VISVESVARAYA TECHNOLOGICAL UNIVERSITY

BELAGAVI – 590018, Karnataka

INTERNSHIP REPORT

ON

"Lip to speech synthesis"

Submitted in partial fulfilment for the award of degree

BACHELOR OF ENGINEERING IN

ELECTRICAL AND ELECTRONIC

Submitted by:

NAME:UMA BAI USN: 1RN20EE046

Varcons Technologies Pvt Ltd

Conducted at

VARCONS TECHNOLOGY PVT LTD

RNS INSTITUTE OF TECHNOLOGY

DEPARTMENT OF ELECTRICAL AND ELECTRONIC Accredited by NAAC, Bangalore, Karnataka 560010

RNS INSTITUTE OF TECHNOLOGY

DEPARTMENT OF ELECTRICAL AND ELECTRONIC ENGINEERING Accredited by NAAC, Bangalore, Karnataka 560010

CERTIFICATE

This is to certify that the Internship titled "**Lip to speech synthesis**" carried out by **Miss UMA BAI,** a Bonafide student of RNS Institute of Technology, in partial fulfilment for the award of **Bachelor of Engineering**, in ELECTRICAL AND ELECTRONIC under Visvesvaraya Technological University, Belagavi, during the year 20212022. 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

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

External Viva:

Name of the Examiner Signature with Date

1) _____

2) _____

D E C L A R A T I O N

I, **UMA BAI,** third year student of computer science Branch, RNS Institute of Technology Bangalore- 560010, declare that the Internship has been successfully completed, in VARCONS TECHNOLOGIES PVT. LTD. This report is submitted in partial fulfilment of the requirements for award of bachelor's degree in Computer science, during the academic year 2021-2022.

Date :18-nov-2022

Place: Bangalore

USN: 1RN20EE046

NAME: UMA BAI

OFFER LETTER

INTERNSHIP OFFER LETTER

Date: **14th October, 2022**

Name: **Uma Bai**
USN: **1RN20EE046**

Dear Student,

We would like to congratulate you on being selected for the **Machine Learning With-Python(Research Based)** Internship position with **Varcons Technologies Pvt Ltd.** effective Start Date **14th October, 2022**. 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 PVT LTD
213, 3rd 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 us adequate 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 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, 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: UMA BAI

USN: 1RN20EE046

ABSTRACT

In this report, we propose a novel lip-to-speech generative adversarial network, Visual Context Attentional GAN (VCA-GAN), which can jointly model local and global lip movements during speech synthesis. Specifically, the proposed VCA-GAN synthesizes the speech from local lip visual features by finding a mapping function of viseme-to-phoneme, 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 modelling 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.

Humans involuntarily tend to infer parts of the conversation from lip movements when the speech is absent or corrupted by external noise. In this work, we explore the task of lip to speech synthesis, i.e., learning to generate natural speech given only the lip movements of a speaker. Acknowledging the importance of contextual and speaker-specific cues for accurate lip-reading, we take a different path from existing works. We focus on learning accurate lip sequences to speech mappings for individual speakers in unconstrained, large vocabulary settings. To this end, we collect and release a large-scale benchmark dataset, the first of its kind, specifically to train and evaluate the single-speaker lip to speech task in natural settings. We propose a novel approach with key design choices to achieve accurate, natural lip to speech synthesis in such unconstrained scenarios for the first time. Extensive evaluation using quantitative, qualitative metrics and human evaluation shows that our method is four times more intelligible than previous works in this space.

Table of Contents

CHAPTER 1

COMPANY PROFILE

A Brief History of Varcons Technologies

Varcons Technologies Private Limited is an unlisted private company incorporated on 11 July 2022. It is classified as a private limited company and is located in, Karnataka. Its authorized share capital is INR 10.00 lac, and the total paid-up capital is INR 10,000.00.

The current status of Varcons Technologies Private Limited is - Active.

Details of the last annual general meeting of Varcons Technologies Private Limited are not available. The company is yet to submit its first full-year financial statements to the registrar.

Varcons Technologies Private Limited has two directors – Chikaegowdanadoddi Kariyappa Somalatha and Haralahalli Chandraiah Spoorthi.

The Corporate Identification Number (CIN) of Varcons Technologies Private Limited is U72900KA2022PTC163646. The registered office of Varcons Technologies Private Limited is at #8/9, 5th Main, 3rd Cross road, Beside Sachidananada Nagar, R R Nagar Bangalore Bangalore , Karnataka.

2. ABOUT THE COMPANY

Communicate.Collaborate. Create

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.

- Speed & Security
- Flexibility & Scalability
- Better Collaboration

With the Right Software, Service and Analytics, Great Things Can Happen

Smart solutions are at the core of all that we do at VCT. Our main goal is to find smart ways of using technology that will help build a better tomorrow for everyone, everywhere. SaaS offers a variety of advantages over traditional software licensing models and We here at VCT tend to include the key features of SaaS in everything we build.

OD services we provide

Traditional Services+ SaaS features= Magic!

Website as Software

We develop websites that behave and interact similar to Sophisticated software. Information +Functionality=Waas

Analytics and Research

Let us 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 to generate a report to help you make better market decisions

Comprehensive Customer Support

With a comprehensive range of services, We can 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. **Smart Automation Tools**

We create API's and tools that help you automate any process with a host of features pertaining to the Device.

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.

- All-In-One, Toolkit
- Integrated, File Sharing
- Total Design, Freedom

3.INTRODUCTION

INTRODUTION 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.

PROBLEM STATEMENT:

Lip to Speech synthesis - Varcons Technologies Pvt Ltd

An already on-development model will be shared with you, which you can use to improve the accuracy and demonstrate the working of the newly developed model Src code will be provided

Src:
https://drive.google.com/drive/folders/15AiZhnChrrRgjnkIhNmI9lIjruCE93m?usp=sharing

Goal: Test the working of LTS and look for ways to improve the accuracy of the same Current accuracy rate: 64/100

1. INTRODUCTION

As the world's communication becomes increasingly digital, it is also becoming increasingly visual. From video calls to movies to YouTube videos, there is a surge in video content consumption. Naturally, understanding and enabling applications for talking-face videos [2, 3, 13, 20, 36, 37] has been an active area of research in recent years. Tasks such as speech/text based lip synthesis [27, 30, 37] have witnessed tremendous advancements. The opposites of these tasks, namely, lip-to-text generation and lip-to-speech generation, both falling under the umbrella of "lip-reading", have proven far more challenging. For the task of lip-to-text generation, multiple impressive works have pushed the boundaries with models that work for any speaker in the wild. However, its sibling task, lip-tospeech synthesis, has not yet witnessed a similar advancement in such unconstrained settings.

Lip-to-Speech Synthesis for Arbitrary Identities:

The goal of lip-to-speech synthesis is to generate meaningful speech for a silent talking-face video. Previous works in this space have focused on training models that work for a fixed set of speakers. They achieve impressive results but rely on videos recorded in laboratory settings [17, 23] or require tens of hours of single-speaker data [36] when working with real-world videos. This makes the previous methods hard to scale to the large number of identities in the wild. Our goal in this work is to perform lip-to-speech synthesis for silent videos of any identity. This allows us to produce results on any speaker at test time. We also show that we can further finetune on videos of a single speaker, if necessary, and achieve similar performance to single-speaker models but with 4× lesser data

CHAPTER 4. SYSTEM ANALYSIS

- **Related Work Lip to Speech Synthesis:** There have been a number of researches and interests in visual-to-speech generation. Ephrat et al.[1] utilized CNN to predict acoustic features from silent talking videos. Then, they augmented the model to two-tower CNN-based encoder-decoder [2] where each tower encodes raw frames and optical flows, respectively. Akbari et al.[3] employed a deep autoencoder for reconstructing the speech features from the visual features encoded by a lip-reading network. Vougioukas et al.[10] directly synthesized the raw waveform from the video by using 1D GAN. Prajwal et al.[4] focused on learning lip sequence to speech mapping for a single speaker in an unconstrained, large vocabulary setting using a stack of 3D convolution and Seq2Seq architecture. Yadav et al.[11] used stochastic modelling approach with variational autoencoder. Michelsanti et al.[12] predicted vocoder features of [13] and synthesized speech using the vocoder. Different from the previous works, our approach explicitly models the local visual feature and global visual context to synthesize accurate speech. Moreover, we try to synthesize the speech from multispeaker which has rarely been handled in the past. Visual Speech Recognition (VSR). Parallel to the development of Lip2Speech, Visual Speech Recognition (VSR) have achieved a great advancement [14, 15, 16, 17, 18, 19]. 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. Chung et al.[9] proposed a large-scale audio-visual dataset and set a baseline model for word-level VSR. Stafylakis et al.[20] proposed an architecture that is combined of residual network and LSTM, which became a popular architecture for word-level lip reading. Martinez et al . [21] replaced the RNN-based backend with Temporal Convolutional Network (TCN). Kim et al. [19, 22] proposed to utilize audio modal knowledge through memory network without audio inputs during inference for lip reading. Assael et al.[23] achieved end-to-end sentence-level lip reading network by adopting the CTC loss [24]. Different from the VSR methods, the Lip2Speech task does not require human annotations, thus is drawing big attention with its practical aspects.

- **Attention Mechanism.** The attention mechanism has affected many research fields, such as image captioning [25, 26, 27], machine translation [28, 29], and speech recognition [30, 31, 32]. It can effectively focus on relative information and reduce the interference from less significant one. There have been several works that incorporate attention mechanism in GAN. Xu et al.[25] proposed a cross modal attention model to guide the generator to focus on different words when generating different image subregions. Qiao et al.[27] further developed it by proposing a global-local collaborative attentive module to leverage both local word attention and global sentence attention and to enhance the diversity and semantic consistency of the generated images. Li et al.[26] introduced channelwise attention-driven generator that can disentangle different visual attributes, considering the most relevant channels in the visual features to be fully exploited. In this paper, we design a cross-modal attention module working with video and audio modalities for context modelling during speech synthesis. By bringing the global visual context through the proposed visual context attention module, speech of high intelligibility can be synthesized.

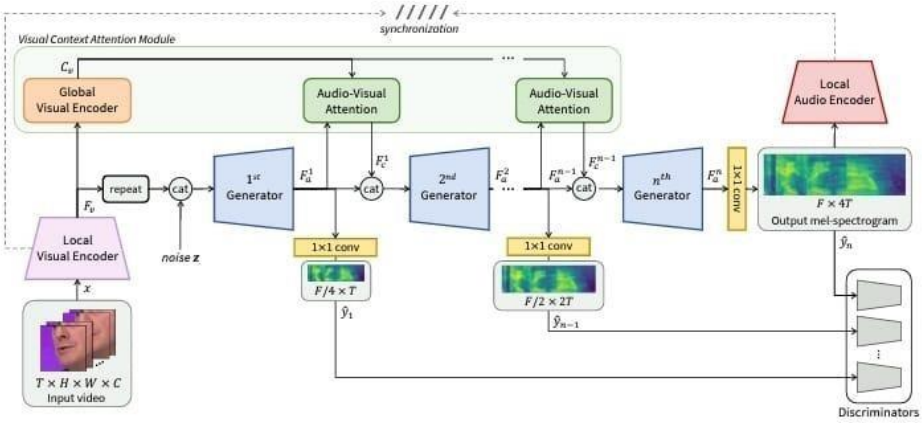

Figure 1: Overview of the VCA-GAN. Global visual context is provided through proposed visual context attention module to the generators to refine the speech representation from low- to high-resolution.

- **Visual context attentional GAN** Considering the entire context from the input lip movements, namely global visual context, can provide additional information that alleviates the ambiguity of homophenes besides the accurate temporal alignment of local visual representations. To achieve this, the generator synthesizes the speech from the local visual features while the global visual context is jointly considered at the intermediate layers of generator through the visual context attention module. Firstly, a local visual encoder φv encodes the video x into local visual features Fv = {f1 v,f2 v,⋯ ,fT v} ∈ RT× D, where D is the dimension of embedding. The local visual encoder φv is composed of combination of 3D and 2D convolutions. Due to the locality of the convolution operator, each local visual 3

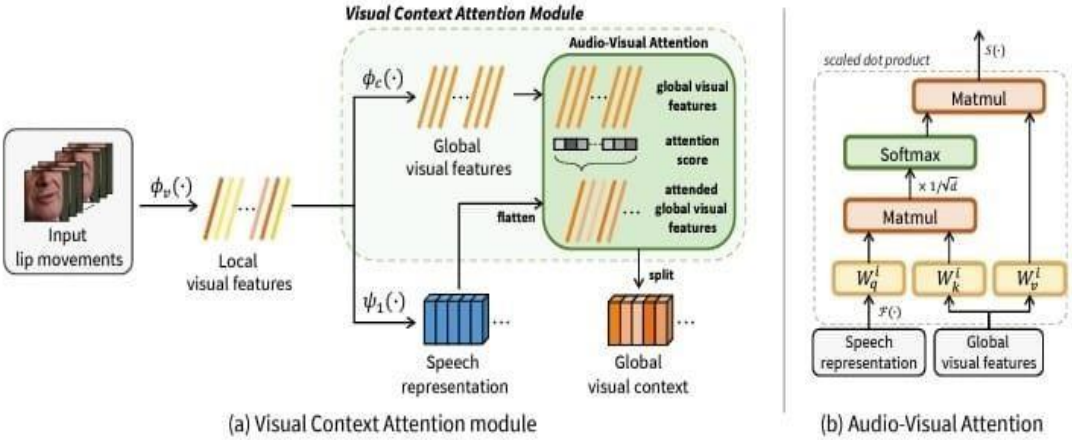

Figure 2: Illustration of visual context modelling through the proposed visual context attention module. (a) visual context module, and (b) audio-visual attention in detail.

CHAPTER 5

REQUIREMENT ANALYSIS

Dataset

GRID corpus [7] dataset is composed of sentences following fixed grammar from 33 speakers. We evaluate our model in three different settings. 1) constrained speaker setting subject of 1, 2, 4, and 29 are used for training and evaluation. We follow the dataset split of the prior works [4, 2, 10, 3, 12]. 2) unseen-speaker setting: 15, 8, and 10 subjects are used for training, validation, and test, respectively. The dataset split from [10, 12] is used. 3) multi-speaker setting: all 33 subjects are used both training and evaluation. For the dataset split, we follow the well-known protocol in VSR of [23]. TCD-TIMIT dataset [8] is composed of uttering videos from 3 lip speakers and 59 volunteers. Following [4], the data of 3 lip speakers are used for the evaluation in constrained speaker setting. LRW [9] is a word-level English audio-visual dataset derived from BBC news. It is composed of up to 1,000 training videos for each of 500 words. Since the dataset was collected from the television show, it has a large variety of speakers and poses, presenting challenges on speech synthesis.

CHAPTER 6

DESIGN ANALYSIS

For the visual encoder, one 3D convolution layer and ResNet-18 [48], a popular architecture in lip reading [49], are utilized. Three generators are used (i.e., n=3) and 2× up sample layer is applied at the last two generators. Each generator is composed of 6, 3, and 3 Residual blocks, respectively. The global visual encoder is designed with 2 layer bi-GRU and one linear layer. For the audio encoder, 2 convolution layers with stride 2 and one Residual block are utilized. The post net is composed of three 1D Residual blocks and two 1D convolution layers. Finally, the discriminators are basically composed of 2, 3, and 4 Residual blocks. Architectural details can be found in supplementary. All the audio in the dataset is resampled to 16kHz, high-pass filtered with a 55Hz cut off frequency and transformed into Mel-spectrogram using 80 Mel filter banks (i.e., F=80). For the dataset composed of 25 fps video (i.e., GRID and LRW), the audio is converted into Mel -spectrogram by using window size of 640 and hop size of 160. For the 30-fps video (i.e., TCD-TIMIT), the window size of 532 and hop size of 133 are used. Thus, the resulting me l-spectrogram has four times the frame rate of the video. The images are cropped to the centre of the lips and resized to the size of 112 × 112. During training, the contiguous sequence is randomly sampled with the size of 40 and 50 for GRID and TCDTIMIT, respectively. During inference, the network generates speech from arbitrary video frame length1. For the multi-scale ground-truth Mel -spectrograms (i.e., y1 and y2), bilinear interpolation is applied to the ground-truth Mel -spectrogram y (i.e., y3). We use Adam optimizer [50] with 0.0001 learning rate. The α, λ recon, and λ sync are empirically set to 2, 50, and 0.5, respectively. The temperature parameter τ is set to 1. For the GAN loss, nonsaturating adversarial loss [34] with R1 regularization [51] is used. Titan-RTX is utilized for the computing.

7 IMPLEMENTATION

- **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 its constraints on implementation, design of methods to achieve the changeover and an evaluation of change over methods apart 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:

• 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.

• 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.

• 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

11 lines (11 sloc) | 80 Bytes

```
*.pkl
*.pt
*.pyc
*.save
__pycache__
*.tar
.vc
.gitattributes
*.pth
.idea
.idea/
```

30 lines (30 sloc) | 433 Bytes

```
audioread==2.1.8
h5py==2.9.0
inflect
librosa==0.7.0
lws==1.2.6
Markdown==3.1.1
matplotlib==3.1.1
multiprocess
numba==0.48
numpy==1.16.4
opencv-python==4.1.1.26
pesq==0.0.1
PyQt5
pystoi==0.2.2
scikit-learn==0.21.2
scikit-image
scipy==1.3.0
sounddevice==0.3.13
SoundFile==0.10.2
tensorboard==1.13.1
tensorboardX==2.0
tensorflow-estimator==1.13.0
tensorflow-gpu==1.13.1
tqdm
umap-learn
Unidecode
visdom
webrtcvad
youtube-dl
torch==1.1.0
```

81 lines (62 sloc) | 2.27 KB

```python
from __future__ import print_function
import os
import torch
from torch.utils.model_zoo import load_url
from enum import Enum
import numpy as np
import cv2
try:
    import urllib.request as request_file
except BaseException:
    import urllib as request_file

from .models import FAN, ResNetDepth
from .utils import *


class LandmarksType(Enum):
    """Enum class defining the type of landm

    ``_2D`` - the detected points ``(x,y)``
    ``_2halfD`` - this points represent the
    ``_3D`` - detect the points ``(x,y,z)``

    """
    _2D = 1
    _2halfD = 2
    _3D = 3


class NetworkSize(Enum):
    # TINY = 1
    # SMALL = 2
    # MEDIUM = 3
    LARGE = 4

    def __new__(cls, value):
        member = object.__new__(cls)
        member._value_ = value
```

```python
ROOT = os.path.dirname(os.path.abspath(__fil

class FaceAlignment:
    def __init__(self, landmarks_type, netwo
                 device='cuda', flip_input=F
        self.device = device
        self.flip_input = flip_input
        self.landmarks_type = landmarks_type
        self.verbose = verbose

        network_size = int(network_size)

        if 'cuda' in device:
            torch.backends.cudnn.benchmark =

        # Get the face detector
        face_detector_module = __import__('f
                                          gl
        self.face_detector = face_detector_m

    def get_detections_for_batch(self, image
        images = images[..., ::-1]
        detected_faces = self.face_detector.
        results = []

        for i, d in enumerate(detected_faces
            if len(d) == 0:
                results.append(None)
                continue
            d = d[0]
            d = np.clip(d, 0, None)

            x1, y1, x2, y2 = map(int, d[:-1]

            inp = images[i][y1 : y2, x1 : x2
            results.append((inp, x1, y1, x2,

        return results
```

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.

• REFERENCE

- Lip to Speech Synthesis with Visual Context Attentional GAN Minsu Kim, Joanna Hong, Yong Man Ro∗ Image and Video Systems Lab KAIST

. https://github.com/80971/lip-to-speech-synthesis-#lip-to-speech-synthesis-

- Learning Individual Speaking Styles for Accurate Lip to Speech Synthesis KRPrajwal∗ IIIT, Hyderabad Rudrabha Mukhopadhyay∗ IIIT, Hyderabad Vinay P. Namboodiri IIT, Kanpur CV Jawahar IIIT, Hyderabad