CHAPTER 1

Introduction

1.1 Background

Lipreading is the process of deciphering text from a speaker's mouth movements. In traditional methods, this task was divided into two main stages: the creation or learning of visual features, and the prediction of the spoken content. In the field of lip-reading classification, the backend systems are specifically designed to anticipate sequential speech elements, such as words or sentences. To achieve this, these systems often employ sequences of processing networks, including Recurrent Neural Networks (RNNs), which can be further defined in terms of Long Short Term Memory Networks (LSTM) or Gated Recurrent Units (GRUs). In addition to RNNs, lip-reading backend systems also utilize alternative classification networks such as Attention-based Transformers (ATT) and Temporal Convolution Networks (TCN). The initial attempts to automate lip-reading mainly relied on non-deep learning techniques, such as Markov Models.

1.2 Relevance

Lipreading: video to speech conversion is highly pertinent to the field of Electronics and Communication Engineering (ECE) and related subjects. In ECE, we learned about various aspects of signal processing, including speech signal processing, which is fundamental to understanding and working with audio data. Our project, which involves the extraction of speech information from visual cues in videos, directly aligns with the principles of signal processing taught in ECE programs. Lipreading video to speech exemplifies the application of technology in improving human-computer interaction and accessibility, which are vital themes within the ECE curriculum. This project not only contributes to advancing communication technology but also highlights the interdisciplinary nature of ECE as it draws from both electrical engineering and computer science to create innovative solutions.

1.3 Literature Survey

The study done by Joon Son Chung et al. in [1] compared the performance of the 'Watch, Listen, Attend and Spell' (WLAS) network with the LRS dataset, which demonstrated that visual cues can significantly improve speech recognition in combination with audio data. Furthermore, the WLAS model was able to outperform prior benchmarks, even outperforming a professional lip reader in BBC videos.

In [2] LipNet, an end-to-end trainable model for sentence-level lipreading is presented. Spatial-temporal convolutions, a recurrent network, and the connectionist temporal classification loss are used. The model converts a variable-length video frame sequence to text, capturing temporal context in lipreading and allowing sentence-level sequence prediction. The model is tested using the GRID corpus, which is a dataset for sentence-level lipreading. The GRID corpus is made up of videos of speakers saying sentences containing various words and phrases. It offers a difficult and realistic testbed for evaluating lipreading models. LipNet performs admirably on the GRID corpus, achieving 95.2% accuracy in sentence-level lipreading. This outperforms skilled human lip readers and exceeds previous state-of-the-art word-level accuracy. In lipreading, the model highlights the necessity of spatiotemporal feature extraction and efficient temporal aggregation. It also emphasizes the value of end-to-end models for applications like quiet dictation and audiovisual speech recognition. The report makes no mention of any potential biases or limits of employing a character-level approach for lipreading, such as the effect of homophones or similar lip movements for distinct phonemes.

A thorough examination of various approaches to improve lip-reading system performance, such as data augmentation, temporal modeling, self-dissecting methods, and word boundary indicators was done [3]. The outcome of the analysis was highly encouraging, as the combination of these approaches resulted in an increase in accuracy of 93.4% compared to the current performance on the LRT dataset by 4.6%. These findings demonstrate the potential for incorporating these methods to significantly improve the accuracy and effectiveness of automated lip-reading systems.

A 3D convolutional vision transformer (3DCvT)-based lip-reading system was introduced in [4] that combines a vision transformer with 3D convolution to extract

spatiotemporal characteristics from continuous images. Following that, the collected features are passed to a Bidirectional Gated Recurrent Unit (BiGRU) for sequence modeling. The approach is tested using the LRW and LRW-1000 largescale lip reading datasets. On the LRW dataset, the best accuracy is 88.5%, whereas on the LRW-1000 dataset, it is 57.5%. Due to the intricacy of the Chinese language, the accuracy of the Chinese lip reading dataset LRW-1000 is relatively poor. To extract more robust spatio-temporal features from continuous images, the suggested 3D convolutional vision transformer (3DCvT) can be improved and optimized.

Xing Zhao et al. focused on the implementation of mutual information constraints on both local and global lip-reading features [5]. This novel approach was assessed using two large-scale benchmark datasets. The results were highly promising, as it led to a new level of performance on both benchmarks. This demonstrates the effectiveness of mutual information limitations in improving lip-reading accuracy and capabilities, which is a major step forward in the field. From a visual point of view, there are still issues with speaker dependency, especially when attempting to use lip-reading on individuals who were not included in the training dataset. Additionally, generic lip-reading systems must be able to handle videos with varying spatial resolution and frame rates, which contain varying amounts of temporal data.

A deep learning approach for audiovisual speech recognition was employed in [7], though the dataset used was not specifically specified. The results of the study were remarkable, as they revealed a significant improvement in performance. Specifically, the word error rate of the ASR system decreased by 6.59% when transcribing spoken language, and the lip reading model reached an impressive 95% accuracy rate, demonstrating the capability of deep learning methods to effectively combine audio and visual signals for speech recognition.

For the first time, the research [8] introduces an attention-based pooling approach to aggregate visual speech representations and employs subword units for lip reading. It also introduces a Visual Speech Detection (VSD) model trained on top of the lip reading network. The suggested models are trained and evaluated using the LRS2, LRS3, and AVA-ActiveSpeaker benchmarks. When trained on public datasets, it obtains state-of-the-art performance on the LRS2 and LRS3 benchmarks, and even outperforms

models trained on large-scale industrial datasets with an order of magnitude less data. On the LRS2 dataset, the top model obtains a word error rate of 22.6%, dramatically narrowing the performance gap between lip reading and automatic speech recognition. Furthermore, in the AVA-ActiveSpeaker test, the VSD model beats several current audio-visual techniques and excels all visual-only baselines. The Visual Speech Detection (VSD) model built on top of the lip reading system provides cutting-edge results, exceeding visual-only baselines and even numerous audio-visual techniques.

A novel Alternating Spatiotemporal and Spatial Convolutions (ALSOS) module was introduced to the methodology, combining spatiotemporal convolutions with spatial convolutions to enhance lip-reading performance was introduced in [9]. The module was integrated with both a Greek and a LRW-500 language dataset, and the results were promising, with the ALSOS module significantly improving lip-reading system performance, particularly in terms of spoken word accuracy. This integration highlights the potential of the ALSOS module as a valuable component of lip-reading technology that can improve results in a variety of language datasets.

The idea of using deep 3D Convolutional Neural Networks (CNNs) as the front-end for visual feature extraction in word-level lipreading was introduced in the paper composed by Xinshuo Weng and Kris Kitani [10]. The authors specifically replace the shallow 3D CNNs + deep 2D CNNs front-end with a two-stream I3D network composed of grayscale video and optical flow streams. Different combinations of front-end and back-end modules are evaluated using the LRW dataset. Pretraining on large-scale image and video datasets, such as ImageNet and Kinetics, is performed to improve classification accuracy. The LRW dataset, which contains short video clips extracted from BBC TV broadcasts, is used for training and evaluation. The dataset consists of 488,766 training videos, 25,000 validation videos, and 25,000 testing videos. The two-stream I3D front-end with a Bi-LSTM back-end achieves an absolute improvement of 5.3% over the previous state-of-the-art on the LRW dataset.

A combination of HPConv (Hierarchical Pyramidal Convolution) and Self-attention (Self-Attention) methods in a new way was accomplished in the recent study. The study used the lip-reading in-wild dataset (LRW). The results were remarkable, with the proposed method achieving an impressive 86.83% accuracy rate, which is a significant

improvement from the current lip-reading system (1.53%). These results demonstrate the power of HPConV and SelfAttention to significantly improve the accuracy and performance of lip reading technology, demonstrating its potential for progress in the field. In order to achieve its goals, the study [14] used a combination of Computer Vision (CV) and CNN (Deep Convolutional Network) models to transcribe spoken sentences into text. To test this method, the researchers used the GRID (Audiovisual Language Interpreter Interpreter) dataset, which consists of 1000 spoken sentences from 34 different speakers. The results showed that the system was able to generate an output string that accurately represented the spoken sentence as written. This success highlights the potential of combining CV and CNN models to rewrite spoken sentences into text, which has applications in speech recognition, accessibility technologies, and more.

In a recent study conducted by Liang Lin et al. [13] proposes a novel reconfigurable part-based model called the And-Or graph model for object shape detection in images. The model is divided into four layers: leaf-nodes for recognising contour fragments, or-nodes for activating leaf-nodes, and-nodes for capturing holistic shape deformations, and a root-node for handling global changes. The authors propose a structural optimization algorithm to train the And-Or model from weakly annotated data. The authors publish a new shape database with annotations that contains over 1500 difficult form cases for recognition and detection. On various challenging datasets, the proposed model outperforms current state-of-the-art algorithms in robust shape-based object detection against background clutter. On various challenging datasets, the proposed model outperforms current state-of-the-art algorithms in robust shape-based object detection against background clutter.

The system utilized in [24] included working out the Signal-to-Noise Ratio (SNR) for both the spotless sound (SNRi) and the sound with foundation commotion (SNRw) utilizing a solitary mouthpiece. The dataset used for this examination was the Methodology corpus, which contains high-goal, high-frame rate video transfers and sound accounts caught under uproarious circumstances. The review primarily focuses on the exposition and in-depth analysis of the results obtained from the deployment of a comprehensive media Automatic Speech Recognition (ASR) engine. This ASR engine was trained and extensively tested using data sourced from the Methodology corpus. This

exploration contributes significant bits of knowledge into the field of general media ASR and its exhibition in genuine world, loud conditions.

The work on lipreading using hidden Markov models (HMMs) with Active Appearance Model (AAM) algorithms for appearance model fitting is attributed to Iain Matthews, Timothy Cootes, J Bangham, Stephen Cox, and Richard Harvey. Their research incorporates feature extraction techniques encompassing lip deformations, head motion, and speech cues, with a focus on a multitalker visual speech recognition task specifically designed for recognizing isolated letters. Through the integration of HMMs, AAM algorithms, and comprehensive feature extraction methods, their work contributes to advancing the field of lipreading and visual speech recognition, particularly in scenarios involving multiple talkers and isolated letter recognition tasks.

To progress in the field of lip-perusing, [18] tackled the force of the MS-TCN technique for lip-perusing related to the AVHuBERT model for highlight extraction. The exploration utilized a significant dataset comprising of Persian word-level lipreading materials, incorporating a tremendous storehouse of roughly 244,000 video tests. Quite, the results of this examination exhibited a noteworthy improvement in exactness while utilizing the AV-HuBERT highlight extraction approach. These discoveries highlight the urgent job that includes extraction, plays in supporting the exhibition of lip-understanding frameworks, especially for the acknowledgment of Persian words.

The Connectionist-hidden Markov model (HMM) system for noise-robust Audio-Visual Speech Recognition (AVSR) was developed by Kuniaki Noda, Yuki Yamaguchi, Kazuhiro Nakadai, Hiroshi Okuno, and Tetsuya Ogata. Their work integrates several innovative components, including a Deep Denoising Autoencoder for preprocessing audio features to mitigate noise effects, deep learning approaches for robust latent feature extraction from both audio and visual modalities, and the incorporation of speakers' lip movements to enhance Automatic Speech Recognition (ASR) performance. This comprehensive approach enhances the system's robustness in noisy environments, making it well-suited for real-world applications. The limited variations of lip region images affected CNN training in the development of the Connectionist-hidden Markov model.

The weakly supervised learning scheme for sign language recognition, as proposed by Oscar Koller, Hermann Ney, and Richard Bowden [35], features Deep Neural Networks (DNNs) as a replacement for traditional methods such as Active Appearance Models (AAM) feature extraction and Gaussian Mixture Models (GMMs). Within this framework, Convolutional Neural Networks (CNNs) serve as the primary machine learning algorithm. The system exploits a variety of visual cues, including mouth shapes, facial features, Histogram of Oriented Gradients (HOG), Scale-Invariant Feature Transform (SIFT), and head pose. Notably, the dataset employed follows a weak supervision approach, meaning it lacks explicit frame labels. This innovative approach enables efficient sign language gesture recognition by harnessing the adaptability and effectiveness of deep learning models while reducing the need for labor-intensive manual labeling.

Driving the way in the field of word-level lipreading, [25] presented an imaginative methodology that rotates around a two-stream model. This model was painstakingly created to catch both static and dynamic elements by utilizing explicit CNN streams, each finely tuned with successful convolutional structures at the front-end. A careful assessment was done on two broadly perceived lipreading datasets to survey the presentation of the proposed model. The consequences of this thorough examination uncovered remarkable results - another best in class execution level on these difficult lipreading datasets, meaning a critical progression in word-level lipreading innovation.

Table 1 offers a consolidated overview of the referenced papers, presenting essential details such as authors, algorithms employed, features utilized, and the corresponding accuracy metrics

Table 1. A summary of research done on existing work

Work	**ML Algorithm**	**Features**	**Dataset**	**Accuracy Metric**
Joon Chung, Andrew Zisserman [1]	Watch, Listen, Attend and Spell (WLAS) network	13-dimensional MFCC features	Lip Reading Sentence (LRS) dataset	WER on GRID is 3.0% and on LRW is 23.8%

Yannis Assael, Nando De Freitas et al [2]	LipNet (end-to-end trainable model)	Spatiotemporal convolutions, recurrent network	GRID corpus (contains entire sentences)	CER is 6.4% and WER is 11.4%
Pingchuan Ma et al [3]	Densely-Connected Temporal Convolutional Network (DC-TCN)	Frame-wise features from a mouth Region-Of-Interest (ROI) encoder	LRW dataset	With Word boundary accuracy is 92.1% and without it is 94.1%
Huijuan Wang et al [4]	D Convolutional Vision Transformer (3DCvT)	Spatio-temporal features of continuous images	LRW and LRW-1000 (large-scale lip reading datasets)	accuracy is 88.5% for LRW and 57.5% for LRW-1000
Xing Zhao et al [5]	Mutual Information Maximization	Local features and global sequence features	LRW (Lip Reading in the Wild) dataset	accuracy is 84.41% for LRW and 38.79% for LRW-1000
Daqing Chen et al. [6]	For visual data processing, 3D convolution is used first, followed by 2D ResNet. For phoneme recognition, Transformers with multi-headed attention are used. For sequential data processing, an RNN is used as the language model.	Lip-reading issues include visual ambiguity, inadequate temporal resolution, efficient storing of spatial-temporal information, speaker reliance, head posture fluctuation, and lighting circumstances.	LRS2 (BBC Lip Reading Sentences)	WER is 40%, CER is 32% and PER is 30%

L Kumar et al. [7]	Deep learning models used for audio visual speech recognition Stack of CNN layers used for lip image classification	Spectrogram features used for audio signals representation. Mel-Frequency Cepstral coefficient (MFCC) and spectrogram commonly used.	LibriSpeech and Grid datasets	WER is 6.59% and WER is 4.5% when compared with SOTA is 4.5%
K Prajwal et al. [8]	Pooling process based on attention, VSD model	Visual speech representations , sub-word units	LRS2, LRS3, AVA-ActiveS peaker	Mean Average Precision is 88.2%
Dimitrios Tsourounis et al [9]	ALSOS ResNet: Residual Network	Spatiotemporal and spatial convolutions	Lip reading datasets in Greek and English languages	WRR is 87.01%
Xinshuo Weng et al. [10]	Two-stream I3D, Bi-LSTM	Grayscale video, optical flow	LRW (Lip Reading in the Wild)	Accuracy on LRW dataset is 84.07%
Tasuya Shirakata et al. [11]	GRU model, auto-encoder neural network	HP, Shape, Exp, AU (Action Unit-based)	OuluVS, CUAVE, CENSREC-1-AV	CENSREC-1-AV accuracy is 77.1% and that of OuluVS and CUAVE is 86.6% and 83.4% respectively
Hang Chen et al. [12]	Hierarchical pyramidal convolution (HPConv), selfattention	Multi-scale processing, spatial feature extraction	Lip Reading in the Wild (LRW) dataset	Accuracy on LRW is 86.83%
Liang Lin et al. [13]	Discriminativ ely trained And-Or graph model	Global shape features, spatial contextual features	Shape database	MeanAP on SYSU-Shape dataset is 0.539
Saakshi	Deep	CNN model	Dataset	

Bhosale et al [15]	Convolutional Neural Network (CNN) Model End-to-end deep learning architecture models.	for image-based problem Dataset includes distance of lips for pronouncing each word	includes words and lip distances for pronunciation . - Lip distances used for word segmentation and classification.	-
Guangxin Xing et al. [16]	LSTM encoder-decoder architecture, spatiotemporal convolutional neural network (STCNN), Word2Vec, Attention model	Mouth shapes, homophones	GRID, LRW, LRW-1000	WER on GRID, LRW and LRW-1000 is 4.40%,11.30% and 59.80% while accuracy on each is 95.60%,88.70% and 40.20%
Javad Peymanfard et al. [18]	AV-HuBERT model	Embedding vectors obtained from AVHuBERT model	Persian word-level lipreading dataset (244,000 videos), OuluVS dataset (phrases), LRW dataset (word-level lip-reading)	Micro Average accuracy and Macro Average accuracy on PLRW is 24.79% and 21.43%
Xubo Liu et al. [19]	Speech-driven lip animation model, VSR model	Lip images, speech utterances	LRS3, Librispeech, CelebA	WER on LRS3 + TTS-LBS-Synth is 32.9%
Marzieh Oghbaie et al. [20]	LCANet, Highway Networks, 3D	Spatial and sequential feature	Wild LRRo, Lab LRRo	WER for LRW is 14.3%, for LRS2 is 49.2% and for LRS3

	CNN, Bidirectional Gated Recurrent Units (Bi-GRU)	extractors		is 59.0%
Souheil Fenghour et al. [26]	Weighted Finite State Transducers (WFSTs), Hidden Markov Models (HMMs)	Visemes	Lip Reading Sentences in the Wild (LRS2)	WER for LRS is 18.0% and 48.3%
Ümit Atila et al [21]	Bi-LSTM (Bidirectional Long Short-Term Memory)	Features are extracted from video frames using pre trained CNN models. Feature vectors are obtained from specific layers of ResNet-18 and GoogLeNet models.	Two new datasets were created, one with 111 words and the other with 113 sentences, for Turkish lip-reading research	Word Accuracy is 84.09% and Sentence Accuracy is 88.55%
Souheil Fenghour et al [22]	Neural network-based lip reading system Transformer with a unique topology for classification of visemes	Purely visual cues, visemes as classes	BBC LRS2 dataset with 45839 sentences for training and 1243 sentences for testing	Visemes classification is over 95% and classification accuracy of words is 65.5%

1.4 Motivation

Our project is motivated by the imperative to enhance the precision and practical applicability of lip reading technology. Lip reading, while academically intriguing, holds significant promise in real-world scenarios such as aiding individuals with hearing impairments, advancing human-computer interaction, and augmenting security protocols. However, the inherent computational challenges within the realm of lip reading pose substantial obstacles to its effective deployment. To address these challenges, our project undertakes the foundational task of refining and meticulously organizing video data. This critical endeavor serves as a prerequisite for the development of robust lip reading systems capable of overcoming existing computational limitations.

Moreover, our project is underpinned by the recognition of a notable gap in the existing research landscape: the absence of adequate attention to Indian-accented English within the domain of lip reading technology. Given the diverse linguistic landscape and prevalence of Indian-accented English, particularly in the context of global communication and technological integration, addressing this oversight becomes imperative. Hence, in addition to our broader motivation to enhance the accuracy and utility of lip reading technology, we are specifically motivated by the opportunity to contribute to the development of solutions tailored to the nuances of Indian-accented English. By acknowledging and addressing this unique aspect, our project aims to foster inclusivity and effectiveness in lip reading technology, ensuring its relevance and accessibility across diverse linguistic contexts.

1.5 Aim of the Project

- Develop an automated lip-reading system using deep learning techniques to achieve precise speech recognition through analysis of lip movements.
- Address significant challenges related to accuracy, adaptability across different speakers and environments, and real-time performance capabilities.
- Focus on Indian-accent English to cater to a broader communication context and transcribe speech patterns specific to this linguistic variation.
- Design and refine a deep learning model that offers accurate transcription of

spoken words from lip movements, with seamless real-time operation and robustness to variations in speakers, environments, and accents.

1.6 Scope and Objectives

The scope of this project encompasses the development and implementation of an automated lip-reading system using deep learning techniques, with a specific focus on Indian-accented English. The project will involve leveraging state-of-the-art technologies to accurately transcribe spoken words from lip movements in real-time, addressing challenges related to accuracy, adaptability to diverse speakers and conditions, and real-time performance. Additionally, a unique aspect of this project involves the creation and utilization of an Indian-accented English dataset to tailor the lip-reading system to this linguistic variation.

Objectives:

- Data Collection and Annotation
- Feature Engineering and Selection
- Model Architecture Design
- Training and Optimization
- Performance Evaluation and Validation
- Robustness Testing and Adaptability
- Deployment and Integration
- Ethical Considerations and Privacy

1.7 Technical Approach

Our technical approach begins with a powerful neural network which we aim to train on the GRID and the Indian dataset for lip reading.

- **Data Preprocessing:** We start by cleaning and standardizing the GRID and the Indian dataset. This involves removing unwanted frames, focusing on the lip

regions, and ensuring all videos have the same size, color, and orientation. Clean data is essential for training the model effectively.We divide the cleaned dataset into two parts: a training set and a testing set. The training set is used to teach our model how to recognize and transcribe lip movements, while the testing set helps us evaluate its accuracy.

- **Training:** We feed the training data, which consists of video frames and corresponding transcriptions, into the neural network. The model learns to identify the lip regions in the frames and generate transcriptions from them by comparing its predictions with the actual transcriptions.

- **Validation and Fine-Tuning:** We continually check how well the model is doing on the testing set. If it's making errors, we fine-tune the model by adjusting its parameters and retraining it until it performs better.

- **Testing:** Once we're satisfied with our model's performance, we apply it to new, unseen videos to convert lip movements into text. The accuracy of these transcriptions tells us how well our lip reading system is working.

- **Deployment:** The deployment of our lip-reading system utilizes Streamlit, providing a user-friendly interface for interaction. The design of the interface allows users to easily select videos from a list of lip reading videos, which processes it through the model and generates text transcriptions from the lip movements. The predicted transcriptions, along with relevant information are then displayed to the user in an intuitive format.

CHAPTER 2

Architecture and Methodology

2.1 Architecture

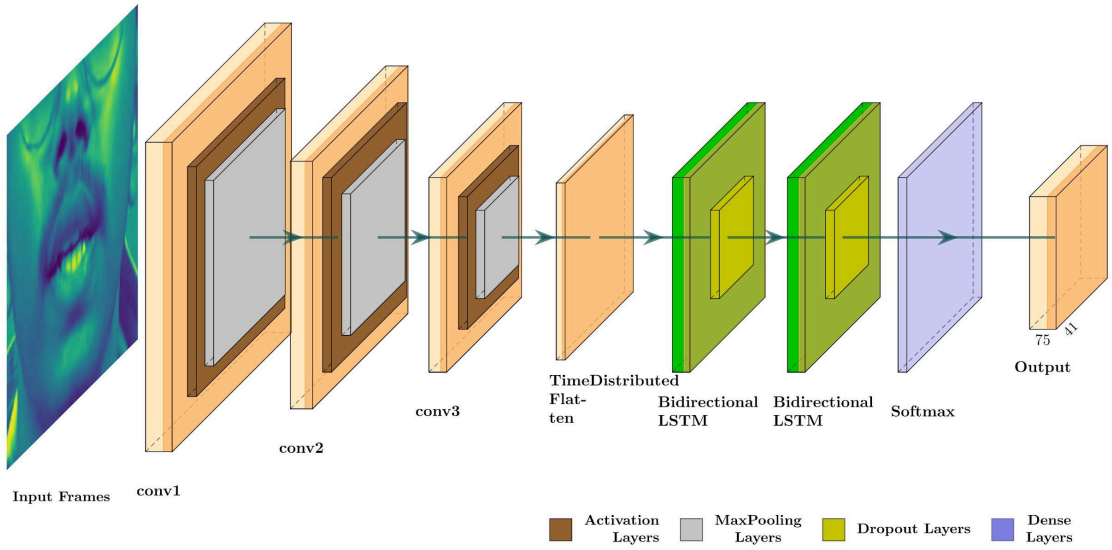


Fig. 1. Model Architecture

The model architecture as shown in figure 1, comprises three main components: The input layer which consists of 75 frames that undergo processing through three 3D convolutional layers. Subsequently, max pooling and ReLU activation are applied to extract spatial and temporal features. The spatial data is compressed and subsequently passed into two LSTM layers to extract temporal correlations. The softmax activation function is applied for classification or prediction purposes. Furthermore, the model undergoes training using CTC loss to enhance optimization.

2.1.1 Convolutional layers

Starting with a sequence of Conv3D layers, the model applies 3D convolutional filters to successive frames of the input video in order to collect spatial characteristics. 3D convolution layers are used due to its higher effectiveness in capturing spatiotemporal features. The output Y of a Conv3D layer can be mathematically stated as follows

$$Y = f(\sum_{i=1}^{N} W_i * X_i + b)$$

where f represents ReLU activation function, input X, W as convolutional filters and b is the bias which gets added during the computation. Conv3D layers are capable of processing video data by convolving across time as well as spatial dimensions, allowing them to capture spatiotemporal features and patterns present in the video. The above equation represents the process of sliding the convolutional kernel over the input feature map, multiplying the kernel values with the corresponding input values, and summing the results to compute the output feature map. Conv3D layers utilize 3D convolutional filters to process input volumes, enabling them to extract spatiotemporal properties from data such as video clips. Conv3D differs from standard 2D convolutions by incorporating a third dimension that can represent either depth or time, in addition to height and breadth. The additional dimension enables the network to handle data that has temporal properties, making it especially valuable for applications such as action detection, medical imaging analysis, or video processing.

During the process of convolution, a three-dimensional kernel moves across the input feature map, doing element-wise multiplication and then adding up the outcomes. Next, a bias term is included, and a non-linear activation function, usually ReLU, is employed to introduce non-linearity to the model. Utilizing the Rectified Linear Unit (ReLU) activation function aids in the removal of negative outputs, leading to more sparse

representations and facilitating the construction of deeper neural networks. Conv3D layers can be customized using several parameters such as kernel size, stride, and padding. The kernel size controls the size of the convolutional filter in each of the three dimensions. Stride regulates how much the filter moves with each step, and padding defines whether to add extra border pixels to maintain the input's dimensions. By altering these settings, Conv3D layers can be fine-tuned for specific applications, balancing spatial and temporal resolution. In many circumstances, Conv3D layers are used in conjunction with pooling layers to downsample the data, lowering dimensionality and computing load while keeping critical characteristics. Pooling can be done in both spatial and temporal dimensions, allowing the model to focus on the most prominent elements across time and space. Overall, Conv3D layers are effective tools for extracting complicated spatiotemporal characteristics, enabling a range of applications in video analysis, action identification, and other areas where understanding the evolution of data over time is crucial.

2.1.2 Temporal Encoding

After the convolutional layers, a flattened representation of the spatial information collected from each frame of the input video is created by applying a TimeDistributed layer to apply a Flatten operation across the temporal dimension. The integration of temporal context into later recurrent layers is made easier by this temporal encoding stage. This temporal encoding stage condenses the spatial characteristics retrieved from each frame into a sequential representation, thereby capturing the temporal dynamics of the lip motions across the whole video series.

A TimeDistributed layer helps assist this conversion by executing a single operation, such as Flatten, to each frame individually across the temporal dimension. The objective of the TimeDistributed layer is to preserve the temporal order of the frames while operating on each frame independently. This structure helps the model to maintain the integrity of the sequence while preparing it for temporal analysis. Once the spatial information has been flattened, it is transmitted to recurrent layers, which are designed to process sequences.

Recurrent layers, such as Long Short-Term Memory (LSTM) or Gated Recurrent Units (GRUs), are well-suited for handling time-series data because they can capture dependencies and interactions between frames. By feeding in the temporally encoded data, the recurrent layers can learn patterns, trends, and sequences that emerge over time.

The process of temporal encoding is critical because it transforms the spatially collected characteristics into a format that can be processed by recurrent models, enabling the capture of temporal dynamics and correlations. This is particularly beneficial in applications where the context and order of events are crucial, such as in lip reading, where small changes in mouth shapes and lip motions over time are key markers for detecting words or phrases.

2.1.3 Recurrent Layers

The heart of a lip-reading model consists of Bidirectional Long Short-Term Memory (BiLSTM) layers, which play a critical role in capturing the temporal relationships and sequential patterns inherent in lip motions throughout time. These BiLSTMs are an extension of the classic Long Short-Term recollection (LSTM) layers, which are designed to handle sequences of data by preserving a recollection of earlier inputs, helping to address the vanishing gradient problem that often affects Recurrent Neural Networks (RNNs). LSTMs achieve this using a sequence of gates—input, forget, and output—that control how information is stored and retrieved within the memory cells. This gating mechanism allows the LSTM to selectively keep significant information while forgetting irrelevant input, thereby making it more efficient at catching patterns over lengthy sequences. Bidirectional LSTMs, or BiLSTMs, enhance this functionality by processing input data in both forward and backward directions. This capacity is particularly helpful for lip reading, because context from both earlier and later frames is crucial for successful prediction. By having access to both past and future information, BiLSTMs can derive a more comprehensive knowledge of the temporal patterns within the lip movement sequences, leading to greater prediction accuracy.

After the BiLSTM layers, Dropout layers are frequently applied to assist mitigate overfitting. Dropout works by randomly removing a fraction of input units during training, which minimizes the model's dependency on specific features or nodes. This randomness promotes the model to generalize better to unseen input, as it learns to rely on redundant paths within the network. By eliminating overfitting, Dropout layers contribute to a more robust model that performs well on new, unexplored lip-reading challenges.

The final stage of the model involves a fully linked (dense) layer followed by a softmax activation function. The thick layer serves to transfer the outputs from the BiLSTMs into the required output space, in this case, the set of possible lip-reading classes. The softmax activation function is then used to turn these outputs into a probability distribution, indicating the likelihood of each class. The final output of the model is a probability distribution over the numerous lip-reading classes, including a special blank symbol that symbolizes no lip movement or quiet. This arrangement allows the model to effectively forecast the most likely class based on the temporal patterns it has learned, offering a stable foundation for lip-reading applications.

2.2 Methodology

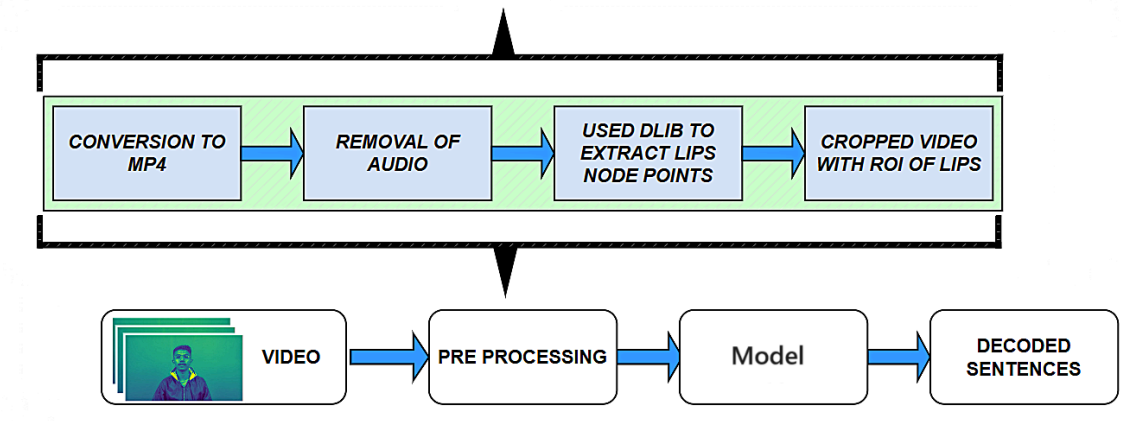


Fig. 2. Overview of the complete process

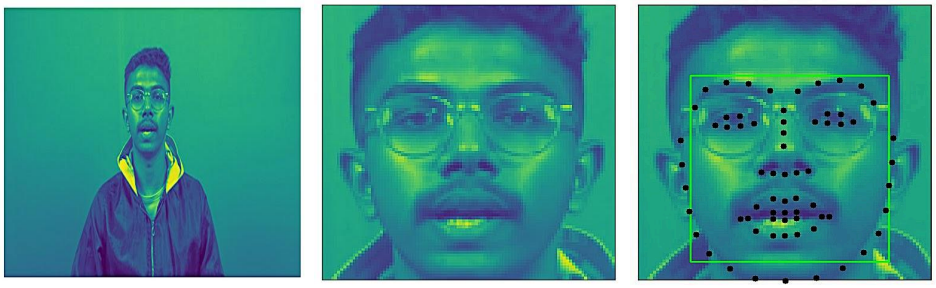
The video data preprocessing pipeline, depicted in Figure 2, ensures compatibility and computational efficiency through a series of meticulously planned steps. The process begins with converting video files to the widely accepted MP4 format using the ffmpeg tool, streamlining subsequent stages of processing. This format conversion not only ensures uniformity but also aids in optimizing storage and playback across different platforms. Once the videos are converted to MP4, the audio component is removed to focus solely on visual data, thus minimizing computational complexity and eliminating any potential auditory distractions (Figure 3.a).

This reduction in data size also contributes to faster processing times in later stages of the pipeline. Next, the pipeline implements accurate face detection using the Dlib library, renowned for its precision and reliability in detecting facial features regardless of the speaker's orientation or position within the frame (Figure 3.b). This step is critical as it anchors the subsequent operations on a stable reference point, allowing for consistent landmark extraction across varying video conditions.

With the face accurately detected, the pipeline proceeds to extract specific facial landmark points, with a keen focus on those that delineate the lips region. These landmarks serve as key reference coordinates for cropping the relevant area from each video frame (Figure 3.c). This precise extraction ensures that the lip movements, which are central to lip-reading, are isolated from other facial features, providing clear and focused input for the model.

Following the landmark extraction, the pipeline employs a cropping operation to isolate the lips region from each frame, using the previously derived landmark coordinates. The cropping process is vital in removing irrelevant visual data, such as background elements or non-lip-related facial features, thereby concentrating the model's attention on the area of interest. To account for variations in lighting and skin tones, the pixel values within the cropped lip region are normalized, ensuring consistent color representation across the

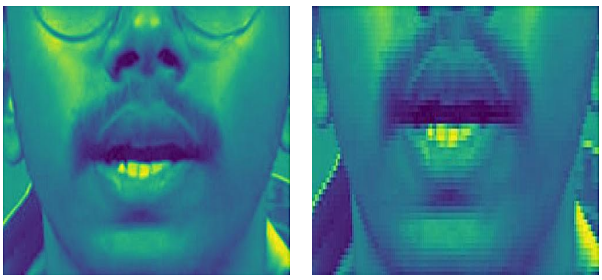
entire dataset. This normalization is crucial for maintaining data integrity and preventing discrepancies due to varying environmental conditions.



(a) (b) (c)

Fig. 3. (a) The initial frame of the video with no audio. (b) Accurate face detection irrespective of the position of the speaker. (c) Using Dlib library we mark out facial landmark points.

This cropping procedure effectively isolates the area containing the lips, thereby optimizing subsequent analysis and modeling efforts. Furthermore the cropped video based on the region of interest (ROI) delineated by the extracted lip landmark points is cropped to 140 x 46 pixels per frame for more accurate cropping of lips as shown in figure 4. This resizing step is essential in ensuring uniform input into the model, allowing it to focus on the temporal dynamics of the lip movements without additional computational overhead. The RGB channels in the entire training set are standardized to have a mean of zero and a variance of one unit.



(a) (b)

Fig. 4. (a) Cropped video frame in accordance with the lip landmark points. (b) Further cropping of the video frame to (140,46) pixels per frame.

Subsequently, videos are judiciously cropped based on the region of interest (ROI) delineated by the extracted lip landmark points, effectively enhancing focus on the pertinent lip region while eliminating irrelevant visual clutter. The processed video is subsequently forwarded to the model, where it undergoes analysis, resulting in the generation of a string of words corresponding to the sentence spoken by the speaker in

the input video. The final step in this preprocessing pipeline involves organizing the frames into sequences, effectively capturing the temporal aspect of the lip movements over time. This step is crucial for training recurrent neural networks or similar architectures that rely on sequential data for context and pattern recognition. Once the frames are organized into sequences, they are ready for model training and analysis, providing a clear and consistent representation of lip movements throughout the video.

In summary, this preprocessing pipeline is designed to deliver high-quality, consistent input data to the lip-reading model by focusing on key steps such as format conversion, audio removal, accurate face detection, landmark extraction, cropping, normalization, resizing, and temporal sequencing. These steps work in harmony to create a robust foundation for the lip-reading task, optimizing the model's performance while minimizing computational overhead.

CHAPTER 3

Dataset

A dataset of 500 videos of mp4 format were created for this project. These videos were recorded where the sentences were taken from the GRID dataset. Each of these 500 videos were precisely edited to have a duration of exactly three seconds. This meticulous process ensured the creation of a well-defined dataset focusing on the Indian English accent.

Each video after processing, consisted of 75 frames, indicating a frame rate of 25 frames per second. The corpus was intentionally distributed among a cohort of 15 or more participants. Each individual was tasked with recording videos in which they articulated a minimum of 20 distinct sentences. This deliberate allocation ensured a diverse range of speakers and content within the dataset, enhancing its richness and representativeness for subsequent analysis. In order to guarantee accuracy in terms of time, the synchronization of all 500 videos was methodically carried out on a frame-by-frame basis. The complex operation was performed using VSDC Free Video Editor, a flexible open-source software platform. Significantly, the length of each video was converted to exactly 3 seconds. Furthermore, the alignment of the videos was meticulously documented and stored in '.align' files, each corresponding to a specific video in the dataset.

These '.align' files contain the time duration of each word in milliseconds within the sentence, with silence denoting instances where no word is spoken at the beginning or at the end of the file.

Table 2 lists some of the main audio-visual datasets that have been utilized for lip-reading. These corpora are made up of hundreds of recordings of humans saying phrases using thousands of different words.

Table 2. Datasets related to Lip Reading

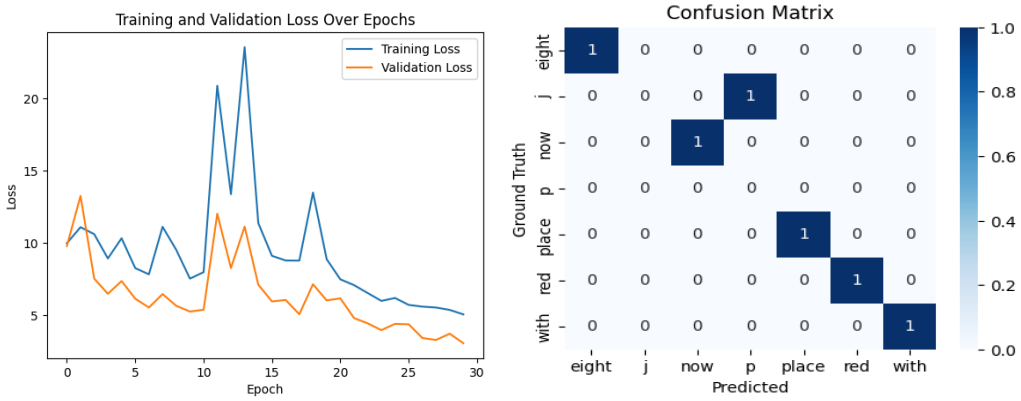
Dataset Name	**Number of classes**	**Number of videos**	**Segment**
LRW [1]	51	500,000+	Sentences
GRID [36]	51	33	Sentences
OuluVS2 [37]	6	80	Words
LRS2 [1]	N/A	100,000+	Words
LRS3 [1]	N/A	3,000,000+	Sentences
LRW-1000 [1]	51	1000	Sentences
GRID-480	51	480	Sentences
MVLRS [38]	N/A	41,222	Sentences
IBMViaVoice	290	24,325	Sentences
Indian Dataset	**-**	**500**	**Sentences**

In real-life circumstances, people may be speaking at various distances from cameras, in various lighting conditions, and with cameras of differing quality. These factors can cause considerable changes in video resolution. To ensure that audio-visual models can properly handle these issues, datasets have included movies of varying resolutions. The LRW [3] dataset, for example, is intended to imitate the variety observed in real-world lip reading events. It comprises films with resolutions ranging from high-definition to lower-quality visuals, akin to what you could see in surveillance footage or recordings shot with various sorts of equipment. Researchers hope to improve the resilience and adaptability of their models by integrating movies with varied resolutions. Models trained on such data learn to recognise and understand lip movements and speech cues over a wide range of visual features. This is critical for practical lip-reading applications such as voice recognition, accessibility technology, and spying.

CHAPTER 4

Results and Discussion

In our study, we evaluated the performance of our model on an Indian as well as GRID speaker dataset. The model trained for around 100 epochs on 2250 videos demonstrated a high degree of accuracy, achieving a character error rate (CER) of 4% and a word error rate (WER) of 15%. The CER is a metric that measures the percentage of incorrectly predicted characters in the transcriptions of spoken language. The confusion matrix (Figure 5.b) is plotted on Indian test dataset for predicted and ground truth sentences. A CER of 4% suggests that our model is able to transcribe characters with a high degree of precision, with only 4 characters out of every 100 being incorrectly predicted. Similarly, the WER is a metric that evaluates the percentage of incorrectly predicted words in the transcriptions. A WER of 15% indicates that our model is able to accurately transcribe 85 out of every 100 words.



(a) (b)

Fig. 5. (a) The plot illustrates the training and validation loss over 30 epochs. (b) Confusion matrix plotted for predicted and the ground truth sentences.

The validation loss of 3.043 (Figure 5.a) also suggests that our model is performing well and has been well-trained on the dataset. This low loss indicates that the model's predictions closely match the actual outcomes during validation. Overall, these metrics highlight the effectiveness of our lipreading model in transcribing speech from Indian speakers with a high level of accuracy.

CHAPTER 5

Conclusions

In conclusion, our project has been dedicated to the development of an automated lip-reading system utilizing deep learning techniques, with a particular focus on Indian-accent English. Through rigorous experimentation and simulation, we have achieved significant milestones and drawn several important conclusions. Firstly, we have successfully trained and refined a deep learning model capable of accurately transcribing spoken words from lip movements, achieving a word error rate (WER) of 15% and a character error rate (CER) of 4%. This represents a notable advancement in accuracy compared to previous systems. Furthermore, by incorporating Indian-accent English as a central aspect of our work, we have tailored our system to interpret and transcribe speech patterns specific to this linguistic variation, thereby enhancing inclusivity in communication tools.

Throughout the project, we have diligently addressed ethical and privacy considerations, ensuring responsible technology development and deployment practices. Alongside achieving measurable outcomes, we have also gained valuable insights into the complexities of lip-reading technology, including challenges related to accuracy, adaptability, and real-time performance.

Looking forward, the applications of our lip-reading system are vast. It holds significant potential for enhancing accessibility for individuals with hearing impairments, providing them with an effective communication tool. Moreover, the system can be applied in various human-computer interaction scenarios, enabling more seamless and intuitive interactions with technology. Additionally, in surveillance applications, the technology can aid in the analysis of video footage, improving the accuracy and efficiency of security systems. In conclusion, our project has not only produced tangible outcomes in terms of accuracy and adaptability but has also provided valuable insights and applications in the domains of accessibility, human-computer interaction, and surveillance.

CHAPTER 6

Future Scope

Despite our efforts to train on diverse datasets, including Indian-accented English, our system may struggle to accurately transcribe speech from lip movements, particularly when faced with diverse linguistic patterns. Processing large volumes of data in real-time presents computational challenges, potentially impacting the system's performance in applications requiring prompt responses. To refine our work, continuous optimization of the deep learning model, supplemented by dataset expansion and architecture fine-tuning, is crucial. Additionally, exploring multimodal approaches and advancements in hardware technology offer promising avenues for enhancing accuracy and scalability. Through these strategies, we aim to improve the effectiveness and applicability of our system across diverse linguistic and computational contexts.

- CCTV Integration: Incorporating CCTV footage as additional data could enrich the training process and improve the system's effectiveness across diverse environments.
- Training on Diverse Data and Languages: Expanding the dataset to include varied speakers, accents, and languages is crucial for enhancing adaptability.
- Technological Advancements: Continued advancements in hardware technology are anticipated to alleviate scalability concerns and facilitate real-time performance.

In reflection, while we have achieved a functional prototype and demonstrated feasibility, refinement is necessary, particularly in terms of accuracy, scalability, and adaptability. This serves as a guide for future researchers, motivating them to build upon our work and address its limitations, ultimately advancing the field of automated lip-reading and benefiting accessibility, communication, and technology.

References

1. Chung, J.S., Zisserman, A. (2017): Lip Reading in the Wild. In: Lai, SH., Lepetit, V., Nishino, K., Sato, Y. (eds) Computer Vision – ACCV 2016. ACCV 2016. Lecture Notes in Computer Science(), vol 10112. Springer, Cham.
2. Yannis, M., Assael., Brendan, Shillingford., Shimon, Whiteson., Nando, de, Freitas. "LipNet: End-to-End Sentence-level Lipreading." arXiv: Learning, undefined (2016).
3. P. Ma, Y. Wang, S. Petridis, J. Shen and M. Pantic, "Training Strategies for Improved Lip-Reading," ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, Singapore, 2022, pp. 8472-8476, doi: 10.1109/ICASSP43922.2022.9746706.
4. H. Wang, G. Pu and T. Chen, "A Lip Reading Method Based on 3D Convolutional Vision Transformer," in IEEE Access, vol. 10, pp. 77205-77212, 2022, doi: 10.1109/ACCESS.2022.3193231.
5. X. Zhao, S. Yang, S. Shan and X. Chen,"Mutual Information Maximization for Effective Lip Reading," 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020), Buenos Aires, Argentina, 2020, pp.420-427, doi: 10.1109/FG47880.2020.00133.
6. El-Bialy, R., Chen, D., Fenghour, S., Hussein, W., Xiao, P., Karam, O. H., & Li, B. (2023). Developing phoneme-based lip-reading sentences system for silent speech recognition.CAAI Transactions on Intelligence Technology, 8(1), 129-138. https://doi.org/10.1049/cit2.12131
7. L Ashok Kumar, D Karthika Renuka, S Lovelyn Rose, M C Shunmuga priya, I Made Wartana, Deep learning based assistive technology on audio visual speech recognition for hearing impaired, International Journal of Cognitive Computing in Engineering, Volume 3, 2022.

8. Prajwal, K., R., Triantafyllos, Afouras., Andrew, Zisserman. (2021). Sub-word Level Lip Reading With Visual Attention. arXiv: Computer Vision and Pattern Recognition
9. Lip Reading by Alternating between Spatiotemporal and Spatial Convolutions. J. Imaging 2021, 7, 91. https://doi.org/10.3390/jimaging7050091.
10. Weng, X., & Kitani, K. (2019). Learning Spatio - Temporal Features with Two-Stream Deep 3D CNNs for Lipreading. ArXiv. /abs/1905.02540.
11. Shirakata, Tasuya, and Takeshi Saitoh. "Lip reading using facial expression features." Int. J. Comput. Vis. Signal Process 1.1 (2020): 9-15.
12. Chen, H., Du, J., Hu, Y., Dai, L., Lee, C., & Yin, B. (2020). Lip-reading with Hierarchical Pyramidal Convolution and Self-Attention. ArXiv./abs/2012.14360
13. L. Lin, X. Wang, W. Yang and J. -H. Lai work on "Discriminatively Trained And-Or Graph Models for Object Shape Detection," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 37, no. 5, pp. 959-972, 1 May 2015, date of issue (doi) : 10.1109/TPAMI.2014.2359888.
14. N. Deshmukh, A. Ahire, S. H. Bhandari, A. Mali and K. Warkari, "Vision based Lip Reading System using Deep Learning," 2021 International Conference on Computing, Communication and Green Engineering (CCGE), Pune, India, 2021, pp. 1-6, date of issue (doi): 10.1109/CCGE50943.2021.9776430
15. Bhosale, Saakshi, et al. "An Application to Convert Lip Movement into Readable Text."
16. Xing, G., Han, L., Zheng, Y., & Zhao, M. (2023). Application of deep learning in Mandarin Chinese Lip Reading recognition. EURASIP Journal on Wireless Communications and Networking, 023(1), 1- 14. https://doi.org/10.1186/s13638-023-02283-y
17. Jin Ting, Chai Song, Hongyang Huang, Taoling Tian, A Comprehensive Dataset for MachineLearningbased Lip-Reading Algorithm, Procedia Computer Science, Volume 199, 2022.
18. J. Peymanfard, A. Lashini, S. Heydarian, H. Zeinali and N. Mozayani, "Word-level Persian Lipreading Dataset, "2022 12th International Conference on Computer and Knowledge Engineering (ICCKE), Mashhad, Iran, Islamic

Republic of, 2022, pp. 225-230, date of issue : 10.1109/ICCKE57176.2022.9960105.

19. X. Liu et al., "SynthVSR: Scaling Up Visual Speech Recognition With Synthetic Supervision," 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 2023,pp. 18806-18815, doi:10.1109/CVPR52729.2023.01803.
20. Oghbaie, M., Sabaghi, A., Hashemifard, K., & Akbari, M. (2021). Advances and Challenges in Deep Lip Reading . ArXiv. /abs/2110.07879.
21. Ü. Atila, F. Sabaz, Turkish lip-reading using Bi-LSTM and deep learning models. Eng. Sci. Technol. Int. J. 35, 101206 (2022).
22. S. Fenghour, D. Chen, K. Guo and P. Xiao, "Lip Reading Sentences Using Deep Learning With Only Visual Cues, " in IEEE Access, volume. 8, pp. 215516-215530, 2020, date of issue (doi): 10.1109/ACCESS.2020.3040906.
23. Rudregowda S, Patil Kulkarni S, H L G, Ravi V, Krichen M. Visual Speech Recognition for Kannada Language UsingVGG16 Convolutional Neural Network Acoustics. 2023; 5(1):343-353.
24. Czyzewski, A., Kostek, B., Bratoszewski, P. et al. An audio-visual corpus for multimodal automatic speech recognition. J Intell Inf Syst 49, 167–192 (2017)
25. Li H, Yadikar N, Zhu Y, Mamut M, Ubul K. Learning the Relative Dynamic Features for Word-Level Lipreading. Sensors. 2022; 22(10):3732. https://doi.org/10.3390/s22103732
26. Fenghour, S., Chen, D., Guo, K., & Xiao, P. (2020). Disentangling Homophemes in Lip Reading using Perplexity Analysis. ArXiv. /abs/2012.07528
27. K. Thangthai and R. Harvey, ‘‘Improving computer lipreading via DNN sequence discriminative training techniques,’’ in Proc. Interspeech, Aug. 2017, pp. 1–5.
28. F. Tao and C. Busso, "End-to-End Audiovisual Speech Recognition System With Multitask Learning,"in IEEE Transactions on Multimedia, vol. 23, pp. 1- 11,2021, doi: 10.1109/TMM.2020.2975922.
29. G. Sterpu, C. Saam and N. Harte, "How to Teach DNNs to Pay Attention to the Visual Modality in Speech Recognition," in IEEE/ACM Transactions on Audio,

Speech, and Language Processing, volume .28, pp.1052-1064, 2020, date of issue (doi): 10.1109/TASLP.2020.2980436.

30. Akhter N, Ali M, Hussain L, Shah M, Mahmood T, Ali A, Al-Fuqaha A. Diverse Pose Lip-Reading Framework. Applied Sciences. 2022; 12(19):9532. https://doi.org/10.3390/app12199532
31. Tao, Fei, and Carlos Busso. "Bimodal Recurrent Neural Network for Audiovisual Voice Activity Detection." INTERSPEECH. 2017.
32. Stafylakis .T & Tzimiropoulos G. (2017). Combining Residual Networks with LSTMs for Lipreading. ArXiv. /abs/1703.04105
33. Noda, Kuniaki, "Audio-visual speech recognition recognition using deep learning." Applied intelligence 42 (2015): 722-737
34. Matthews, I. et al. “Extraction of Visual Features for Lipreading.” IEEE Trans. Pattern Anal. Mach. Intell. 24 (2002): 198-213.
35. Koller, Oscar et al. “Deep Learning of Mouth Shapes for Sign Language.” 2015 IEEE International Conference on Computer Vision Workshop (ICCVW) (2015): 477-483
36. M. Cooke, J. Barker, S. Cunningham, and X. Shao. An audio-visual corpus for speech perception and automatic speech recognition. The Journal of the Acoustical Society of America, 120(5):2421–2424, 2006.
37. I. Anina, Z. Zhou, G. Zhao and M. Pietikäinen, "OuluVS2: A multi-view audiovisual database for non-rigid mouth motion analysis," 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), Ljubljana, Slovenia, 2015, pp. 1-5, doi: 10.1109/FG.2015.7163155.
38. Chung, J., & Zisserman, A. (2017). Lip reading in profile. Ritish Machine Vision Conference, 2017.
39. T. Afouras, J. S. Chung, A. Zisserman Deep Lip Reading: A comparison of models and an online application INTERSPEECH, 2018.