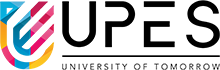
**MINOR-1 PROJECT**

**SYNOPSIS REPORT**

**Lipnet AI**



|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |

**School of Computer Science**

**UNIVERSITY OF PETROLEUM AND ENERGY STUDIES**

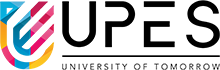
**B. Tech CSE, 5th Sem, B-3 AIML**

**Approved By**

**Dr. Alok Agarwal**

**Dr. Alok Agarwal Dr. Anil Kumar**

(**Project Guide) (Cluster Head)**



**School of Computer Science**

University of Petroleum & Energy Studies, Dehradun

**Table Of Content**

|  |  |
| --- | --- |
| **1.** | **Project title** |
| **2.** | **Abstract** |
| **3.** | **Introduction** |
| **4.** | **Literature** |
| **5.** | **Objective** |
| **6.** | **Problem Statement** |
| **7.** | **Methodology** |
| **8.** | **Dataflow** |
| **9.** | **NN Architecture** |
| **10.** | **References & GIT Link** |
|  |  |
|  |  |

***ABSTRACT***

* Lipreading, the skill of interpreting text from a speaker's lip movements, has experienced a remarkable transformation in recent years. Conventional methods traditionally divided this task into two distinct phases: the development or learning of visual features and subsequent prediction. Recent research has shed light on a captivating phenomenon in this domain.
* Motivated by these insights, we present Lip-Net—a groundbreaking model designed to map sequences of video frames of varying lengths to textual content. Lip-Net harnesses spatiotemporal convolutions, recurrent networks, and employs the connectionist temporal classification loss function, all seamlessly integrated into a comprehensive end-to-end training framework. Notably, Lip-Net stands as an innovative achievement, being the first-ever end-to-end model for sentence-level lipreading. It adeptly combines the extraction of spatiotemporal visual features with sequence modeling.
* In rigorous testing, Lip-Net aspires to achieve an outstanding accuracy rate of 90% in sentence-level lipreading tasks, surpassing the capabilities of even seasoned human lipreaders. This remarkable advancement in lipreading technology signifies a substantial leap toward enhancing human-computer interaction, promoting accessibility, and facilitating communication for individuals with hearing impairments and beyond.

**Literature**

* AI-driven lipreading faces formidable challenges, as it must interpret subtle visual cues and variations in lip and facial movements. Factors like lighting conditions, speaker variations, and background noise can affect the accuracy of AI-based lipreading systems.
* Recent advancements in AI have propelled lipreading technology forward. Deep learning models, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), have significantly improved accuracy. These AI-driven lipreading systems find diverse applications, including real-time transcription services for the deaf and hard of hearing, enhancing human-computer interaction, and bolstering security systems.

**PROBLEM STATEMENT**

* Title: Improving Lip-Net for Recognizing Speech at Different Levels: Sentences, Phonemes, and Movie Dialogue
* Background : Lip-Net is a technology that uses computers to understand what people are saying by looking at their lip movements. Originally, it was designed to understand full sentences. But we want to make it even better and use it to understand smaller parts of speech and even what actors are saying in movies.
* Objective: Our goal in this project is to explore how Lip-Net can be used to understand speech in different ways. We want to use it to recognize full sentences, individual speech sounds (phonemes), and what people are saying in movies.
* Problem Description:
* 1. Sentence-Level Recognition: - We aim to make Lip-Net really good at understanding and writing down full sentences, just by looking at how people move their lips. To do this, we will teach Lip-Net using videos of people speaking sentences. We'll check how well it works and if it can handle different speaking styles.
* 2. Phoneme-Level Analysis: - We want Lip-Net to be able to recognize individual speech sounds, like 'ba' or 'pa.' To do this, we'll break down videos into these small speech sounds and train Lip-Net to tell them apart accurately.
* 3. Movie Monologue/Dialogue Recognition: We'll use Lip-Net to understand what actors are saying in movies. This is a bit tricky because there are often many actors talking at once, and they might be emotional. We want Lip-Net to handle these situations well.

**MOTIVATION**

Our "Lip-Net AI" college minor project aims to improve end-to-end phrases, phonemes, and dialogues to improve human-computer interaction. Our project's goal is to increase sentence creation for genuine dialogues, phoneme identification accuracy for accurate speech-to-text conversion, and dialogue production for a more human-like connection. Our project's goals are to improve AI applications' usability, accuracy, and adaptability, therefore redefining human-machine communication. Our project's impact will be that chat-bots and virtual assistants will be able to speak organically with users, changing human-machine communication.

**CONTRIBUTION**

* Enhancing Sentence Generation: We are enhancing the 'Lip-Net AI's' capacity to produce sentences that are both naturally sounding and contextually suitable.
* This suggests that AI may engage in more insightful and human-like conversations with customers, much like a chatbot or virtual assistant that has a deeper grasp of your needs.
* Increasing Phoneme Recognition Accuracy: We are striving to increase 'Lip- Net AI's' ability to accurately recognize even the smallest spoken sounds, called phonemes.
* This development is essential for tasks like accurately transcribing spoken words, which aids voice assistants and transcription services.

**OBJECTIVE**

* Lipreading model that will predict subtitles without audio using phoneme dictionary.
* To achieve accuracy more than 50% which is much much better than human level performance.
* Increase our understanding among Algorithms.
* (optional) Working on different languages for example English India, Hinglish, Hindi (Hinglish and Hindi models will be an innovation).

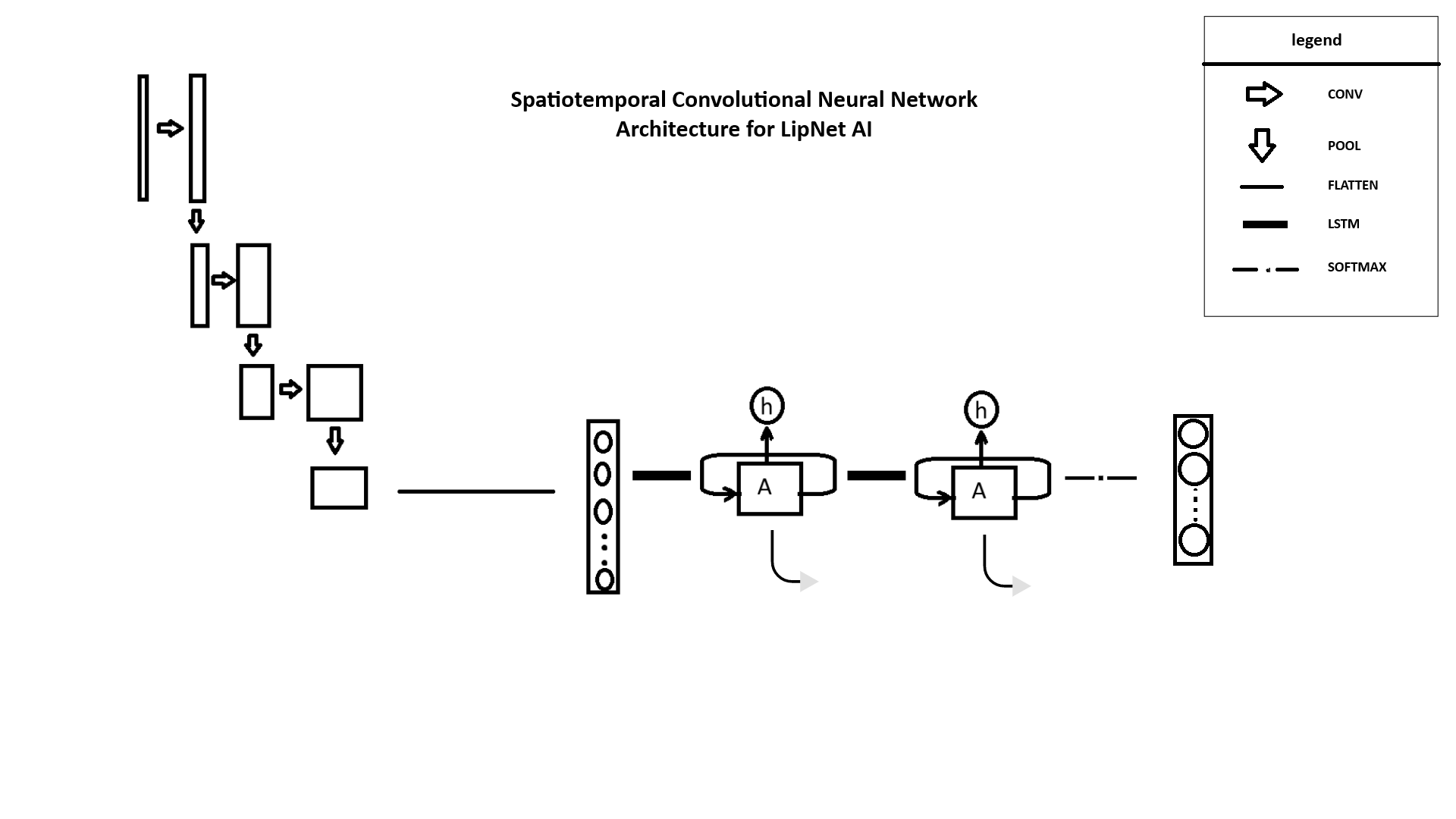
**Data Flow**

* At first, video data is converted to a custom extension .ALIGN or donwloaded directly. .ALIGN files represent frames of video data in alignment.
* Pipelines will be created then.
* After data will feed into Saptial Convolutional Neural Network Architecture
* First three layers will be traditional combinations of CONV and MAX POOL layers
* After FLATTEN layer BIDIRECTIONAL layers will be used for understanding and predicting phonemes.
* Finally SOFTMAX layer is used as output layer.
* Model develpment phase is over, now we can evaluate or make predictions

**Methodology**

* Dataset Collection - Dataset will collected from web source, for English, (optional) Hindi or Hinglish will be developed
* Data Preprocessing - It will be done by making functions for different algorithms such as BILINEAR INTERPOLATION, etc
* Data Distribution - This step will give final touch to data so that it can be feeded to model
* Model Architecture Development - Spatiotemporal Convolutional Neural Network Architecture
* Model Training and Development - Running iterations to achive accuracy
* Model Evaluation - Validation and Testing
* Model Deployment (optional) - Deployment on localhost, basically web development

**NN Architecture**

****