VISVESVARAYA TECHNOLOGICAL UNIVERSITY

INTERNSHIP REPORT

ON

"Lip to Speech synthesis"

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

BACHELOR OF ENGINEERING IN
CSE BRANCH

Submitted by:

SRINIVAS G

1AY19CS109

Conducted at
VARCONS TECHNOLOGIES PVT LTD

ACHARYA INSTITUTE OF TECHNOLOGY

Department of CSE

Accredited by NBA, New Delhi

Acharya Dr Sarvepalli Radhakrishnan Rd, Soladevanahalli,
Karnataka 560107

ACHARYA INSTITUTE OF TECHNOLOGY
Department of CSE
Accredited by NBA, New Delhi

Acharya Dr Sarvepalli Radhakrishnan Rd, Soladevanahalli,
Karnataka 560107

CERTIFICATE

This is to certify that the Internship titled **"Lip to Speech synthesis"** carried out by **Mr SRINIVAS G,** a bonafide student of **Acharya Institute of Technology**, in partial fulfillment for the award of **Bachelor of Engineering**, in **CSE** under Visvesvaraya Technological University, Belagavi, during the year 2022-2023. 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 with Date

1)__________________

2)__________________

DECLARATION

I, **SRINIVAS G**, final year student of CSE, Acharya Institute of Technology - 560 107,declare that the Internship has been successfully completed, in **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.

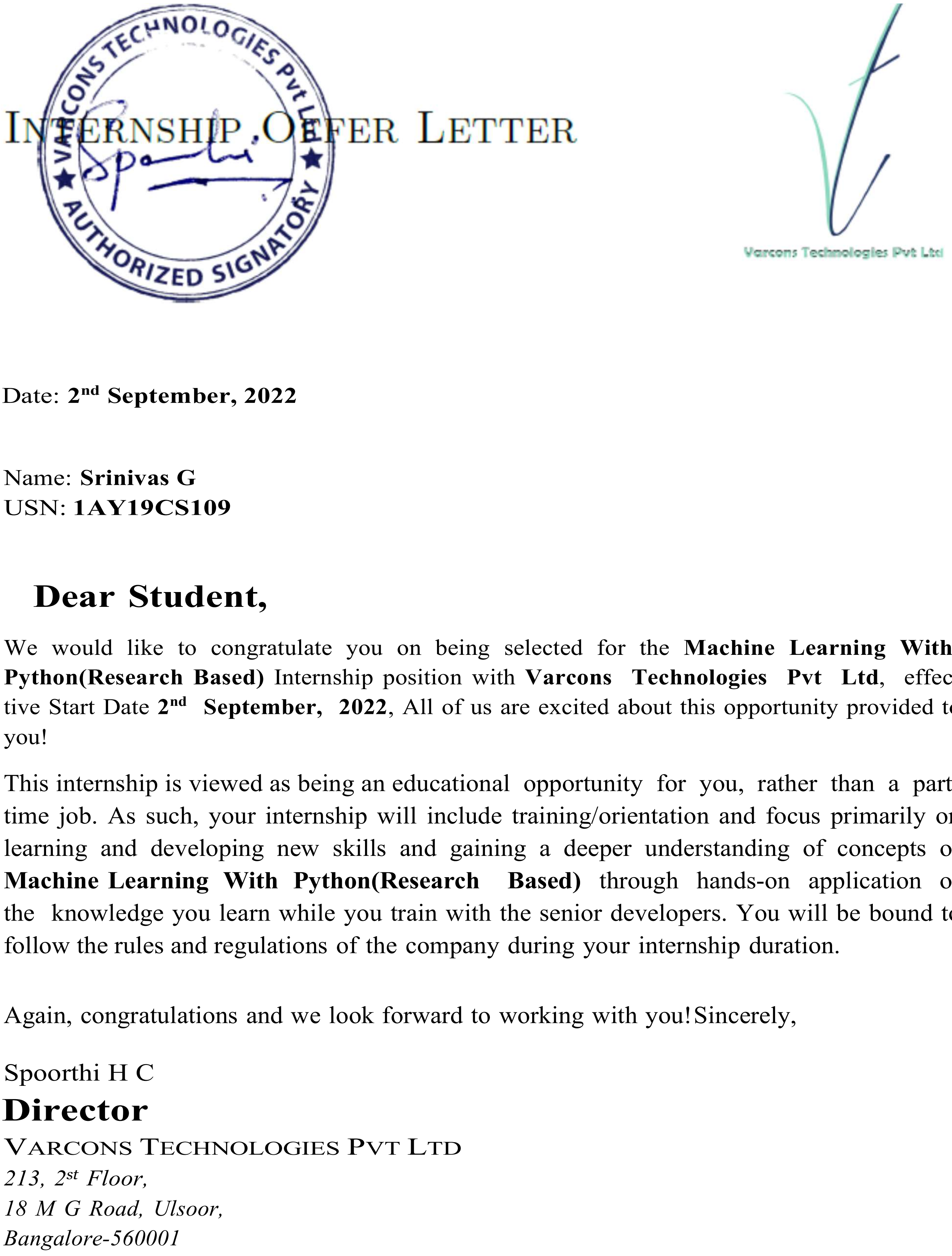
Date : 01/10/2022 :

Place : Bangaluru

USN : 1AY19CS109

NAME : SRINIVAS G

OFFER LETTER



VARCONS TECHNOLOGIES Pvt Ltd
AUTHORIZED SIGNATORY

INTERNSHIP OFFER LETTER

Varcons Technologies Pvt Ltd

Date: **2nd September, 2022**

Name: **Srinivas G**
USN: **1AY19CS109**

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 **2nd September, 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, 2st Floor,
18 M G Road, Ulsoor,
Bangalore-560001

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 **Dr Rajath hegade**, for providing usadequate facilities to undertake this Internship.
We would like to thank our Head of Dept – **Dr Ajith Padyana** , for providing us an opportunity to carry out Internship and for his valuable guidance and support.

We express our deep and profound gratitude to our guide, Mr Jovin Deglus, 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.

SRINIVAS G

(1AY19CS109)

ABSTRACT

In this project 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 VCAGAN 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 homophene. 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 stateof-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

A Brief History of Varcons Technologies

Varcons 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. Varcons 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 want to achieve.

CHAPTER 2
ABOUT THE COMPANY

Varcons Technologies Pvt Ltd

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, Varcons 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 technologicalinnovations 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 specifiic 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

Varcons 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 Varcons Technologies gives you the facility of skilled employees so that you do not feel unsecured aboutthe 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 Varcons 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 Varcons 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

Introduction to ML

Arthur Samuel, an early American Leader in the field of computer gaming and artificial intelligence,
Coined the term "Machine Learning" in 1959 while at IBM, He defined machine learning as "the field of study that gives computers the ability to learn without being explicitly programmed".

- Machine learning is programming computers to optimize a performance criterion using example data or past experience. We have a model defined up to some parameters, and learning is the execution of a computer program to optimize the parameters of the model using the training data or past experience. The model may be predictive to make predictions in the future, or descriptive to gain knowledge from data.
- The field of study known as machine learning is concerned with the question of how to construct computer programs that automatically improve with experience.

Types of Machine Learning

Machine learning algorithms differ in their methodology, the type of data they input and produce, and the objective or issue that they are designed to address. Machine Learning may be divided into four areas.

- Supervised Learning
- Unsupervised Learning
- Reinforcement Learning
- Semi-supervised Learning

Machine learning allows for the examination of vast amounts of data. While it typically delivers more quickly, precise outcomes Additional information may be required to discover profitable possibilities or risky threats, time and money to properly train it**.**

Supervised Learning

Supervised Learning is a sort of learning in which we are given a data set and already know what the proper output should be, with the assumption that there is a link between the input and the result. It is essentially a learning task that involves learning a function that maps an input to an output based on example input/output pairs. It derives a function from labeled training data, which consists of a collection of training samples.

Unsupervised Learning

Unsupervised Learning is a sort of learning that allows us to tackle issues with little or no prior knowledge of how the problem should appear. The structure may be derived by clustering the data based on a relationship between the variables in the data. There is no feedback based on prediction results in unsupervised learning. It is essentially a sort of selforganized learning that aids in the discovery of previously unknown patterns in data sets with no pre-existing label.

Reinforcement Learning

Reinforcement learning is a type of learning in which the learner interacts with its environment by performing actions and discovering faults or rewards. The most important aspects of reinforcement learning are trial and error search and delayed reward. This technology enables machines and software agents to automatically find the optimal behavior in a given scenario in order to optimize their performance. For the agent to understand which action is better, simple reward feedback is necessary.

PROBLEM STATEMENT

Humans involuntarily tend to infer parts of the conver-sation 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. We investigate the problem of silent lip videos to speech generation in large vocabulary, unconstrained settings for the first time

CHAPTER 4

SYSTEM ANALYSIS

1. Existing System:

The initial approaches to this problem extracted the visual features from sensors or active appearance models, the recent works employ end-to-end approaches.Vid2Speech and Lipper generate low-dimensional LPC (Linear Predictive Coding) features given a short sequence of K frames (K < 15). The facial frames are concatenated channel-wise and a 2D-CNN is used to generate the LPC features. We show that this architecture is very inadequate to model real-world talking face videos that contain significant head motion, silences and large vocabularies. Further, the low dimensional LPC features used in these works do not contain a significant amount of speech. both these works do not make use ofthe well-studied sequence-to-sequence paradigm that isused for text-to-speech generation thus leaving a largeroom for improvement in speech quality and correctness

Reply

Finally, all the above works show results primarily on the GRID corpus which has a very narrow vocabulary of 56 tokens and very minimal head motion. We are the first to study this problem in a large vocabulary setting with thousands of words and sentences. Our datasets are collected from YouTube video clips and hence contain a significant amount of natural speech variations and head movements. This makes our results directly relevant to several real-world applications.

2. Proposed System

Lip reading, also known as lipreading or speechreading, is a technique of understanding speech by visually interpreting the movements of the lips, face and tongue with information provided by the context, language, and any residual hearing. Each speech sound has a particular facial and mouth position (viseme), although many phonemes share the same viseme and thus are impossible to distinguish from visual information alone. Thus a speechreader must use cues from the environment and knowledge of what is likely to be said.

3. Objective of the System

In this study, we ask the human participants to manually identify and report (A) the percentage of mispronunciations,(B) the percentage of word skips and (C) the percentage of mispronunciations that are homophenes. Word skips denotes the number of words that are either completely unintelligible due to noise or slurry speech. We choose 10 predictions from the unseen test split of each speaker in our Lip2Wav dataset to get a total of 50 files. We report the mean numbers of (A), (B), and (C) in Table 5. Model (A) (B) (C) GAN-based [36] 36.6% 24.3% 63.8% Ephrat et al [9] 43.3% 27.5% 60.7% Lip2Wav (ours) 21.5% 8.6% 49.8% Table 5. Objective Human evaluation results. The participants manually identified the percentage of (A) Mispronunciations, (B) Word skips and (C) Homophene-based errors in the test samples. Our approach makes far fewer mispronunciations than the current state-of-the-art method. It also skips words 3× lesser, however, the key point to note is that the issue of homophenes is still a dominant cause for errors in all cases indicating there is still scope for improvement in this area..

CHAPTER 5
REQUIREMENT ANALYSIS

Hardware Requirement Specification

- Pentium 200-MHz computer with a minimum of 64 MB of RAM
- Monitor with a refresh rate of at least 40Hz for a smooth GUI experience (optional).
- Intel core i5
- Ram :8 GB
- Hard disk: 1TB

Software Requirement Specification

- Jupyter Notebook.
- Google Collab.
- Sklearn
- Google Chrome or Microsoft Edge of latest version.
- Windows 8 or Above.
- Webcam or video Tape.

CHAPTER 6
DESIGN ANALYSIS

Technology selection:

We used the following tools to implement the project

o Pandas

o Sklearn

o Matplotlib

Prior works in lip to speech regard their speech representation as a 2D-image in the case of melspectrograms or as a single feature vector in the case of LPC features.In both these cases, they use a 2D-CNN to decode these speech representations. By doing so, they violate the ordering in which they model the sequential speech data, i.e. the future time steps influence the prediction of the current time step. In contrast, we formulate this problem in the standard sequence-to-sequence learning paradigm.Concretely,each output speech time-step Sk is modelled as a conditional distribution of the previous speech time-steps S<k and the input face image sequence

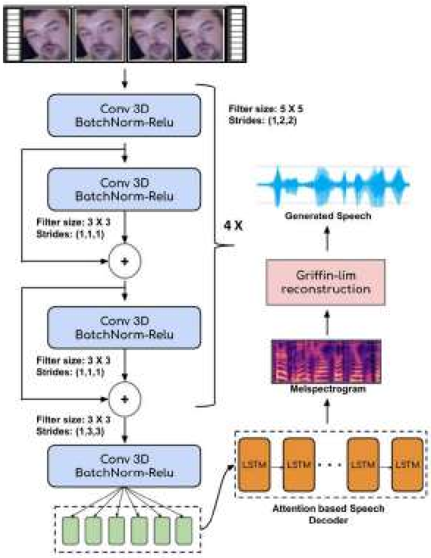
I = (I1, I2, . . . , IT).

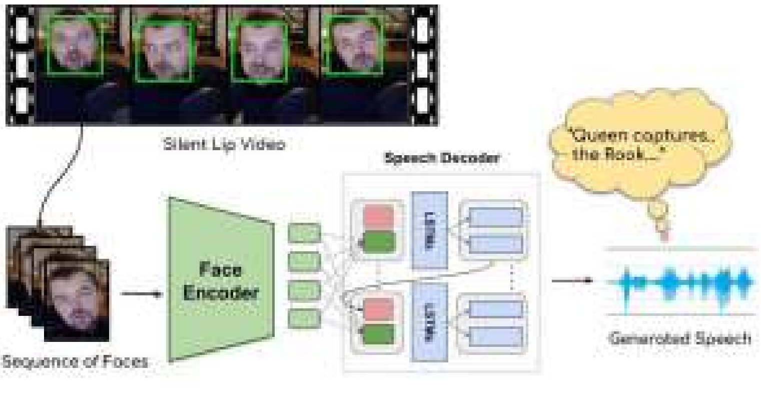
The probability distribution of each output speech time-step is given by:

P(S|I) = Πk(Sk|S<k, I) (1)

Lip2Wav, as shown in Figure 3 consists of two modules:

(i) Spatio-temporal face encoder (ii) Attention-based speech decoder. The modules are trained jointly in an end-to-end fashion. The sequence-to-sequence approach enables the model to learn an implicit speech-level language model that helps it to disambiguate homophenes.





Spatio-temporal Face Encoder:

Our visual input is a short video sequence of face images. The model must learn to extract and process the finegrained sequence of lip movements. 3D convolutional neural networks have been shown to be effective [18, 33, 36] in multiple tasks involving spatio-temporal video data. In this work, we try to encode the spatio-temporal information of the lip movements using a stack of 3D convolutions (Figure 3). The input to our network is a sequence of facial images of the dimension T × H × W × 3, where T is the number of time-steps (frames) in the input video sequence, H, W correspond to the spatial dimensions of the face image. We gradually down-sample the spatial extent of the feature maps and preserve the temporal dimension T. We also employ residual skip connections [14] and batch normalization [16] throughout the network. The encoder outputs a single D-dimensional vector for each of the T input facial images to get a set of spatio-temporal features T × D to be passed to the speech decoder. Each time-step of the embedding generated from the encoder also contains information about the future lip movements and hence helps in the subsequent generation.

Attention-based Speech Decoder:

The recent breakthroughs [27, 30] in text-to-speech generation. We adapt the Tacotron 2 [30] decoder which has been used to generate melspectrograms conditioned on text inputs. For our work, we condition the decoder on the encoded face embeddings from the previous module. We refer the reader to the Tacotron 2 [30] paper for more details about the decoder. The encoder and decoder are trained end-to-end by minimizing the L1 reconstruction loss between the generated and the ground-truth melspectrogram.

Figure 2. Our Lip2Wav dataset contains talking face videos of 5 speakers from chess analysis and lecture videos. Each speaker has about 20 hours of YouTube video content spanning a rich vocabulary of 5000+ words.

Gradual Teacher Forcing Decay:

In the initial stages of training, up to ≈ 30K iterations,we employ teacher forcing similar to the text-to-speech counterpart. We hypothesize that this enables the decoder to learn an implicit speech-level language model to help in disambiguating homophenes. Similar observations are made in lip to text works [2] which employ a transformer-based sequence-to-sequence model. Over the course of the training, we gradually decay the teacher forcing to enforce the model to attend to the lip region and to prevent the implicit language model from over-fitting to the train set vocabulary.

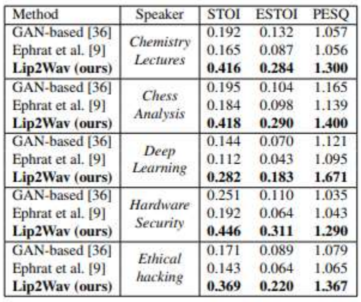
CHAPTER 7

IMPLEMENTATION

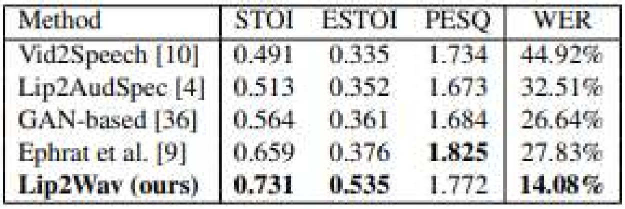
Implementation is the stage where the theoretical design is turned into a working system. Themost crucial stage in achieving a new successful system and in giving confidence on the new system for the users that it will work efficiently and effectively. We obtain results from our Lip2Wav model on all the test splits as described above. For comparing related work, we use the open-source implementations provided by the authors if available or re-implement a version ourselves. We compare our models with the previous lip to speech works using three standard speech quality metrics: Short-Time Objective Intelligibility (STOI) and Extended ShortTime Objective Intelligibility (ESTOI) for estimating the intelligibility and Perceptual Evaluation of Speech Quality (PESQ) to measure the quality. Using an outof-the-box ASR system4 , we obtain textual transcripts for our generated speech and evaluate our speech results using word error rates (WER) for the GRID and TCD-TIMIT lip speaker corpus . We, however do not compute WER for the proposed Lip2Wav corpus due to the lack of text transcripts. We also perform human evaluation and report the mean opinion scores (MOS) for the proposed Lip2Wav model and the competing methods. Next, we also perform extensive ablation studies for our approach and report our observations. Finally, as we achieve superior results compared to previous works in single-speaker settings, we end the experimental section by also reporting baseline results for word-level multi-speaker lip to speech generation using the LRW dataset and highlight its challenges as well.

TESTING

We now move on to evaluating our approach in unconstrained datasets that contain a lot of head movements and much broader vocabularies. They also contain a significant amount of silences or pauses between words and sentences. It is here that we see a more vivid difference in our approach compared to previous approaches. We train our model independently on all 5 speakers of our newly collected Lip2Wav dataset. The training details are mentioned in the sub-section. For comparison with previous works, we choose the best performing models [9, 36] on the TIMIT dataset based on STOI scores and report their performance after training on our Lip2Wav dataset. We compute the same metrics for speech intelligibility and quality that are used . The scores for all five speakers for our method and the two competing methods across all three metrics.



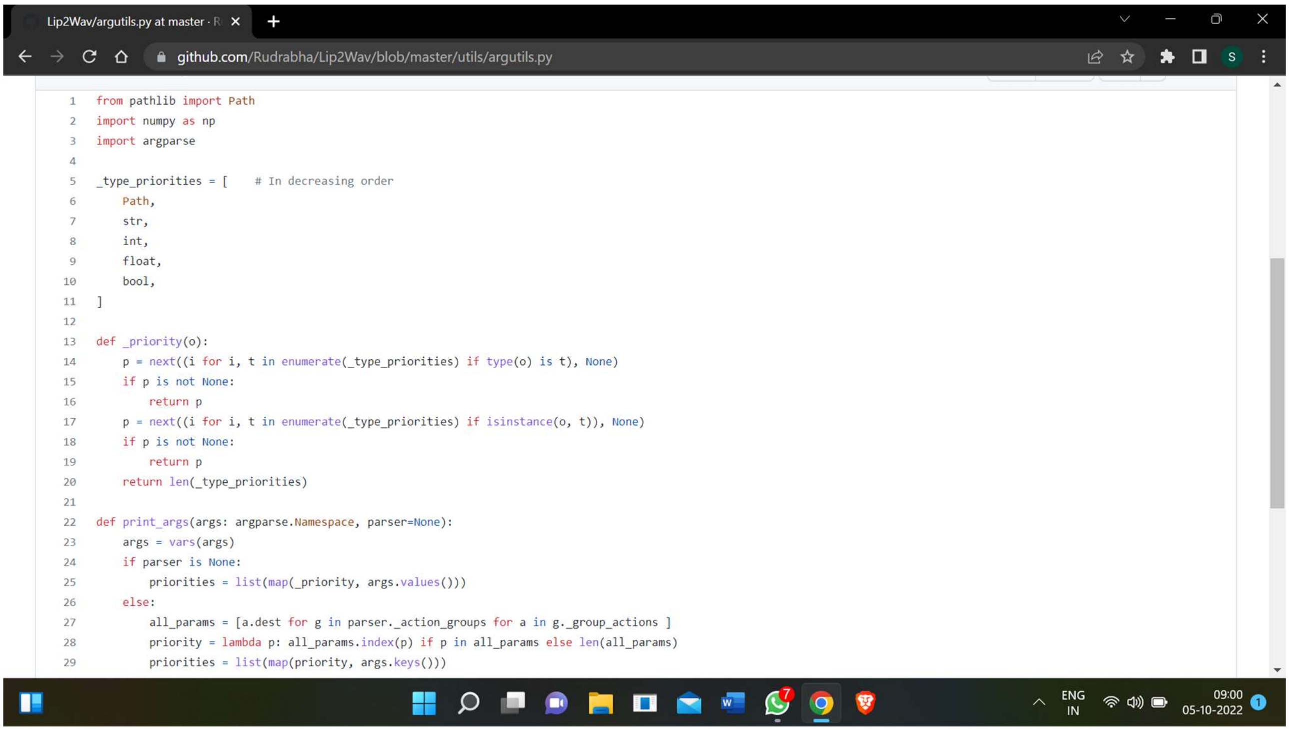
Method	Speaker	STOI	ESTOI	PESQ
GAN-based [36]	*Chemistry Lectures*	0.192	0.132	1.057
Ephrat et al. [9]		0.165	0.087	1.056
Lip2Wav (ours)		**0.416**	**0.284**	**1.300**
GAN-based [36]	*Chess Analysis*	0.195	0.104	1.165
Ephrat et al. [9]		0.184	0.098	1.139
Lip2Wav (ours)		**0.418**	**0.290**	**1.400**
GAN-based [36]	*Deep Learning*	0.144	0.070	1.121
Ephrat et al. [9]		0.112	0.043	1.095
Lip2Wav (ours)		**0.282**	**0.183**	**1.671**
GAN-based [36]	*Hardware Security*	0.251	0.110	1.035
Ephrat et al. [9]		0.192	0.064	1.043
Lip2Wav (ours)		**0.446**	**0.311**	**1.290**
GAN-based [36]	*Ethical hacking*	0.171	0.089	1.079
Ephrat et al. [9]		0.143	0.064	1.065
Lip2Wav (ours)		**0.369**	**0.220**	**1.367**



Method	STOI	ESTOI	PESQ	WER
Vid2Speech [10]	0.491	0.335	1.734	44.92%
Lip2AudSpec [4]	0.513	0.352	1.673	32.51%
GAN-based [36]	0.564	0.361	1.684	26.64%
Ephrat et al. [9]	0.659	0.376	**1.825**	27.83%
Lip2Wav (ours)	**0.731**	**0.535**	1.772	**14.08%**

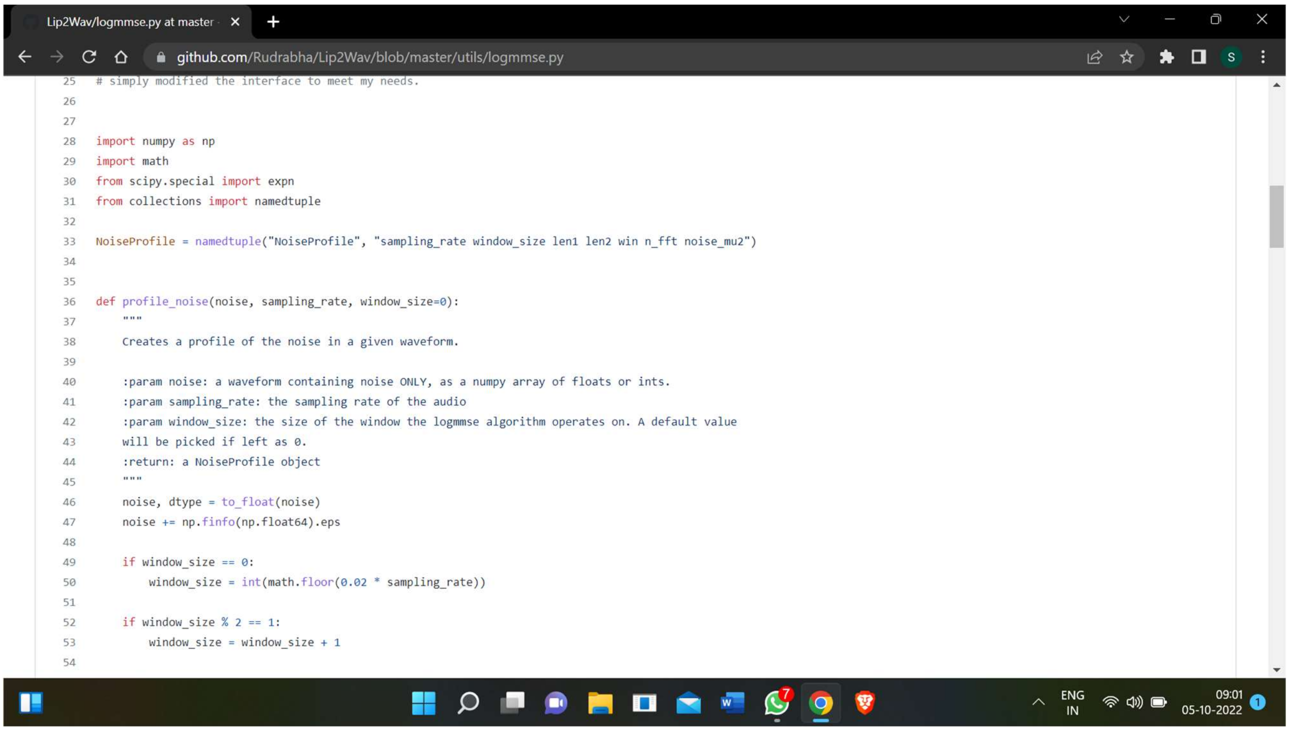
CHAPTER 8

SNAPSHOTS


Lip2Wav/argutils.py at master · R
github.com/Rudrabha/Lip2Wav/blob/master/utils/argutils.py
ENG
IN
09:00
05-10-2022

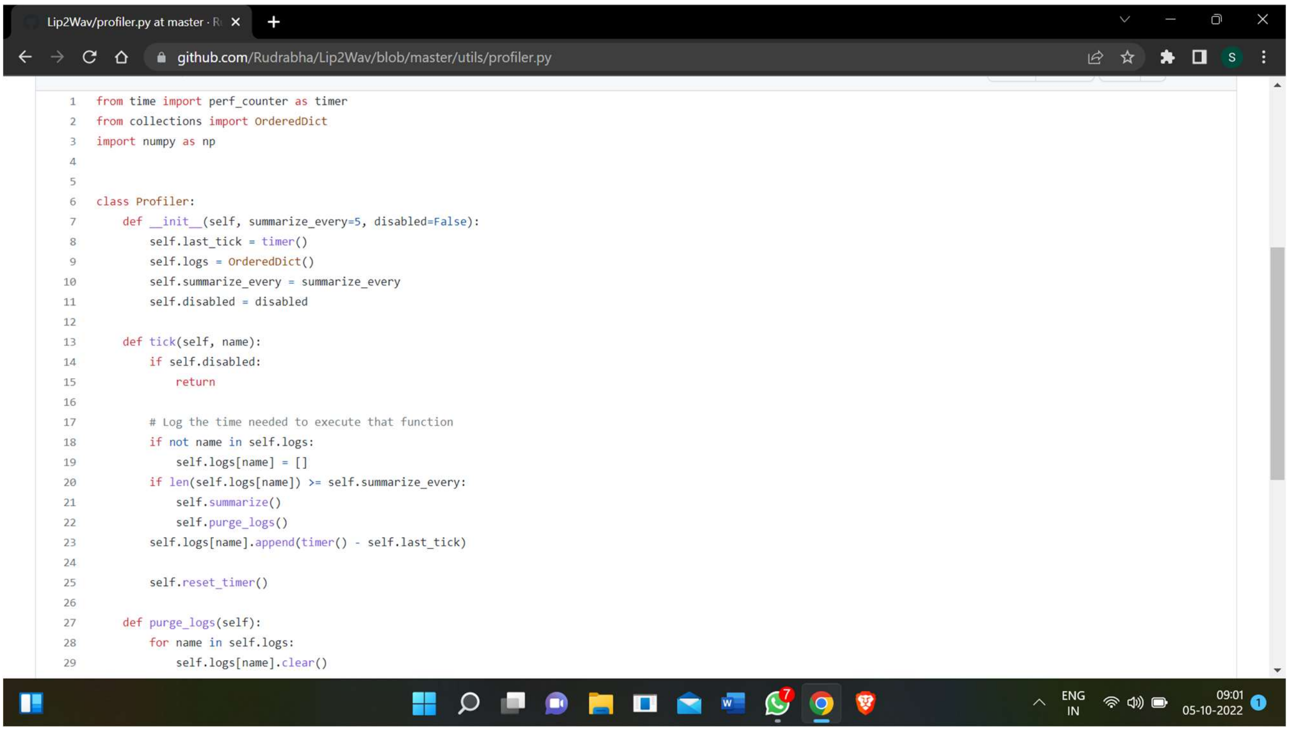
```python
from pathlib import Path
import numpy as np
import argparse

_type_priorities = [    # In decreasing order
    Path,
    str,
    int,
    float,
    bool,
]

def _priority(o):
    p = next((i for i, t in enumerate(_type_priorities) if type(o) is t), None)
    if p is not None:
        return p
    p = next((i for i, t in enumerate(_type_priorities) if isinstance(o, t)), None)
    if p is not None:
        return p
    return len(_type_priorities)

def print_args(args: argparse.Namespace, parser=None):
    args = vars(args)
    if parser is None:
        priorities = list(map(_priority, args.values()))
    else:
        all_params = [a.dest for g in parser._action_groups for a in g._group_actions ]
        priority = lambda p: all_params.index(p) if p in all_params else len(all_params)
        priorities = list(map(priority, args.keys()))
```

Fig 1: main.py


Lip2Wav/logmmse.py at master ·
github.com/Rudrabha/Lip2Wav/blob/master/utils/logmmse.py
ENG
IN
09:01
05-10-2022

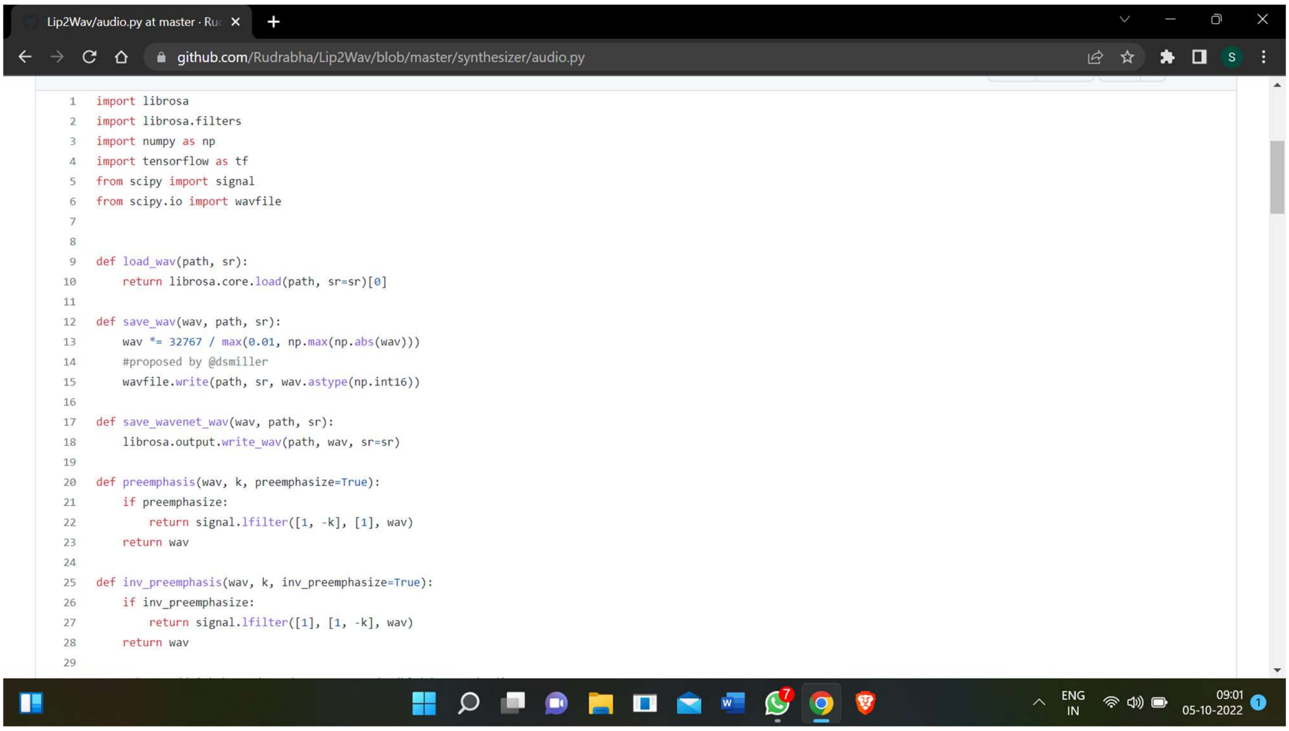
```python
# simply modified the interface to meet my needs.


import numpy as np
import math
from scipy.special import expn
from collections import namedtuple

NoiseProfile = namedtuple("NoiseProfile", "sampling_rate window_size len1 len2 win n_fft noise_mu2")


def profile_noise(noise, sampling_rate, window_size=0):
    """
    Creates a profile of the noise in a given waveform.

    :param noise: a waveform containing noise ONLY, as a numpy array of floats or ints.
    :param sampling_rate: the sampling rate of the audio
    :param window_size: the size of the window the logmmse algorithm operates on. A default value
    will be picked if left as 0.
    :return: a NoiseProfile object
    """
    noise, dtype = to_float(noise)
    noise += np.finfo(np.float64).eps

    if window_size == 0:
        window_size = int(math.floor(0.02 * sampling_rate))

    if window_size % 2 == 1:
        window_size = window_size + 1

```

Fig 2: detect.py


Lip2Wav/profiler.py at master · R
github.com/Rudrabha/Lip2Wav/blob/master/utils/profiler.py
ENG
IN
09:01
05-10-2022

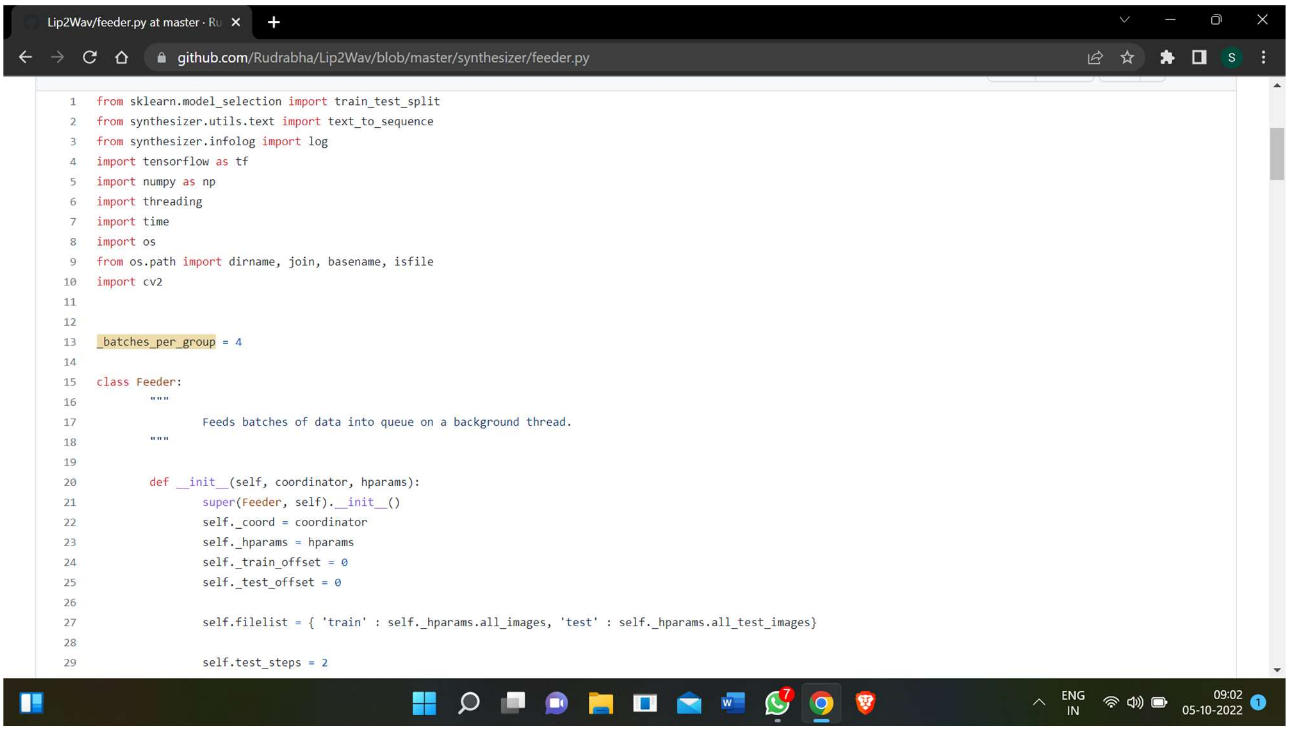
```python
from time import perf_counter as timer
from collections import OrderedDict
import numpy as np


class Profiler:
    def __init__(self, summarize_every=5, disabled=False):
        self.last_tick = timer()
        self.logs = OrderedDict()
        self.summarize_every = summarize_every
        self.disabled = disabled

    def tick(self, name):
        if self.disabled:
            return

        # Log the time needed to execute that function
        if not name in self.logs:
            self.logs[name] = []
        if len(self.logs[name]) >= self.summarize_every:
            self.summarize()
            self.purge_logs()
        self.logs[name].append(timer() - self.last_tick)

        self.reset_timer()

    def purge_logs(self):
        for name in self.logs:
            self.logs[name].clear()
```

Fig 3: fuctions.py


Lip2Wav/audio.py at master · Ru
github.com/Rudrabha/Lip2Wav/blob/master/synthesizer/audio.py
ENG
IN
09:01
05-10-2022

```python
import librosa
import librosa.filters
import numpy as np
import tensorflow as tf
from scipy import signal
from scipy.io import wavfile


def load_wav(path, sr):
    return librosa.core.load(path, sr=sr)[0]

def save_wav(wav, path, sr):
    wav *= 32767 / max(0.01, np.max(np.abs(wav)))
    #proposed by @dsmiller
    wavfile.write(path, sr, wav.astype(np.int16))

def save_wavenet_wav(wav, path, sr):
    librosa.output.write_wav(path, wav, sr=sr)

def preemphasis(wav, k, preemphasize=True):
    if preemphasize:
        return signal.lfilter([1, -k], [1], wav)
    return wav

def inv_preemphasis(wav, k, inv_preemphasize=True):
    if inv_preemphasize:
        return signal.lfilter([1], [1, -k], wav)
    return wav

```

Fig 4: speech.py


Lip2Wav/feeder.py at master · Ru
github.com/Rudrabha/Lip2Wav/blob/master/synthesizer/feeder.py
ENG
IN
09:02
05-10-2022

```python
from sklearn.model_selection import train_test_split
from synthesizer.utils.text import text_to_sequence
from synthesizer.infolog import log
import tensorflow as tf
import numpy as np
import threading
import time
import os
from os.path import dirname, join, basename, isfile
import cv2


_batches_per_group = 4

class Feeder:
    """
        Feeds batches of data into queue on a background thread.
    """

    def __init__(self, coordinator, hparams):
        super(Feeder, self).__init__()
        self._coord = coordinator
        self._hparams = hparams
        self._train_offset = 0
        self._test_offset = 0

        self.filelist = { 'train' : self._hparams.all_images, 'test' : self._hparams.all_test_images}

        self.test_steps = 2
```

Fig 5: yolopy.py

CHAPTER 9

CONCLUTION

Overall for this project, the concept of image or video analysis had been taken seriously in determining the success of this project. From starting of this project, many concepts to analyze of image had been done. Several techniques such as colour marker detection, optical flow analysis and snake tracking been taken look. After making a research, the colour marker detection is much easier and much more reliable to determine the movement of the lip. The other technique more reliable to be use for others application and for different situation. From the research and experiment that had been done, the result may be use to make a more reliable lip reading system. Factor like intensity of light in the test environment need to be consider. Besides that, when operating with the system that need more space of memory, our device of the laptop been use in this experiment need to be consider more. From the test had been done, after making several of test, the laptop may have a problem to processing the image. It may come out the result with an error or sometime the laptop itself will automatically shut down because of the problem of memory dump. For determining the successful of the project, a lot of researches need to be done and a lot of tests need to be working on. When facing with the signal processing concept, it needs a lot of understanding of the concept and need to analyze lot of mathematical problem to determine our flow of the project.

10.REFERENCE

- D. McGurck and J. MacDonald. Hearing lips and seeing voices. Nature, 264, December 1976.
- A. Greenwald. Lipreading made easy. Alexander Graham Bell Association for the Deaf, 1984.
- T. Chen, "Audiovisual speech processing. Lip reading and lip synchronization," IEEE Signal
- Processing Mag., vol. 18, January 2001, pp. 9-21.
- S. Lucey, S. Srindharan and V. Chandran (2001): "An investigation of HMM classifier
- combination strategies for improved audio-visual speech recognition," In: [Dalsgaardet al. (2001) Dastard, Lindberg and Benner], pp. 1185-1188.
- J. C. Wojdel and L. J. M. Rothkrantz (2001): "Using aerial and geometric features in automatic ipreading," In: [Dalsgaard et al.(2001)Dalsgaard, Lindberg and Benner], pp. 2463-2466.
- G.Potamianos and C.Neti (2001): "Automatic speechreading of impaired speech," In: [Massaroet al.(2001)Massaro, Light and Geraci], pp. 177-182.
- Github link = [https://github.com/Rudrabha/Lip2Wav]