VISVESVARAYA TECHNOLOGICAL UNIVERSITY

BELAGAVI – 590018, Karnataka

INTERNSHIP REPORT

ON

LIP TO SPEECH SYSNTHESIS

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

BACHELOR OF ENGINEERING IN
YOUR BRANCH

Submitted by:

P Likith Kumar

1BG19CS069

Conducted at
Varcons Technologies Pvt Ltd

B.N.M Institute of Technology

Department of Computer Science and Engineering
Accredited by NBA, New Delhi

12th Main Road, 27th Cross, Banashankari Stage II, Banashankari, Bengaluru, Karnataka 560070

Department of Computer Science and Engineering
Accredited by NBA, New Delhi

12th Main Road, 27th Cross, Banashankari Stage II, Banashankari, Bengaluru, Karnataka 560070

CERTIFICATE

This is to certify that the Internship titled **Lip to speech synthesis** carried out by **Mr. P Likith Kumar,** a bonafide student of **B.N.M Institute of Technology**, in partial fulfilment for the award of **Bachelor of Engineering**, in **Computer Science and Engineering** 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) _____

_____

D E C L A R A T I O N

I, **P Likith Kumar**, final year student of Computer science and Engineering, B.N.M Institute of Technology, declare that the Internship has been successfully completed, in Varcons Technologies. This report is submitted in partial fulfilment of the requirements for award of Bachelor Degree in Computer science and Engineering, during the academic year 2022-2023.

Date: 25-092022

Place: Bangalore

USN: 1BG19CS069

NAME: P Likith Kumar

OFFER LETTER

INTERNSHIP OFFER LETTER

Date: **23rd August, 2022**

Name: **Likith Kumar**
USN: **1BG19CS069**

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 **23rd August, 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

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 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.

P Likith Kumar 1BG19CS069

ABSTRACT

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 varication 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 essential problem is to and the correspondence between the audio and visual streams, which is the goal of this paper. We propose the use of a coupled 3D convolutional neural network (3D CNN) architecture that can map both modalities into a representation space to evaluate the correspondence of audio-visual streams using the learned multimodal features. The proposed architecture will incorporate both spatial and temporal information jointly to effectively and the correlation between temporal information for different modalities. By using a relatively small network architecture and much smaller data set for training, our proposed method surpasses the performance of the existing similar methods for audio-visual matching, which use 3D CNNs for feature representation. We also demonstrate that an effective pair selection method can significantly increase the performance. The proposed method achieves relative improvements over 20% on the equal error rate and over 7% on the average precision in comparison to the state-of-the-art method.

Table of Contents

CHAPTER 1

COMPANY PROFILE

1. COMPANY PROFILE

A Brief History of Varcons Technologies

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

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

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

2. ABOUT THE COMPANY

Varcons 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 web promotion and tailor-made software products, designing solutions best suiting clients requirements. The organization where they have a right mix of professionals as a stakeholders to help us serve our clients with best of our capability and with at par industry standards. They have young, enthusiastic, passionate and creative Professionals to develop technological innovations in the field of Mobile technologies, Web applications as well as Business and Enterprise solution. Motto of our organization is to "Collaborate with our clients to provide them with best Technological solution hence creating Good Present and Better Future for our client which will bring a cascading a positive effect in their business shape as well". Providing a Complete suite of technical solutions is not just our tag line, it is Our Vision for Our Clients and for Us, We strive hard to achieve it.

Products of Varcons 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 tutorials. 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 software development is possible by using specialized Android applications.

Web Application

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

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

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

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

Web design

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

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

Departments and services offered

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 about the academics. Personality development and academic status are some of those things which lie on mentor's hands. If you are trained well then you can do well in your future and knowing its importance of 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

3. INTRODUCTION

Introduction to ML

AIML stands for **Artificial Intelligence Markup Language**. AIML was developed by the Alicebot free software community and Dr. Richard S. Wallace during 1995-2000. AIML is used to create or customize Alicebot which is a chat-box application based on A.L.I.C.E. (Artificial Linguistic Internet Computer Entity) free software. AIML can be used in the financial services industry for various tasks such as developing chatbots, predictive modeling, developing virtual assistants, etc. AIML develops chatbots for customer service or support in the financial services industry. As businesses and other organizations undergo digital transformation, they're faced with a growing tsunami of data that is at once incredibly valuable and increasingly burdensome to collect, process and analyze. New tools and methodologies are needed to manage the vast quantity of data being collected, to mine it for insights and to act on those insights when they're discovered

Problem Statement

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 varication 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 essential problem is to and the correspondence between the audio and visual streams, which is the goal of this paper. We propose the use of a coupled 3D convolutional neural network (3D CNN) architecture that can map both modalities into a representation space to evaluate the correspondence of audio-visual streams using the learned multimodal features

The proposed architecture will incorporate both spatial and temporal information jointly to effectively and the correlation between temporal information for different modalities. By using a relatively small network architecture and much smaller data set for training, our proposed method surpasses the performance of the existing similar methods for audio-visual matching, which use 3D CNNs for feature representation. We also demonstrate that an effective pair selection method can significantly increase the performance. The proposed method achieves relative improvements over 20% on the equal error rate and over 7% on the average precision in comparison to the state-of-the-art method.

CHAPTER 4

SYSTEM ANALYSIS

4. SYSTEM ANALYSIS

1. Existing System

The main characteristic of CNNs is their locality, i.e., the convolution operation is applied to specific local regions in an image must be locally correlated in the sense of time and frequency on both axes respectively. The main characteristic of CNNs is their locality, i.e., the convolution operation is applied to specie local regions in an image. As a visual inference of this locality property, the neighbour features should be correlated in some sense. Since the input speech feature maps are treated as images when a CNN architecture is used, the features

2. Proposed System

The network input is a pair of features that represent lip movement and speech features extracted from 0.3-second ofa video clip. The main task is to determine if a stream of audio corresponds with a lip motion clip within the desired stream duration.The difficulty of this task is the short time interval of the video clip (0.3-0.5 second) considered to evaluate the method. This setting is close to real-world scenarios because,in some biometrics or forensics applications, only a short amount of captured video or audio might be available to distinguish between different modalities. Temporal video and audio features must correspond over the time interval they
cover. This correspondence is discussed in the next two sections.

3. Objective of the System

The propose the use of a coupled 3D convolutional neural network (3D CNN) architecture that can map both modalities into a representation space to evaluate the correspondence of audio-visual streams using the learned multimodal features. The proposed architecture will incorporate both spatial and temporal information jointly to effectively and the correlation between temporal information for different modalities. By using a relatively small network architecture and much smaller data set for training, our proposed method surpasses the performance of the existing similar methods for audio-visual matching, which use 3D CNNs for feature representation.

CHAPTER 5

REQUIREMENT ANALYSIS

5. REQUIREMENT ANALYSIS

Hardware Requirement Specification

The most common set of requirements defined by any operating system or software application is the physical computer resources, also known as hardware.

CPU: Intel or AMD processor

Cores: Dual-Core (Quad-Core recommended)

RAM: minimum 4GB (>4GB recommended)

Graphics: Intel Integrated Graphics or AMD Equivalent

Secondary Storage: 250GB

Display Resolution: 1366x768 (1920x1080 recommended)

Software Requirement Specification

Software requirements deal with defining software resource requirements and prerequisites that need to be installed on a computer to provide optimal functioning of an application.
The following are the software requirements for the application :

Operating System: Windows 10

Compiler: GNU C/C++ Compiler

Development Environment: Visual Studio 2019 Community Edition

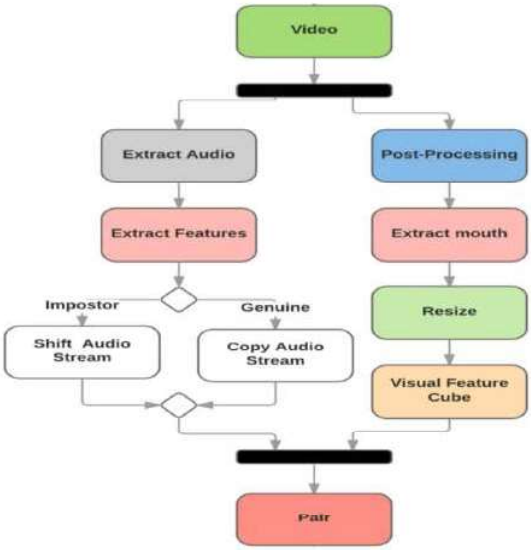
PI: OpenGL API & Win32 API for User Interface and Interaction

CHAPTER 6

DESIGN ANALYSIS

6. DESIGN & ANALYSIS

The processing pipeline of both datasets is shown in Fig. 1.The pipeline is subdivided into two visual and audio sections. In the visual section, the videos are post-processed to have equal frame rate of 30 f/s. Then, face tracking and mouth area extraction is performed on the videos using the dlib library [31]. Finally, all mouth areas are resized to have the same size, and concatenated to form the input featurecube. The dataset does not contain any audio _les. In the audio section, the audio _les are extracted from videos usingthe FFmpeg framework [32]. Then the speech features will beextracted from audio _les. The library that has been used forspeech feature extraction task is SpeechPy [33].



The network input is a pair of features that represent lip movement and speech features extracted from 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. The difficulty of this task is the short time interval of the video clip (0.3-0.5 second) considered to evaluate the method. This setting is close to real-world scenarios because, in some biometrics or forensics applications, only a short amount of captured video or audio might be available to distinguish between different modalities. Temporal video and audio features must correspond over the time interval they cover. This correspondence is discussed in the next two sections.

CHAPTER 7

IMPLEMENTATION

7. IMPLEMENTATION

For audio-visual matching using the *Lip Reading in the Wild* dataset, 500 words (subjects) are available. To make the train and test sets mutually exclusive, the _rst 400 words are used for creating the training set and the remaining 100 words are used for test set generation. For each of the train/test sets, only 50 utterances of each word are chosen for data generation. The compiled initial training data contains generated genuine and impostor pairs. The reason this is described as initial training data is that not all the generated data is used for training. The method of selecting pairs was described in Section VII-A. Genuine pairs (audio/video) are created by matching the 9-channel visual feature cube with the corresponding audio feature cube as we discussed earlier in Section IV. For impostor pair generation, the audio feature map for a video is shifted alongside its time axis. The shifting is random, and could bup to 0.5-second at maximum. This shifting method allows the network to learn the matching between

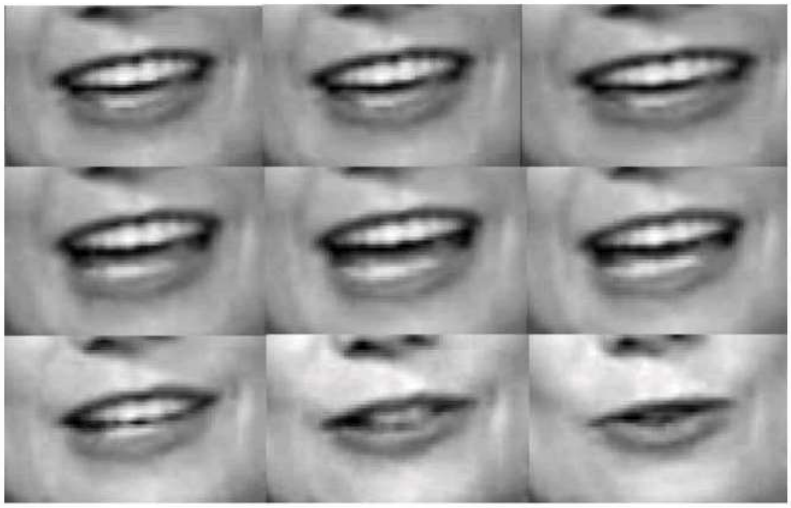
TESTING

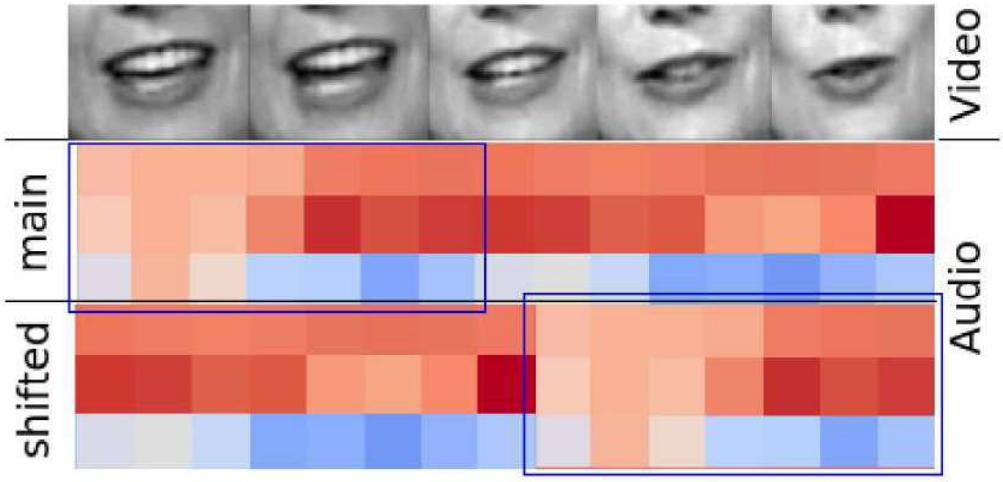
In this section, the effect of choosing different data representation (MFCC/MFEC features & using temporal derivatives) on the performance results is investigated. We compare our method with the state-of-the-art in [26] in which regular CNN architecture with MFCCs for speech feature representation is used. In [26] only one channel MFCC has been used. We modify the structure represented in [26] to accept 3-channels input speech features as we use MFCC alongside with its _rst and second order derivatives. For having a more comprehensive comparison, we compare our method with two common audio-visual synchrony approaches, based on CCA [28] and CoIA [40]. Prior to CCA/CoIA transformation, since the audio features are extracted at a faster rate, we interpolated the visual lip motions to have the same frame rate as audio features. Empirical evidence showed that not all the canonical correlations carry useful information. Considering the aforementioned evidence, only 20 dimensions of the correlation feature vector extracted from CCA or CoIA operations on audio-visual features are chosen which are corresponding to the higher correlation coef_cients. For speech feature representation, addition to static MFCC features, _rst and second orderderivatives have been used as well.

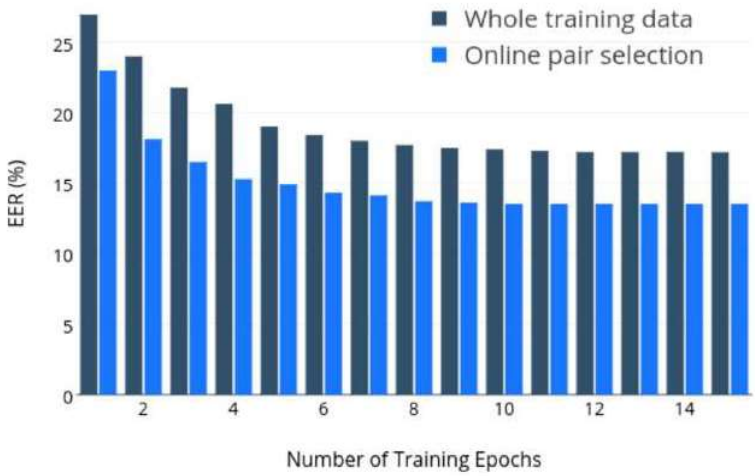
CHAPTER 8

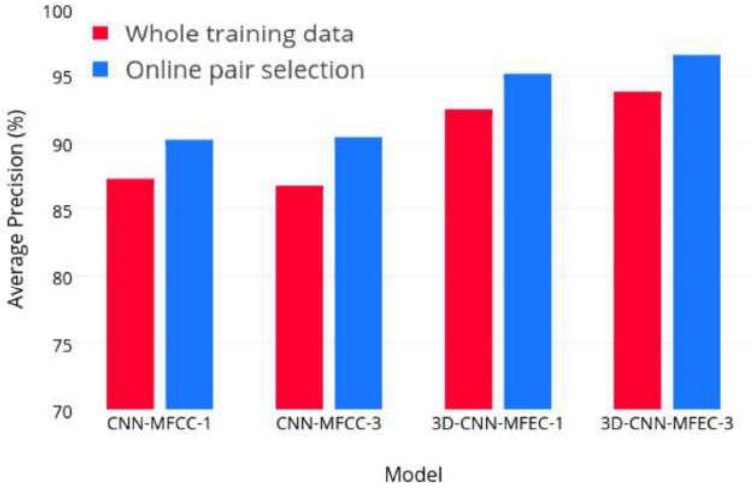
SNAPSHOTS

8. SNAPSHOTS






Whole training data
Online pair selection
EER (%)
25
20
15
10
5
0
2
4
6
8
10
12
14
Number of Training Epochs


Whole training data
Online pair selection
Average Precision (%)
100
95
90
85
80
75
70
CNN-MFCC-1
CNN-MFCC-3
3D-CNN-MFEC-1
3D-CNN-MFEC-3
Model

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 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.

10. REFERENCE

[1] G. Hinton *et al.*, ``Deep neural networks for acoustic modelling in speech recognition: The shared views of four research groups," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82_97, Nov. 2012.

[2] Q.V. Le *et al.*, ``Building high-level features using large scale unsupervised learning," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*,
May 2013, pp. 8595_8598.

[3] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, ``Multimodal deep learning," in *Proc. ICML*, 2011, pp. 1_8.

[4] P. Wu, H. Liu, X. Li, T. Fan, and X. Zhang, ``A novel lip descriptor for audio-visual keyword spotting based on adaptive decision fusion," *IEEE Trans. Multimedia*, vol. 18, no. 3, pp. 326_338, Mar. 2016.

[5] J. Huang and B. Kingsbury, ``Audio-visual deep learning for noise robust speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2013, pp. 7596_7599.

[6] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, ``Multimodal deep learning," in *Proc. 28th Int. Conf. Mach. Learn. (ICML)*, 2011,
pp. 689_696.

[7] C. Neti *et al.*, ``Audio visual speech recognition," IDIAP, Tech. Rep. EPFLREPORT- 82633, 2000.

[8] E. Erzin, Y. Yemez, and A. M. Tekalp, ``Multimodal speaker deification using an adaptive classi_er cascade based on modality reliability," *IEEE*
Trans. Multimedia, vol. 7, no. 5, pp. 840_852, Oct. 2005.

[9] S. Zeiler, R. Nicheli, N. Ma, G. J. Brown, and D. Kolossa, ``Robust audio-visual speech recognition using noise-adaptive linear discriminant analysis,"
in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 2797_2801.

[10] J. Wang, J. Zhang, K. Honda, J. Wei, and J. Dang, ``Audio-visual speech recognition integrating 3D lip information obtained from the Kinect,"*Multimedia Syst.*, vol. 22, no. 3, pp. 315_323, 2016.