACHELOR OF ENGINEERING IN COMPUTER SCIENCE

Submitted by:

NAME: Xavier Emmanuel Dias

USN: 2JR19CS098

Conducted at
Varcons Technologies Pvt Ltd

Jain College Of Engineering And Research

Department of computer science

Accredited by NBA, New Delhi

Udyambag, Angol, Belagavi, Karnataka 590008

Jain College of Engineering And Research

Department of computer science

Accredited by NBA, New Delhi

Udyambag, Angol, Belagavi, Karnataka 590008

CERTIFICATE

This is to certify that the Internship titled **"**Lip to speech synthesis**"** carried out by **Mr.** Xavier Dias (2JR19CS098)**,** a bonafide student of Jain College of Engineering And Research Belgaum, in partial fulfillment for the award of **Bachelor of Engineering, in Computer Science** under Visvesvaraya Technological University, Belagavi, during the year 2022-2023. It is certified that all corrections/suggestions indicated have been in corporate 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 with Date

1)_____
_____

2)_____
_____

DECLARATION

I, Mr. Xavier Emmanuel Dias, final year student of Jain College of Engineering And Research Belgaum - 590008, declare that the Internship has been successfully completed, **Varcons Technologies Pvt Ltd**, This report is submitted in partial fulfillment of the requirements for award of Bachelor Degree in Branch name, during the academic year 2022-2023.

Date: 25/08/2022

Place: Belgaum

USN: 2JR19CS098

NAME: Xavier E Dias

ACKNOWLEDGEMENT

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 us adequate facilities to undertake this Internship.

We would like to thank our Head of Computer Science Department, 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 department, for helping 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: Xavier Emmanuel Dias

USN: 2JR19CS098

OFFER LETTER

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 audio-visual 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 Varcons Technologies Private Limited

Varcons Technologies Private Limited is a Private incorporated on 11 July 2022. It is classified as Non-govt Company and is registered at Registrar of Companies, Bangalore. Its authorized share capital is Rs. 1,000,000 and its paid up capital is Rs. 10,000. It is inolved in Other computer related activities [for example maintenance of websites of other firms/ creation of multimedia presentations for other firms etc.

Directors of Varcons Technologies Private Limited are Chikaegowdanadoddi Kariyappa Somalatha and Haralahalli Chandraiah Spoorthi

Varcons Technologies Private Limited's Corporate Identification Number is (CIN) U72900KA2022PTC163646 and its registration number is 163646.Its Email address is ca.mittalankushjain@gmail.com and its registered address is #8/9, 5th Main, 3rd Cross road, Beside Sachidananada Nagar, R R Nagar Bangalore Bangalore KA 560098

CHAPTER 2

ABOUT THE COMPANY

2. ABOUT THE COMPANY

Varcons Technologies is a leading provider of cutting-edge technologies and services, offering scalable solutions for businesses of all sizes. Founded by a group of friends who started by scribbling their ideas on a piece of paper, today we offer smart, innovative services to dozens of clients. We develop SaaS products, provide Corporate Seminars, Industrial trainings and much more

Built for Creatives, by Creatives

At VCT, We make sure every product/service that we offer is built keeping in mind the practical usability of the product/Service, We're a startup focused on Creativity and Customizability, and We also provide subscription models for Software that we have already built, Since the application is already configured, the user has a ready-to-use application. This not only reduces installation and configuration time but also cuts down the time wasted on potential glitches linked to software deployment

Services provided by Varcons Technologies.

VCA provides a host of services to its customers/users/clients, Enabling business success driven by technology Harnessing the power of technology, we create a measurable difference for our clients across various industries & multiple geographies.

- Website as Software

  We develop websites that behave and interact similar to sophisticated software. Search Engine Optimisation We help you manage your SEO campaign more efficiently and effectively. We help you gain market share by leveraging our expertise. our holistic approach to identify anything that may be hurting your traffic or rankings and show you just how to outrank the competition.

- Comprehensive Customer Support

  With a comprehensive range of services, we guarantee your technology needs are not just met, but exceeded. We shall work with your customers/users closely to understand the way your users/customers use/make use of products/services Branding and Design We offer professional Graphic design, Brochure design & Logo design.

- Analytics and Research

  We analyse the way your users/customers interact with you/your business by gathering, studying and understanding the consumer voice and their perception of the product/service

- Embedded Systems and IOT

  We work with Consumer Electronics, Lighting, Home Automation, Metering, Sensor-Technology, Home Appliance and Medical Device companies to help them create smart and connected products. Through its integrated Embedded and IoT services, VCA helps build intelligent & connected devices that can be remotely monitored and controlled while leveraging edge and cloud computing for a host of intelligent applications and analytics.

CHAPTER 3

INTRODUCTION

3. INTRODUCTION

Introduction to ML

Machine learning (ML) is a branch of artificial intelligence (AI) that enables computers to "self-learn" from training data and improve over time, without being explicitly programmed. Machine learning algorithms are able to detect patterns in data and learn from them, in order to make their own predictions. In short, machine learning algorithms and models learn through experience. In traditional programming, a computer engineer writes a series of directions that instruct a computer how to transform input data into a desired output. Instructions are mostly based on an IF-THEN structure: when certain conditions are met, the program executes a specific action.

Machine learning, on the other hand, is an automated process that enables machines to solve problems with little or no human input, and take actions based on past observations. While artificial intelligence and machine learning are often used interchangeably, they are two different concepts. AI is the broader concept – machines making decisions, learning new skills, and solving problems in a similar way to humans – whereas machine learning is a subset of AI that enables intelligent systems to autonomously learn new things from data.

Supervised learning, as the name indicates, has the presence of a supervisor as a teacher. Basically supervised learning is when we teach or train the machine using data that is well labelled. Which means some data is already tagged with the correct answer. After that, the machine is provided with a new set of examples (data) so that the supervised learning algorithm analyses the training data (set of training examples) and produces a correct outcome from labelled data.

Unsupervised learning is the training of a machine using information that is neither classified nor labeled and allowing the algorithm to act on that information without guidance. Here the task of the machine is to group unsorted information according to similarities, patterns, and differences without any prior training of data.

Problem Statement

The crucial part of an AVR algorithm is the feature selection for both audio and visual modalities, which has a direct impact on the performance of the audio-visual recognition task. Regarding the speech modality, most speech recognition systems employ Hidden Markov Models (HMMs) to extract the temporal information of speech and Gaussian Mixture Models (GMMs) to discriminate between different HMMs states for acoustic input representation Similar approaches have been employed in the analysis of multi-model voice and face data, which resulted in an improvement of speech recognition performance. The inference based on common sense is that the lip motions and the heard voice which is represented by speech features are highly correlated as a human is usually able to match the heard sound to a given set of lip motion. However, the visual lip motions and their corresponding audio stream still can have non-negligible uncorrelated information.

The main problem is to recognize whether the visual lip motions of a speaker corresponds to the accompanying speech signal. The aforementioned root problem is the precedent to audiovisual synchrony verification, as recognizing the consistency between the audio-visual streams is desired. The problem of audio-visual synchrony recognition has been addressed in different research efforts such as for identity verification, and liveness recognition of the audio-visual streams. To address the problem, we propose to use the 3D Convolutional Neural Networks models that have recently been employed for action recognition, scene understanding, and speaker verification and demonstrated promising results. 3D CNNs concurrently extract features from both spatial and temporal dimensions, so the motion information is captured and concatenated in adjacent frames. We use 3D CNNs to generate separate channels of information from the input frames. The combination of all channels of correlated information creates the final feature representation. The focus of the research effort described in this paper is to implement two non-identical 3D-CNNs for audio-visual matching (Section V). The goal is to design nonlinear mappings that learn a non-linear embedding space between the corresponding

Audio-video streams using a simple distance metric. This architecture can be learned by evaluating pairs of audio-video data and later used for distinguishing between pairs of matched and non-matched audio-visual streams. One of the main advantages of our audio-visual model is the noise-robust audio features, which are extracted from speech features with a locality characteristic (Section IV), and the visual features, which are extracted from spatial and temporal information of lip motions. Both audio-visual features are extracted using 3D CNNs, allowing the temporal information to be treated separately for better decision making.

CHAPTER 4

SYSTEM ANALYSIS

4. SYSTEMANALYSIS

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 large-scale 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.

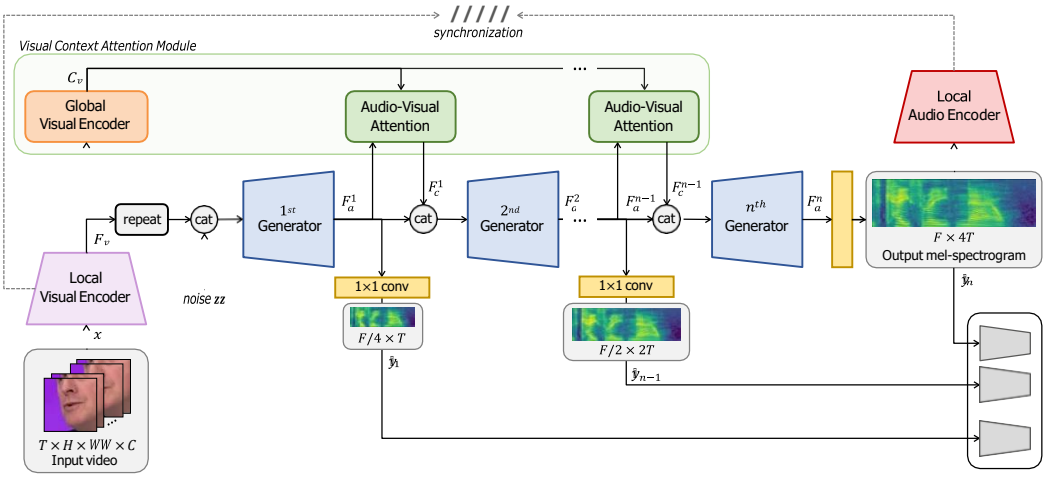


Figure: Overview of the VCA-GAN.

CHAPTER 5

REQUIREMENT ANALYSIS

5. REQUIREMENT ANALYSIS

Specific requirements

A Software requirements definition is an abstract description of the service which the system should provide, and the constraints and which system must operate. It should only specify the external behavior of the system.

User Requirement

- Easy to understand and should be simple.
- The built-in functions should be utilized to maximum extent.

Hardware Constraints

- Processor: Intel
- RAM: 512MB
- Hard Disk: 20GB(approx)

Software Constraints

- OperatingSystem:Windows10/2000/XP/Vista/UBUNTU
- Language: XML,JAVA
- Compiler: Android studio

CHAPTER 6

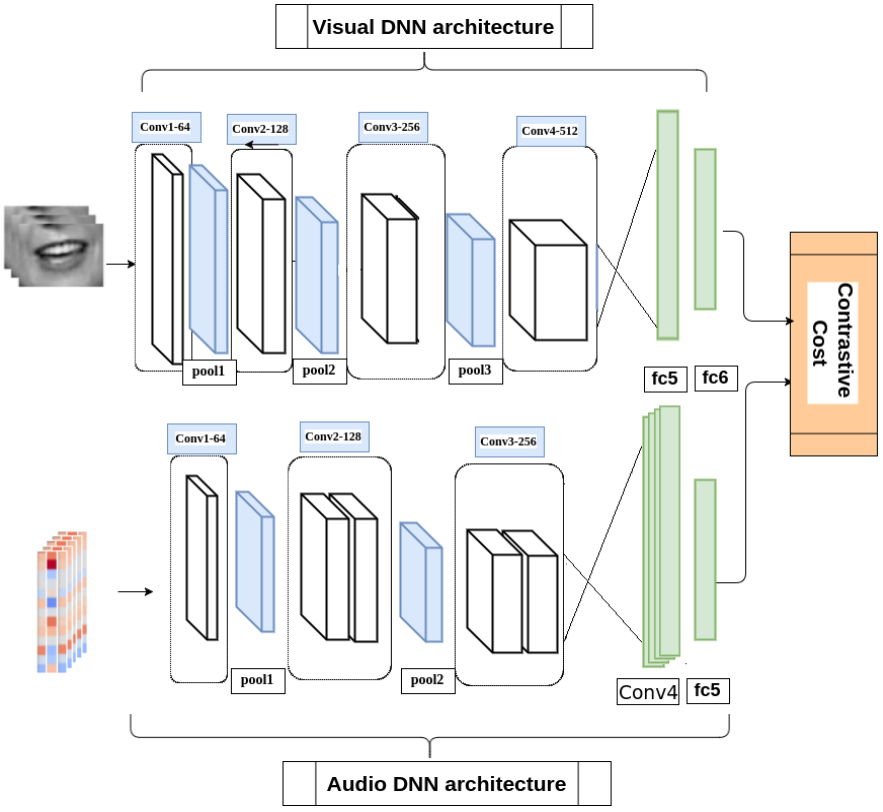
DESIGN ANALYSIS

6. DESIGN&ANALYSIS

input is a pair of features that represent lip movement and speech features extracted from In the visual section, the videos are post-processed to have an equal frame rate of 30 f/s. Then, face tracking and mouth area extraction are performed on the videos using the dlib library. Finally, all mouth areas are resized to have the same size and concatenated to form the input feature cube. The dataset does not contain any audio files. The audio files are extracted from videos using FFmpeg frame work .

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. In the proposed 3D CNN architecture, the convolutional operations are performed on successive temporal frames for both audio-visual streams.



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 apart from planning.

Two major tasks of preparing the implementation are education and training of the user's 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 the objectives 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 divided into 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.

Please run the **run.sh**file as a demo to the project. It can be run as below:

```
./run.sh
```

The following command line script, only execute the lip tracking operation:

```
./run.sh lip_tracking path/to/file.mp4
```

Basically, the path/to/file.extis the reletive path to the input video file. Example is data/sample_video.mp4.

General View

Audio-visual recognition (AVR) has been considered as a solution for speech recognition tasks when the audio is corrupted, as well as a visual recognition method used for speaker verification in multi-speaker scenarios. The approach of AVR systems is to leverage the extracted information from one modality to improve the recognition ability of the other modality by complementing the missing information.

The Problem and the Approach

The essential problem is to find the correspondence between the audio and visual streams, which is thegoal of this work. We proposed the utilization of a coupled 3D Convolution Neural Network (CNN) architecture that can map both modalities into a representation space to evaluate the correspondence of audio-visual streams using the learned multimodal features.

How to leverage 3D Convolution Neural Networks?

The proposed architecture will incorporate both spatial and temporal information jointly to effectively find the correlation between temporal information for different modalities. By using a relatively small network architecture and much smaller dataset, our proposed method surpasses the performance of the existing similar methods for audio-visual matching which use CNNs for feature representation. We also demonstrate that effective pair selection method can significantly increase the performance.

Code Implementation

The input pipeline must be provided by the user. The rest of the implementation consider the datasetwhich 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:

```
cd code/lip_tracking
```

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

The run the dedicated python fileas 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 argumentsare defined by the following python script which have been defined in the

VisualizeLip.pyfile:

```
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
```

Finally, the train.pyfile must be executed:

```
python train.py
```

For evaluation phase, a similar script must be executed:

```
python test.py
```

CHAPTER 8

SNAPSHOTS

8. SNAPSHOTS

Fig: Given input Obama audio and a reference video, we synthesize photorealistic, lip-synced video of Obama speaking those words.

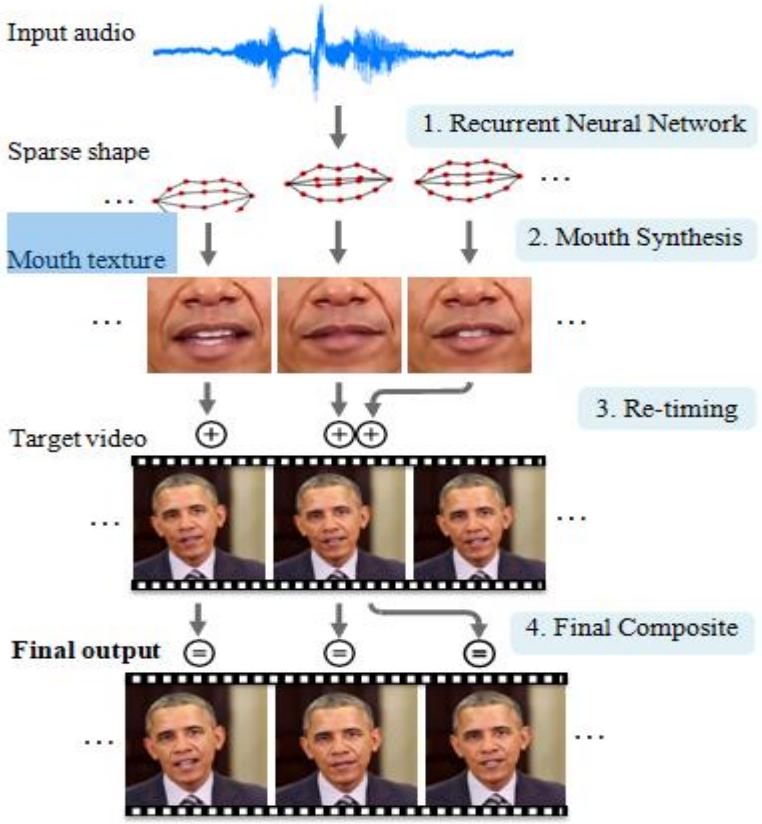


Fig: Our system first converts audio input to a time-varying sparse mouth shape. Based on this mouth shape, we generate photo-realistic mouth texture, that is composited into the mouth region of a target video. Before the final composite, the mouth texture sequence and the target video are matched and re-timed so that the head motion appears natural and fits the input speech.

Original Video for Input Audio

A) Our result

Original Video for Input Audio

B) Our result

Fig: Comparison of our mouth shapes to the ground-truth footage of the input audio A) Is a weekly address on climate change and B) is on health care

CHAPTER 9

CONCLUTION

9. CONCLUTION

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 over comes 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.

10. REFERENCE

[1] 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. Video-driven 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.

[13] Masanori Morise, Fumiya Yokomori, and Kenji Ozawa. World: a vocoder-based high-quality speech syn-thesis system for real-time applications. *IEICE TRANSACTIONS on Information and Systems*

[14] Triantafyllos Afouras, Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. Deep audio-visual speech recognition. *IEEE transactions on pattern analysis and machine intelligence*

[15] Chenhao Wang. Multi-grained spatio-temporal modeling for lip-reading.

[16] Jingyun Xiao, Shuang Yang, Yuanhang Zhang, Shiguang Shan, and Xilin Chen. Deformation flow basedtwo-stream network for lip reading.

[17] Yuanhang Zhang, Shuang Yang, Jingyun Xiao, Shiguang Shan, and Xilin Chen. Can we read speech beyond the lips? rethinking roi selection for deep visual speech recognition.

[18] Minsu Kim, Joanna Hong, Se Jin Park, and Yong Man Ro. Multi-modality associative bridging through memory: Speech sound recollected from face video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*

[19] Themos Stafylakis and Georgios Tzimiropoulos. Combining residual networks with lstms for lipreading.