2. LITERATURE STUDY

This chapter carries out a detailed analysis of the work that has been carried out in Automatic Lip Reading (ALR). The first section talks about the foundation and evolution of ALR. The second section explains the 4-step ALR process. The next section explains about 3-step ALR process based on deep learning. Section four dives into CNN based ALR. Section 5 talks about the datasets that have been created for ALR process in various spoken languages. Finally, the last two sections describe the works that have been carried out for Indian languages and challenges in ALR.

2.1 THE FOUNDATION OF AUTOMATIC LIP READING

In audio-video communication, both visual and voice signals carry information. While voice signals often carry more detailed information, the integration of visual signals enhances understanding and context. The efficiency and speed of today's wireless communication systems are largely driven by advancements in processing and transmitting these signals. Voice channels carry one-dimensional information, while visual channels carry two-dimensional information in the form of images or videos. Visual Speech Recognition (VSR) is a technology that interprets speech signals by analysing visual cues, such as the movements of the face, lips, tongue, and teeth. This technology has broad applications, including improving speech recognition in noisy environments and assisting individuals with hearing impairments. A specific and crucial aspect of VSR is Automatic Lip Reading (ALR), which focuses on understanding speech by observing the speaker's lip movements. Speech signals can easily be inferred in noisy environment while visual information needed for ALR will not be affected. My research concentrates on this subset of VSR, utilizing machine learning techniques to enhance the accuracy and effectiveness of ALR.

As discussed earlier, in 1954, Sumby and Pollack demonstrated that visual information, such as facial and lip movements, enhances speech intelligibility, laying foundational work for later developments in Automatic Lip Reading (ALR) [18]. Building on these early insights, the research field expanded significantly. For instance, in 1981, Don Pearson's work in 'Visual Communication Systems for the Deaf' explores the development and implementation of technologies designed to aid communication for the deaf community.

Pearson's research highlights the importance of visual cues in understanding speech, which is fundamental to the advancement of ALR systems [28].

In 1984, Petajan developed the first Audio-Visual Automatic Speech Recognition (AV-ASR) system, where he collected video recordings along with audio to capture lip movement of speakers. Various image processing techniques were used for frame extraction, Region of Interest (ROI) detection, image enhancement, and feature extraction. Different neural network architectures were tested and trained to learn the relationship between visual and audio features. This methodology demonstrated that visual information extracted from lip movements could significantly enhance speech recognition accuracy, even in noisy environment [25]. This work laid the foundation for future advancements in ALR systems. Before discussing the work done in lip reading, first explore the different approaches to Automatic Lip Reading (ALR)

In the 21st century, with the evolution of Artificial Intelligence and Deep learning methods, these can be classified into the following main four approaches:

1. **Traditional approach:** This approach was developed when only basic image processing methods and statistical methods were available. They are further divided into the following two categories.
    a. **Feature-Based Method:** In this method, image processing techniques are used for feature extraction from lip movements like lip contour, shape and motion.
    b. **Model-Based Method:** In this method, different statistical models are used to model temporal features of lip movements.
2. **Machine learning approach:** This approach was developed with the evolution of Artificial Intelligence, which is further divided into two categories based on the type of data used.
    a. **Supervised Learning:** In this method labeled data are used to train the machine and Support Vector Machine or Neural Network techniques are used.
    b. **Unsupervised Learning:** Here unlabeled data are used for clustering and learning patterns.
3. **Deep Learning approach:** With the further advancement in AI, machine learning, and computer vision, deep learning methods developed. In this approach, further three approaches are there.
    a. **Convolution Neural Networks (CNNs):** Different Levels of CNNs are used for feature extraction from image sequences of lip movements.

b. **Recurrent Neural Network (RNNs) and Long Short-Term Memory (LSTMs) Networks:** Capture temporal dependencies in image sequences of lip movements.
c. **Hybrid Models:** Combination of CNNs and RNNs are used, CNNs for feature extraction and RNNs/LSTMs for Temporal dynamics.

4. **Multimodal Approach:** Generally, this approach is used in noisy environment. Audio and visual inputs both are integrated to enhance recognition rate.

Traditional and machine learning approaches for Automatic Lip Reading both use similar feature extraction techniques, such as color segmentation, texture analysis, and edge detection. However, traditional methods rely on handcrafted rules and simple algorithms, while machine learning-based methods rely on manually engineered features extracted from the data, providing greater adaptability to variations in lip appearances. In contrast, deep learning-based methods, particularly using Convolutional Neural Networks (CNNs), automatically learn hierarchical features directly from the raw data, integrating feature extraction and classification into a single end-to-end model, leading to higher adaptability and accuracy.

In traditional and machine learning approaches, Automatic Lip Reading (ALR) involves the following sequences:

1. **Lip Detection and Extraction**
2. **Feature Extraction**
3. **Feature Transformation**
4. **Classification**.

On the contrary, deep learning approaches combine these steps into three main stages:

1. **Lip Detection and Extraction**
2. **Front-end (Feature Extraction and Transformation)**,
3. **Back-end (Classification)**.

In the following section, firstly, the four-stage process of Automatic Lip Reading (ALR) has been discussed followed by traditional and machine learning approaches. At last, the process followed by deep learning approaches has been outlined.

2.2 4-STEP ALR PROCESS

2.2.1 Lip Detection and Extraction

The first step of Automatic Lip Reading (ALR) is lip detection and extraction. In this step, the lip area or ROI is extracted from the given image or video and then passed on for further processing. Pre-processing steps significantly affects the overall performance of ALR recognition. Different approaches employ various methods for this task.

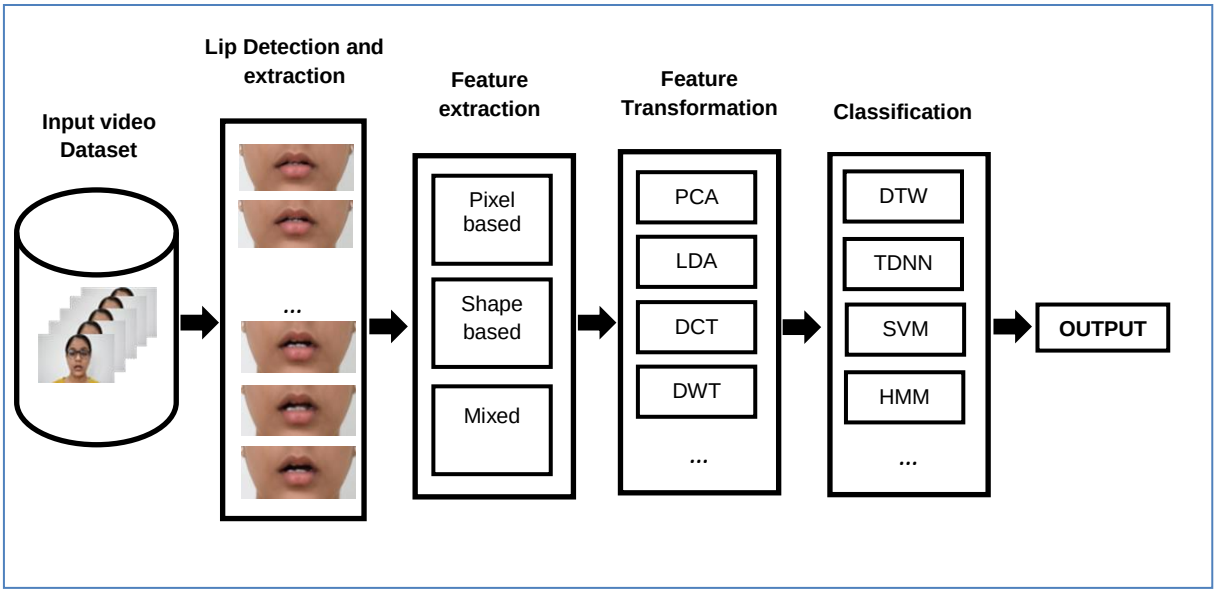


Figure 1 ALR Process based on image processing methods

Pre-processing Steps for Lip Detection and Extraction:

1. **Face Detection**: Detect the face within an image or video to narrow down the search area for the lips.
2. **Facial Landmark Detection**: Identify key facial landmarks, specifically those around the lip area, after detecting the face.
3. **ROI Extraction**: Extract the region of the image that contains the lips based on the detected facial landmarks.
4. **Image Resizing and Normalization**: Resize the extracted lip region to a fixed size and normalize pixel values to standardize the input for the neural network.

Various techniques based on pixel information, structure, shape, template, and model are used for pre-processing, ensuring that lip images remain in a consistent and suitable form for subsequent feature extraction and classification.

Pixel Information-based method

Pixel information-based methods utilize various parameters such as color and intensity values of individual pixels to detect and extract lips. The process involves several steps.

In **colour space transformation,** lip detection starts with transformation of image from RGB to different colour space like YCbCr, HSV, YIQ, and many more. In this transformation lip region can be more easily distinguished from the surrounding skin and other facial features. Once colour space is defined, next step is **pixel classification** based on colour value (intensity) using Histogram analysis or Thresholding. Next step is **region grouping,** which involves grouping of lip region pixels from the rest of the face. Techniques like edge detection and contour fitting applied for smoothing and geometric constraints for shaping and sizing. Finally **masking and cropping** are applied to extract lips from the face region.

Machine learning approaches often influence geometric properties of the face for lip detection, incorporating color segmentation and texture-based methods like thresholding, histogram analysis, Gabor filtering, and Local Binary Patterns (LBP) into the pixel-based approach.

While these methods are simple and fast, they may encounter challenges with changing lighting conditions, different skin tones, and variations in lip color due to makeup.

Several studies have explored different approaches within pixel information-based methods:

- **Positioning Method**: Wark et al. [29] introduced a lip detection method based on the R/G ratio, where points outside a certain range of R/G ratio are excluded from the ROI.
- **Red Exclusion Algorithm**: Lewis et al. [30] proposed a red exclusion algorithm to distinguish between skin and lip regions based on differences in the green and blue components of pixel values.
- **Color Contrast and Clustering**: Skodras et al. [31] utilized color contrast and K-clustering methods for detecting key points of lips.
- **Fuzzy C-Means Clustering**: Ghaleh et al. [32] employed the fuzzy c-means clustering method in the RGB color space to extract the ROI.
- **Color Transformation**: Gritzman et al. [33] experimented with 33 color transformation methods for lip segmentation and found HSV to be the most effective for this task.

(i) Structure-based method

In the structure-based method, the process begins with face detection using techniques like the Viola-Jones Face detector with Haar-like features.

Viola-Jones Face detector is developed by Paul Viola and Michael Jones in 2004 [34]. This method efficiently identifies facial features and serves as the foundation for **locating the lip region**. The lip region is located according to the distribution properties of each facial organ. Edge detection methods like Canny edge detector, Sobel edge detector are used to **identify boundaries of Lip area**. Contour tracing and Snake methods are used to identify continuous boundaries detected in edge detection [34].

Structure-based methods focus on geometric features, such as the relative positions of the eyes, nose, and mouth, to locate the lip area by utilizing the structural characteristics of the lips. Statistical models like AAM (Active Appearance Model) and ASM (Active Shape Model), along with Shape predictors and Regression trees, are used for facial landmark detection.

Unlike pixel-based methods, which rely on colour information, structure-based methods utilize geometric properties, rendering them less sensitive to changes in lighting conditions and more adaptable to different skin tones.

However, facial occlusions such as glasses or facial hair may interfere with detection accuracy. Additionally, excessive computational requirements can increase the complexity of the method.

Mita et al. [35], and Wang et al. [36] have used Haar-like features with the AdaBoost cascade classifier for face detection. Puviarasan and Palanivel [37] have used top, height, left, and width values to locate the face and lower part of face is detected as a mouth region using certain mathematical calculations. In figure H_f, H_m, W_f, W_m are height of Face, Height of mouth, Width of Face and width of mouth respectively.

Figure 2 Face structure based method for lip detection

(ii) Model-based method

Model-based approach is used to locate ROI according to shape and appearance of detected lips. Both traditional and machine learning approaches for model-based lip detection and extraction use predefined models to represent lip shapes and appearances. Traditional methods rely on statistical models like ASMs and AAMs or template matching, while machine learning approaches leverage models like Haar cascades, Dlib's 68-point model [27], Support Vector Machine (SVM) with Histogram of Oriented Gradients (HOG), CNNs, regression trees, and Generative Adversarial Network (GAN) to learn complex patterns and improve detection accuracy. Traditional methods typically use handcrafted features and predefined models, whereas machine learning methods learn features directly from data, offering greater flexibility and robustness.

In traditional approach, model based methods use predefined models to represent the shape and appearance of the lips. ASMs are statistical model that capture the variability of lip shapes and use facial landmarks to define the shape of lips. AAMs extend ASMs by incorporating texture information or appearance along with the shape. In template matching predefined templates of lip shapes are used for matching. Luettin and Thacker [38] have applied ASM for lip detection and extraction while, Nguyen and Milgram et al. [39] have proposed multi feature ASM. Main drawback of single feature ASM is that it can fail into local minimum when speakers have beard, wrinkles, and low contrast skin tones. Cootes et al. [40] and Rothkrantz et al. [41] have implemented AAM for lip extraction.

ACM model which is also known as Snake model is proposed by Kass et al. [42], which was based on energy minimizing spline guided by external forces and influenced by image forces. Image forces help Snakes to lock onto nearby lines, edges, and contours and localize it

In the **machine learning approach**, model-based methods use trained models to detect and extract lips from images. These models are often more flexible and can learn complex patterns from large datasets.

Haar cascades use a series of classifiers trained with positive and negative samples to detect objects in images. For lip detection, they can be trained to recognize facial features including the mouth region. **Dlib's 68-point model** identifies 68 key points on the face, which can be used to precisely locate the lips. Both the methods use pre-trained models. **SVM with HOG** combining used for lip detection (SVM) and feature extraction (HOG).

A **CNN-based model** can be trained to detect and extract lips by providing large datasets. **Regression trees** and **Shape predictors** are used to predict specific landmarks on the face, such as corners and edges of lips. Kazemi and Sullivan [43] used regression trees for facial landmark detection, providing accurate lip localization.

2.2.2 Feature Extraction and Transformation

Feature extraction is the task of identifying and isolating relevant features from ROI for further analysis or classification. Raw lip images need to be processed to capture important information and patterns.

For feature extraction and transformation task, traditional approach depends on predefined methods and machine learning approach learns this process from the data.

In traditional approach feature extraction methods can be classified into pixel-based methods, shape-based methods, and mixed feature extraction methods.

Pixel-based methods: All pixels of the ROI are considered as the original feature space. Various methods are used to reduce the dimension of this feature space to extract important features. **Linear transformation methods** transform the lip feature vector, remove noise, and reduce its size. These methods include Principal Component Analysis (PCA) [44], [45], Discrete Cosine Transform (DCT) [46], Discrete Wavelet Transform (DWT) [37], [47], [48], Linear Discriminant Analysis (LDA) [49], Local Sensitive Discriminant Analysis (LSDA) [48], [50], and Maximum Likelihood Linear Transformation (MLLT) [51]. The main task of these methods is to reduce the size of the lip feature vector by removing redundant and non-useful information.

Intra-frame linear transformation extracts visual language information from a single image, while inter-frame linear transformation extracts information of lips between video

frames dynamically. Applying both methods together provides effective space-time information. Hierarchical Linear Discriminant Analysis (HILDA) is a two-stage approach introduced by Potamianos et al. [52]. In the first stage, transformations are applied to audio-only and visual-only features. In the second stage, both features are concatenated into a single modality feature for further processing.

Pixel-based methods can limit recognition accuracy as they use all pixel information, making them sensitive to changes in skin color, illumination, rotation, and scaling. To overcome these problems, **local pixel feature methods** perform linear transformations on original pixel values. One of the basic algorithms for local pixel transformation is Local Binary Patterns (LBP), which can process single two-dimensional images only, while lip-reading images are in sequence. Zhou et al. [53] introduced a variant of LBP named LBP-TOP (Local Binary Pattern from Three Orthogonal Planes) [54], which was used to extract spatiotemporal information.

Another pixel-based method is the optical flow method. The **optical flow method** tracks the changes of pixels in the time domain by comparing the current frame to the preceding frame. This approach is commonly used in studies [55], [56], [57] for ALR. However, the main disadvantage of this method is its computational intensity, requiring substantial calculation. Additionally, it is sensitive to changes in the speaker's position and lighting conditions.

Shape-based methods build a model based on the shape of the lips and other key parameters while speaking. Shape-based methods can be classified based on geometric features and contour features. Geometric features like height, width, perimeter, area, and shape of the lip contour are used. Ma et al. [58] used six lip points to extract five geometric features of lips. Contour features, like edges of the lips, are used to locate key feature points of lips. The Active Contour Model (ACM), also known as the snake model [42], is based on the point distribution model. In 1997, the ASM was used for the first time for ALR by Luettin and Thacker [38].

Shape-based methods are more accurate than pixel-based methods. They offer good controllability and interoperability irrespective of illumination conditions, skin color, or lip rotation and scaling. However, extracted features are based on the shape of the lips, which may cause information loss. This model requires high-quality images, making calculations complex and time-consuming [59].

Mixed feature extraction methods combine shape-based and pixel-based approaches. The Active Appearance Model (AAM) is a mixed feature extraction method proposed by Cootes et al. [40] in 2001. In 2016, Watanabe et al. [60] proposed 3D AAM from three perspectives—front, left, and right—which was capable of recognizing lip images from any angle. The AAM method is highly accurate but requires a greater number of iterations to calculate feature parameters. Local optimization is also a drawback of AAM [60].

In machine learning approaches, feature extraction involves identifying and selecting important attributes or characteristics from raw data that can be used for classification or regression tasks. PCA, LDA, DCT and DWT, HOG, and optical flow are used for feature extraction and transformation in machine learning approach

In both traditional and machine learning approaches, the methods for feature extraction and transformation often overlap. Machine learning approach emphasizes the use of data-driven models automated learning from labelled data. Techniques like PCA and LDA, and others focus on not just to reduce dimensionality but also to enhance the model's ability to generalize from the data. Machine learning models are mode adaptable to variations in lip appearance, lighting conditions, and other factors because they learn from data. This leads to greater adaptability and better performance in real-world scenarios.

2.2.3 Classification

In both traditional and machine learning approaches to ALR, various classification methods are used to recognize visual features extracted from lip movements. Traditional approaches rely on methods like template matching and HMMs, while machine learning approaches leverage data-driven models such as SVM, RF, and ANNs to optimize feature extraction and classification. Here, different classifiers used in traditional and machine learning approach both have been discussed.

Template matching methods

The Template matching method is a very simple and traditional method of classification. It uses static images to extract features and match with existing templates. This method was used by Petajan [25] for lip reading classification. As each person has different speed of speaking resulting into different dynamic features of pronunciation, Petajan et al. [61] tried to ease this problem by introducing Dynamic time-warping.

DTW is used to measure similarity between temporal sequences which may vary in speed. It aligns sequences of visual features extracted from lip movements to a template, compensating for variations in speaking speed. DTW is effective for isolated word recognition but has limitations in continuous speech recognition [62].

Artificial Neural Network

ANNs are used in traditional approach and machine learning approach both for classification. In the traditional approach, Artificial Neural Networks (ANNs) were used primarily as shallow networks due to computational and hardware limitations. They were often used in a simple, straightforward manner, without extensive layers or sophisticated architectures. A classic application of ANNs in traditional approaches is the Time-Delay Neural Network (TDNN), which captures temporal dependencies in sequential data. In Automatic Lip Reading (ALR), TDNNs process lip image sequences to recognize dynamic lip movements. However, traditional TDNNs were constrained by limited computational power, resulting in simpler models with restricted performance.

In the machine learning approach, ANNs have evolved into more sophisticated architectures, often involving multiple layers and advanced training techniques. With the advent of improved computational resources and large datasets, ANNs in machine learning can be deeper and more complex, leading to better performance.

Modern ANNs in machine learning utilize multiple hidden layers to capture abstract features and complex patterns. Techniques like backpropagation, dropout, and batch normalization enhance their training and generalization capabilities.

TDNNs are still used today but have been enhanced with more layers and sophisticated training methods. They now belong to a broader family of recurrent neural networks (RNNs) and other sequence models, offering improved handling of temporal dependencies in lip movements. Modern techniques such as backpropagation, dropout, and batch normalization further enhance their performance and generalization capabilities [63], [64].

Modern ANNs in machine learning achieve significantly better performance by training deeper networks with more data. They generalize better and handle complex variations in lip movements and appearances, resulting in higher accuracy in ALR.

Hidden Markov Model

Hidden Markov Model is widely used in ALR Traditional and machine learning approach for classification. HMM was proposed by Baum in 1960 and applied for speech recognition. Later it started to use for ALR also. The basic idea is to model the lip movement as a sequence of observable states that correspond to the semantic information of speech. HMMs represent the temporal sequence of lip movements using states and transition probabilities. Each state represents a particular configuration of lip shapes or positions, and transitions between states capture the dynamic nature of speech.

In machine learning approaches, HMMs are enhanced with more sophisticated techniques and integrated with other models to improve performance. Machine learning approaches use supervised learning with large labelled datasets and methods like backpropagation to train the DNN components of HMMs, improving their accuracy. Moreover, Machine learning leverages advanced RNNs, such as LSTMs, to better model temporal dependencies, enhancing the effectiveness of HMMs in capturing lip movement sequences.

Sujatha and Krishnan [65] applied Discrete Cosine Transform (DCT) to extract dynamic visual features, which were then utilized to estimate the parameters of an HMM. During testing, the features of the test image sequence were inputted to obtain prediction results. Similarly, Thangthai et al. [66] employed a DNN-HMM model to represent the extracted features over time. Unlike traditional GMM-HMM, they replaced the generation probability estimation with a DNN, where each unit's output represents the posterior probability of a state.

Overall, while traditional HMMs laid the foundation for modeling temporal sequences in ALR, machine learning approaches have significantly enhanced their capabilities, leading to more accurate and robust ALR systems

Support Vector Machine (SVM) is a popular machine learning algorithm for classification tasks. It finds the hyperplane that best separates different classes in the feature space. In ALR, SVMs are used to classify visual features extracted from lip movements into different phonetic categories. SVMs can handle non-linear relationships using kernel functions, making them suitable for complex patterns in ALR data.

Random Forests is an ensemble learning method that constructs multiple decision trees during training and outputs the mode of the classes for classification. It is robust to

overfitting and can handle high-dimensional data, which is useful for ALR where visual features can be complex and numerous.

Multi-layer Perception is a type of feedforward artificial neural network that consists of multiple layers of neurons. MLPs are used to classify lip features by learning complex patterns in the data. They are versatile and can be applied to various classification tasks in ALR.

2.3 3-STEP ALR WITH DEEP LEARNING APPROACH

Traditional and machine learning approaches rely heavily on manually crafted features. Handcrafted features might not generalize well across different speakers, lighting conditions, and environments, reducing the robustness and accuracy of the models. Simple models of traditional approach like HNNs may not capture complex dependencies and variations in lip movements. Early ANNs like TDNN were designed with limited computational resources. This less complex models cannot fully exploit the available data. Traditional Hidden Markov Models (HMMs) and early machine learning models were not very effective for continuous speech, such as sentences and phrases, because they struggled to capture long-term dependencies and contextual nuances in the sequences of lip movements. Traditional methods often require a significant amount of pre-processing and feature engineering for each new dataset, making it challenging to scale and adapt to new data efficiently. Machine learning models may not perform well on large-scale datasets due to limitations in their ability to automatically learn from raw data. In short, the limitations of traditional and machine learning approaches in feature extraction, model complexity, temporal modelling, and scalability highlight the need for deep learning approaches. The advancements in deep learning, coupled with emerging technologies, offer promising solutions to overcome these challenges, leading to more accurate, robust, and scalable automatic lip-reading systems.

ALR methods based on deep neural networks use end-to-end approaches, automatically learning the characteristics of lip movement information from video to achieve classification. As mentioned above, in deep learning approach, ALR is divided into three steps (See figure 3):

1. **Lip detection and extraction**

2. **Front-end** (**Feature extraction and transformation**)

3. **Back-end** (**Classification**)

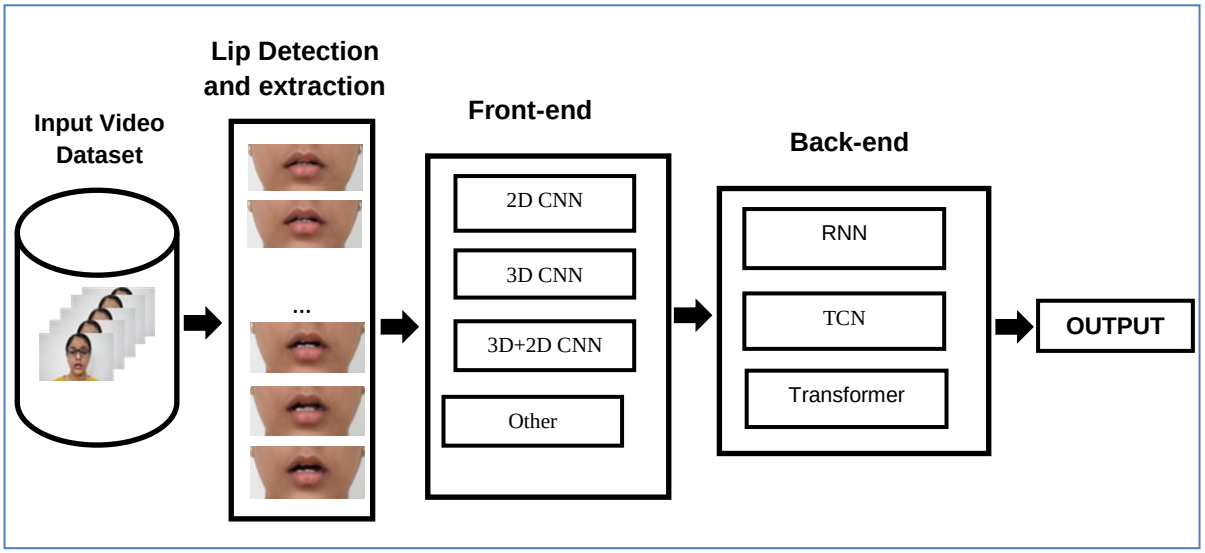


Figure 3 Lip reading process based on Deep Neural Network

2.3.1 Lip detection and extraction

Many pre-trained models are available for deep neural network based ALR. Haar cascades, Dlib frontal face detector, Multi-task cascaded Convolution Neural Network (MTCNN), Caffe Model etc. Haar cascades were introduced in 2001 by Paul viola and Michel Jones [56] which is simple and efficient classifier that can handle large set of features very effectively. MTCNN proposed by Kaipeng Zhang et al. in 2017, [67] is a cascaded structure with three levels of CNN. It detects the face from given image with reference to five facial landmarks including eyes, nose, and left and right lip corners. Caffe model is based on Single Shot-Multibox Detector (SSD) and uses ResNet-10 as backbone [68].

Dlib is a C++ toolkit which offers a wide range of functionality through various machine learning algorithms. Dlib is based on HOG and linear SVM and it gives outstanding results in frontal face detection. Dlib Library can be used to detect 68 landmarks of the face. The areas between 49–68 landmarks is a lip area, which can be cut down as the input for front-end network [27].

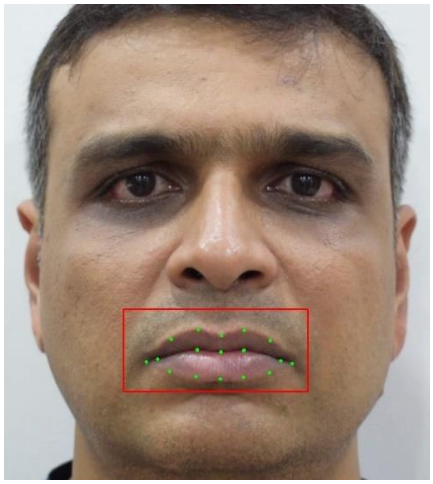


Figure 4 Face landmark points for lip detection

2.3.2 Front-End

Front-end uses different deep neural networks to extract lip movement features from lip images. The output of front-end network affects the classification. Different networks like Feedforward Neural Network (FNN) [69], [70], [71], Deep Belief Network (DBN) [72], Boltzmann Machines (BM) [73], Auto-encoder [74], and Convolution Neural Network (CNN) are used for feature extraction and transformation. CNN has been the most common and effective network architecture for feature extraction. The classic structures of CNNs are LeNet[75], AlexNet[76], VGGNet[77], GoogLeNet[78], ResNet[79], DenseNet[26], and MobileNet[80] etc. In CNN, neurons of higher layers perceives local areas, collect low-level features (like edges), synthesized at higher layers to extract high-level features from deeper layers. This stacked layer structure of CNN is effective for image recognition. To implement ALR with video data, more depth is added to CNNs, resulting in 3D CNNs [81].

2.3.3 Back-End

The main task of the back-end is to model the features extracted from the front-end network over time. The back-end learns long-term dependencies to perform recognition. This is crucial for video data where sequence of frames matters. Recurrent Neural Network (RNN) and Long-Short Term Memory (LSTM) networks [82] are designed to handle sequential data. By capturing dynamic nature of speech and lip movements, these networks can learn features changes over time.

An extension of RNNs called Bi-RNN processes input data in both forward and backward directions. Speech and lip movements inherently exhibit bidirectional temporal

dependencies. When lip movements depend on both prior and subsequent movements, Bi-RNNs can effectively recognize them. By understanding the context of the entire sequence, Bi-RNNs can more accurately recognize words and phrases, even with variations in speaking styles and speech [83], [84], [85].

2.4 LITERATURE STUDY OF ALR WITH DIFFERENT STRUCTURE OF CNN

Here, different implementations of ALR using 2D CNN, 3D CNN, combination of 2D and 3D, and other special deep learning models are discussed. After the development of Petajan's AV-ASR system and before the implementation of deep learning models for ALR, the process required extensive pre-processing of frames to extract image features, temporal pre-processing to extract video features, or the use of handcrafted vision pipelines. The ALR literature is extensive, a few key studies are [61], [86], [87], [88].

In 1994, Goldschen implemented an Automatic Lip Reading (ALR) system to extract visual features from lip movements to enhance Automatic Speech Recognition (ASR). Later in 1997, he further developed the ALR system with more sophisticated techniques, including improved algorithms for lip extraction (potentially involving image processing or facial feature tracking algorithms) and advanced statistical models (likely HMM) for feature extraction and recognition from video datasets.[55],[89]

After this, in 2000, Neti et al. used HMMs with hand-engineered features to implement the first sentence-level audio-visual speech recognition system using the IBM ViaVoice Dataset [90].

2.4.1 ALR with 2D-CNN

In the second decade of the 21st century, instead of hand-crafted algorithms and statistical models, researchers have turned their focus to CNN-based models for ALR. Initially, these approaches performed classification tasks for words, phonemes, or digits. In 2014, Noda et al. [91] proposed a work using 2D CNN for visual feature extraction and GMM + HMM for isolated Japanese word recognition. The experimental results showed that visual features obtained by 2D CNN were superior to traditional approaches [91]. In their subsequent work, Noda et al. incorporated a voice module to implement an Audio-Visual Speech Recognition (AVSR) system, achieving satisfactory results using 2D CNN [92].

Garg et al. [93] employed a 2D CNN in the front end and LSTM in the back end, using a pre-trained Very Deep CNN (VGGNet), and applied it to celebrity faces sourced from IMDB and Google Images [94]. Li et al. [95] utilized seven image frames (the current frame and ±3 adjacent frames) to create dynamic feature images for input to a 2D CNN, followed by classification using HMM.

In 2017, SyncNet was developed by Chung and Zisserman et al. with five convolutional layers, two fully connected layers, and an LSTM architecture [96]. ImageNet [97] developed a network with five convolutional layers, three fully connected layers, and one LSTM layer. ImageNet demonstrated a lower recognition rate compared to SyncNet, as training the convolutional network directly on the dataset yielded more accurate results than pre-training.

Saitoh et al. [98] introduced Concatenated Frame Image (CFI) as a novel approach for representing sequence images, demonstrating its ability to preserve spatiotemporal information across the entire image sequence. They used Network in Network (NIN) [99], AlexNet, and GoogLeNet to extract CFI features. NIN incorporates a micro-network named Mlpconv, where the layer structure resembles both convolutional and multi-layer perceptron layers. AlexNet consists of five convolutional layers and three fully connected layers, while GoogLeNet employs a 22-layer deep network with a sparse connection architecture, outperforming the others according to experimental results [59].

Zhang et al. [100] proposed LipCH-Net, an end-to-end visual speech recognition architecture designed for recognizing Chinese sentences. The architecture includes two deep neural network modules. It uses fixed-size grayscale images as input, extracts lip features using a CNN based on VGG-M, and processes them through ResNet with two layers of LSTM.

2.4.2 ALR with 3D-CNN

Lip movement is crucial in the ALR, making both timing and sequence important. A 2D CNN can only extract spatial information, not temporal information. A 3D CNN, however, uses a 3D kernel to convolve the image cube, which is an array of contiguous frames. This allows the feature map in the convolution layer to connect with multiple contiguous frames from the previous layer, thereby capturing motion information. The performance of 3D CNNs in ALR is demonstrated in [101], [102], [103], and [104].

In 2016, Assael et al. [105] proposed an end-to-end sentence-level model named LipNet. LipNet maps a variable-length sequence of video frames to text using three layers of 3D convolutional networks in the front-end and two layers of Bi-GRU [106] in the back-end. Classification and network training tasks were implemented using the Softmax and Connectionist Temporal Classification (CTC) loss functions, respectively. Fung and Mak [107] proposed a similar architecture with eight layers of 3D CNNs and max-out activation units.

Many researchers have utilized 3D CNNs for automated ALR. Torfi et al. [108] proposed an audio-visual speech recognition system coupled with 3D CNNs. Chung et al. [109] proposed a 3D CNN and VGG-based architecture with four modules: Watch, Listen, Attend, and Spell (WLAS). In this architecture, Watch and Listen are used in the front-end, and Attend and Spell are used in the back-end. Xu et al. [110] proposed another end-to-end ALR system named LCANet, which includes a 3D CNN, highway network, and Bi-GRU in the front-end, and a CTC network in the back-end. Highway networks, designed by Shrivastava et al. [111], overcome the training problem of deep networks using LSTM and adaptive gate units to regulate information flow. Deep highway networks, with hundreds of layers, can be trained using a simple gradient descent method [111]. Yang et al. [112] developed the largest Chinese dataset for lip reading, LRW, and proposed a D3D model based on DenseNet.

2.4.3 ALR with Hybrid 2D-3D Convolutional Neural Network

2D CNNs can extract spatial features, while 3D CNNs can capture temporal dynamics. Combining these characteristics can reduce the high computation and storage costs associated with 3D CNNs alone. With this in mind, Stafylakis and Tzmiropoulos [113] developed a word-level end-to-end visual speech system. The front-end network in their architecture consists of one layer of 3D CNN and 34 layers of 2D ResNet, while the back-end network comprises two layers of Bi-LSTM. Margam et al. [114] developed a 3D-2D CNN architecture for character-level recognition.

Several researchers, including Xiao et al. [115], Luo et al. [116], Zhang et al. [117], and Zhao et al. [118], have designed various architectures combining 3D and 2D CNNs for lip-reading tasks. A wide range of architectures is available in deep learning networks.

2.4.4 Other Deep Learning Network for ALR

There are some end-to-end architectures which are not based on CNN or RNN. Ngiam et al. [119], have utilized auto-encoders and RBMs in their multimodal deep learning experiments. They developed models that can process and integrate information from different data modalities, such as audio and video. Autoencoders were used to pre-train the network by learning to reconstruct input data, while RBMs were utilized to model the distribution of the features and capture the correlations between different modalities. Their work does not primarily focus on 2D CNNs but rather on non-CNN approaches to effectively handle multimodal data [119].

Following this, many researchers have proposed various end-to-end architectures for sentence recognition. Petridis et al. [74] proposed four end-to-end systems composed of autoencoders and LSTMs [74], [120], [121], and [122].

2.5 AN OVERVIEW OF DATASETS AND LANGUAGES USED IN ALR

Standard datasets are crucial for implementing machine learning models for ALR. The size and quality of a dataset are vital parameters that affect the training phase of a model. Additionally, the spoken language of a dataset is a key factor. In lip reading, the recognition rate depends on the oral motion of the language, regardless of scene length [123]. Various datasets, differing in size and spoken languages, have been used for lip reading. The complexity of these datasets is determined by the number of speakers, their gestures and postures, the background, and the content of the datasets, such as alphabets, numbers, words, phrases, and sentences. Lip reading datasets based on uttered text [123] are categorized in Figure.

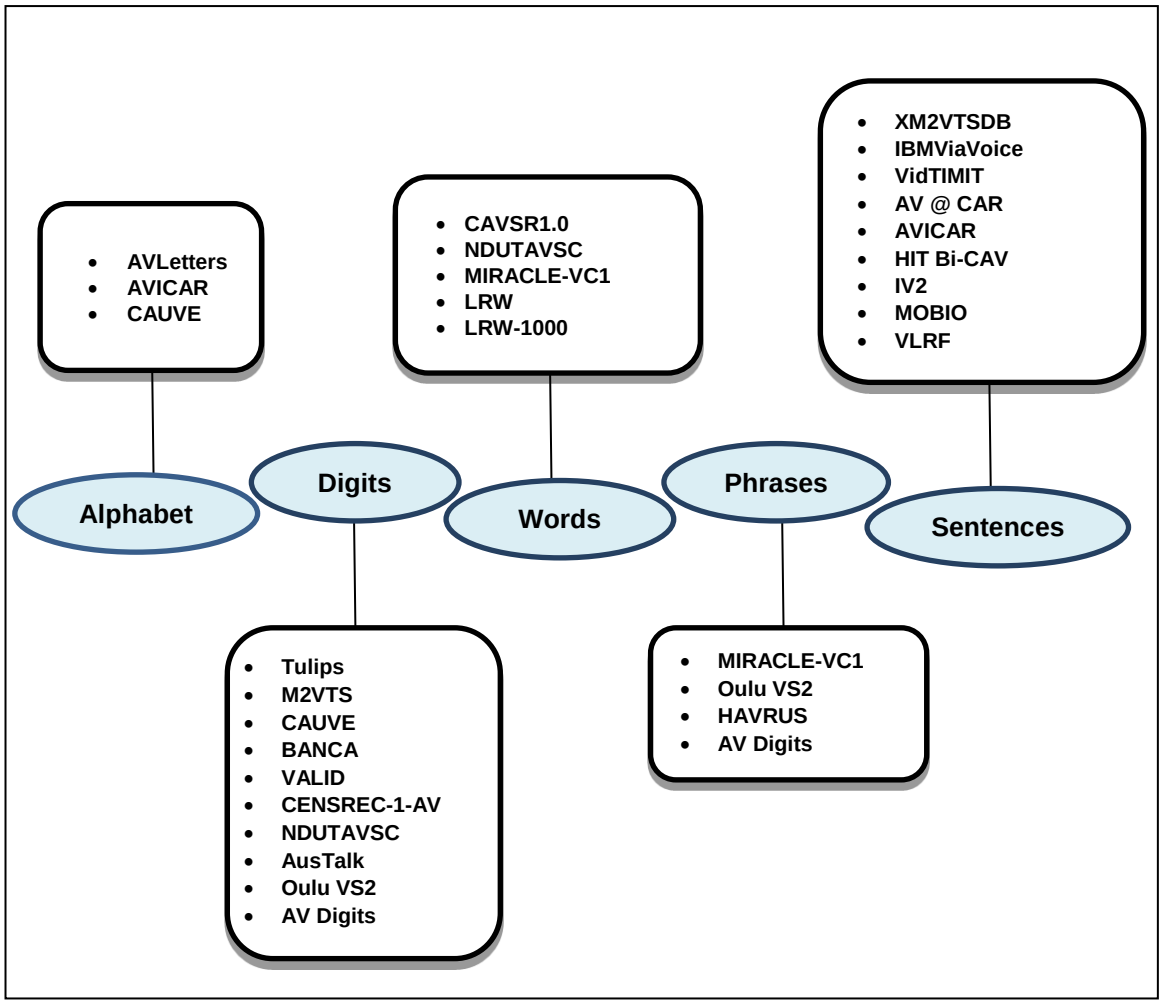


Figure 5 Lip Reading Datasets categories based on uttered text

ALR datasets can be categorized based on their recording settings. Some datasets are recorded in controlled environments, ensuring consistent lighting, background, and recording conditions. Others are captured from real-world sources such as news broadcasts, YouTube videos, and online lectures, presenting a variety of challenges such as varying lighting, diverse backgrounds, and different recording conditions. This type of data is referred as Lip Reading in the Wild (LRW) [5]. Various datasets such as Tulips, XM2VTSDB, AVLetters, CAUVE, AVICAR, OuluVS, OuluVS2, GRID, MIRACLE-VC, and VidTIMIT are created under controlled settings. Datasets like LRW, LRW-1000, LRS, LRS2-BBC, MV-LRS, and LSVSR are datasets collected in the wild.

2.5.1 ALR Datasets for Alphabets and Digits

As my work is focuses on the Gujarati Alphabet, I will discuss various lip reading datasets for alphabets and digits. Table 1 lists the ALR datasets created so far for alphabets and digits, and Table 2 is for deep learning models used in different datasets with accuracy.

Dataset Name	Year	Language	Task
Tulips [124]	1995	English	Digits
M2VTS [125]	1999	French	Digits
XM2VTSDB [126]	1999	English	Digits
AVLetters [127]	2002	English	Alphabet
CUAVE [128]	2002	English	Digits
BANCA [129]	2003	Multiple	Digits, Alphabet
AVI@CAR [130]	2004	Spanish	Digits, Alphabet
AVICAR [131]	2004	English	Digits, Alphabet
IBMIH [132]	2004	English	Digits
VALID [133]	2005	English	Digits
AVLetter2 [134]	2008	English	Alphabet
IBMSR [135]	2008	English	Digits
CENSREC-1-AV [136]	2010	Japanese	Digits
QuLips [137]	2010	English	Digits
NDUTAVSC [138]	2010	Dutch	Digits
LTS5 [139]	2011	French	Digits
AGH AV [140]	2012	Polish	Digits
AusTalk [141]	2014	English	Digits
OuluVS2 [142]	2015	English	Digits
AV Digits [143]	2018	English	Digits

Table 1 ALR Datasets for Alphabets and Digits

2.5.1.1 Datasets for Alphabets

AVLetters: AVLetters [127] is an audio-visual dataset published by Matthews et al. This popular dataset consists of 780 utterances of isolated English letters A-Z by 10 speakers (5 women and 5 men). In this small-scale dataset, each speaker utters each letter three times. Zhao et al. [142] achieved a 62.80% WRR using LAB-TOP for feature an SVM for classification. Pei et al. [144] used an RFMA-based system and achieved a WRR of 69.60%. Petridis and Pantic [145] combined DBNF and DCT features and applied an LSTM model, achieving a classification accuracy of up to 58.10% on the AVLetters dataset [145]. Hu and Li [146] proposed a system based on multimodal RBMs, named Recurrent Temporal Multimodal Restricted Boltzmann Machines (RTMRBMs). They applied this system to AVLetters, AVLetters2, and AVDigits using different modalities (audio, video, and both), achieving a mean accuracy of 64.63% for AVLetters [146].

AVICAR: It is an audio-visual speech recognition in a Car (AVICAR) dataset [131] published by Lee et al. This dataset is recorded inside a car, where four cameras are placed on the dashboard. There were 100 speakers (50 women and 50 men) who were reading 13 numbers, letters and 20 TIMIT sentences [147]. Dataset consists of text annotation of isolated digits, isolated letters, and ten-digit phone numbers along with TIMIT sentences.

In AVICAR, 60% of subjects are Native American and the others have Latin American, European, East Asian, and South Asian backgrounds. The speakers are divided in 10 groups of five males and five females. For each group, a different script set is prepared, where 118 utterance are recorded for each script set [5]. Fu et al. [148] have used a Locality Discriminant Graph (LDG) in a novel lip reading framework on the AVICAR dataset, achieving accuracy of up to 37.87%.

Name	Year	Model		Task	Accuracy
		Front End	Back End		
AVLetters	2002	LBP-TOP	SVM	Alphabets	62.80%[142]
	2013	RFMA		Alphabet	69.60%[144]
	2016	DBNFs +DCT	LSTM		58.10%[145]
	2016	RMRBM			64.63%[146]
CAUVE	2009	AAM	HMM	Digits	83%[149]
	2011	Auto-encoder + RBMs			68.70%[119]
	2017	DBNF	GMM-HMM		63.40%[150]
	2017	Auto-encoder+ LSTM	Bi-LSTM		78.60%[74]
LRW	2016	CNN +LSTM	LSTM + Attention	Words	76.20%[109]
	2017	3D CNN			98.50%[108]
	2017	3D CNN + ResNet	Bi-LSTM		83.00%[165]
	2017	3D CNN + ResNet + word boundaries	Bi-LSTM		88.08%[166]
OuluVS2	2016	DCT + PCA	HMM	Phrases	63.00%[168]
	2016	SDF + STLF	SVM		87.55%[167]
	2016	CNN	LSTM		83.80%[168]
	2017	Auto-encoder + Bi-LSTM	Bi-LSTM		96.90%[122]
GRID	2016	Eigenlips HOG Feed Forward	SVM SVM LSTM	Sentences	69.50%[69] 71.20%[69] 79.60%[69]
	2016	3D CNN	Bi-LSTM +CTC		95.20%[105]
	2018	3D CNN + HW	Bi-GRU + Attention-CTC		97.10%[110]
	2019	3D CNN + 2D CNN	Bi-LSTM-CTC		98.70%[114]

Table 2 : Models used for datasets of Alphabets and Digits, words, phrases, and sentences

2.5.1.2 Datasets for Digits

Tulips: The first ALR dataset, released in 1995 by Movellan et al. [124], consists of 12 speakers (3 women and 9 men) and 96 samples. Each speaker utters 1-4 digits in English twice. Using a simple Hidden Markov Model (HMM) on the Tulips dataset [124] achieves an accuracy of up to 89.93%.

M2VTS: M2VTS [125] is a French dataset released by Vanegas et al. in 1994. This dataset contains 37 speakers (12 women, 25 men) who utter the numbers 0-9 in French five times. Using HMM, the recognition rate with original lip movement is 74.5%, while with lip tracking, it is 76.6% [125].

CAUVE: For digit recognition, CAUVE [128] is the most widely used dataset. It is a speaker-independent corpus with 7000 utterances of both connected and isolated digits. Papandreou et al. [149] used an AAM-HMM-based model on the CAUVE dataset and achieved an accuracy of up to 83.00%. Ngiam et al. [119] used a structure based on an Autoencoder and Boltzmann machine, achieving a WRR of 68.70%. Rahmani and Almasgani [150] implemented lip-reading using a DNN-HMM hybrid system on the CAUVE dataset and obtained a WRR of 63.40%. Petridis et al. [74] achieved a WRR of 78.60% for the CAUVE dataset using LSTM and BLSTM.

CENSREC-1-AV: To design this dataset, more than 5000 Japanese connected digit utterances were collected, with each utterance consisting of 1-7 digits. Speech signals were recorded in an office environment using a lapel microphone connected to two cameras. One camera captured color videos, while the other simultaneously captured infrared images using a lens filter. Infrared images may be useful when the lighting conditions change drastically [136]. Ninomiya et al. used a deep learning model with Deep Bottleneck Features (DBFs) as the front end, and GMM and HMM as the back end. This proposed "multi-stream DBNF-GMM-HMM" model achieved a WER of 18.9% [151].

NDUTAVSC: It is the largest Dutch dataset designed for Automatic Lip Reading (ALR). The corpus consists of 10 hours and 38 minutes of continuous recordings of 66 speakers, including 20 females and 46 males. Speakers utter random sentences, numbers, letters, spellings, and open questions [138]. The recognition rate for the NDUTAVSC dataset, with AAM as the front-end and HMM as the back-end, is 84.27% [152].

AusTalk: The AusTalk corpus was designed and created through the 'Big Australian Speech Corpus (Big ASC)', a collaborative project with the goal of creating a large audio-

visual dataset for Australian English. Up to 1000 geographically and socially diverse speakers were recorded in locations across Australia. This corpus includes three one-hour recordings containing words, digits, and sentences [141]. Sui et al. [153] have used Autoencoder, DCT, and LDA in front-end, HMM in back-end, and achieved 67.8% accuracy, while Sui et al. [154] have used DBM, DCT, and LDA in front-end, HMM in back end and achieved 69.10% accuracy for AusTalk digits.

AVDigits: This audio-visual dataset contains normal, whispered, and silent speech from 53 speakers. Frontal, 45°, and profile views were used to record digits and phrases. Petridis et al. achieved 68.00% accuracy using an auto encoder and BLSTM [143].

2.5.2 Other Datasets

Following text briefly review other well-known and most widely used datasets.

The TIMIT corpus was created in 1990 to provide a standard dataset for the evaluation of Automatic Lip Reading (ALR) by the Massachusetts Institute of Technology (MIT), Texas Instruments (TI), and the Stanford Research Institute (SRI). It consists of recordings of 630 speakers from eight major dialects of American English, each reading ten phonetically rich sentences. TIMIT is used for training and testing in Automatic Speech Recognition (ASR), speaker recognition, phonetic analysis, and various other speech processing tasks. The TIMIT corpus is renowned for its high-quality recordings and detailed annotations, making it a benchmark dataset for evaluating the performance of speech recognition systems [147].

VidTIMIT [155] is based on TIMIT sentences. This audio-visual dataset includes recordings of 19 females and 24 males, each uttering 10 TIMIT sentences. VidTIMIT can be useful for research on topics such as multi-view face recognition, multi-modal speech recognition, and person identification.

GRID [156] is another well-known dataset, designed by Cooke et al. in 2006. It includes a total of 34 speakers, with each speaker uttering 1000 sentences, each sentence lasting approximately 3 seconds. The sentences in this dataset have a fixed form: <verb> + <color> + <position> + <digit> + <letter> + <adverb>. In this form, verbs, colors, positions, and adverbs each consist of four words, digits range from 0-9, and there are 25 letters in total, making the dataset size 51. The letter 'w' is excluded because its utterance is longer than other letters.

OuluVS [142] is an audio-visual phrase-based dataset containing 1000 sentences. It includes 20 speakers from different countries, each uttering 10 greeting phrases in English one to five times. OuluVS2 [157] dataset contains phrases from OuluVS, along with 10 randomly generated sequences of 10 digits, and 10 randomly chosen TIMIT [147] sentences.

MIRACLE-VC [158] is a visual dataset in which 15 speakers (10 women, 5 men) each speak 10 words and 10 phrases, 10 times. As an RGB-D camera was used for recording, the images contain more depth information, making this dataset useful for a variety of research fields such as face detection and biometrics.

The LRW [159] dataset is one of the established English word-based audio-visual lip-reading datasets collected by the Visual Geometry Group (VGG) at Oxford University. Videos in the dataset are sourced from BBC TV news programs in the UK. There are more than 1000 speakers, each with different postures and lighting conditions. Each speaker utters 500 words, with each word containing between 5 to 10 letters. All videos are 29 frames in length. Short words are excluded due to homophones, which can make the lip-reading process difficult. Additionally, 23 pairs of singular and plural forms of the same words and 4 pairs of present and past forms are avoided.

The LRW-1000 [160] dataset was collected from 51 TV programs across 26 TV stations in China. Unlike LRW, the sample resolution in LRW-1000 is not fixed. This is the largest Chinese dataset, containing more than 1000 classes and a total of 718,018 samples

LRS [161] and LRS2-BBC [162] are audio-visual datasets prepared by the VGG group. In LRS, videos collected from BBC programs recorded between 2010 and 2016 are divided into sentences and phrases based on the punctuation marks in the transcripts. LRW and LRS focus solely on news and debate programs, while LRS2-BBC includes a wider range of programs.

The LRS3-TED [163] dataset, also from the VGG group, is based on TED and TEDx talks in English. In this dataset, videos are clipped to 100 characters or six seconds. The subjects in the training, test, and validation sets are not identical in LRS3-TED. In LRW, LRS, and LRS3-TED, the train, test, and validation sets have overlapping data because many speakers appear in various programs.

MV-LRS (Multi-View LRS) [164] is based on LRS with two differences. There is a small face angle deviation in LRW and LRS2-BBC, whereas in MV-LRS, the angle ranges from

0° to 90°. MV-LRS includes dramas and factual programs where speakers are engaged in conversations, creating multi-view data.

2.6 WORK CARRIED OUT FOR INDIAN LANGUAGES

Using ALR technology, deaf children can be taught their mother tongue easily during childhood and later learn other languages through lip reading. As observed in the previous section, primary research on ALR has predominantly utilized datasets from the English language. However, in the last two decades, ALR has also been implemented for other languages such as Chinese, Japanese, Spanish, and French, due to technological advancements. In the last decade, there have been a few attempts to develop lip-reading systems for Indian languages as well [169]. Initially, these efforts focused solely on acoustic signals.

2.6.1 Literature Review for Indian Languages

In 1976, Ali et al. applied adaptive pattern recognition theory for the recognition of speech signals of 36 Hindi syllables [170]. In 1983, Paliwal et al. performed speech recognition for speech signals of Hindi digits [171]. Patil et al. [172], showed that ILSL12 phone-set, which is widely used for Indian Language speech recognition, has limitations to represent features of the speech like, voicing, fricatives, etc. for Indian Languages. They addressed the issue by considering the voiced and unvoiced features for Hindi speech recognition. They incorporated finer representations at the time of lexicon expansion and successfully tested for Hindi word recognition showed that it has significant improvement in WER [172].

In the development of ALR system for Indian languages, Faruquie et al. [173] presented a novel method of for Translingual visual speech synthesis. Their approach to adapting a speech recognition system from a base language (e.g., English) to a novel language (e.g., Hindi) through phonetic and viseme vocabulary adaptation layers offers a practical solution for handling the diverse phonetic landscapes of Indian languages[173].

In 2017, Kandagal et al. implemented Visual Speech Recognition for the digits 0-9 in English, Kannada, and Telugu languages. They used the canny edge detection algorithm for ROI extraction, the Gray Level Co-occurrence Matrix (GLCM) and Gabor Convolve algorithm for feature extraction, and an Artificial Neural Network (ANN) for classification. With 120 samples, they achieved an accuracy of 90% [174].

Brahme et al. [62] have implemented VSR for identifying first three digits of Marathi language. For this, Lip extraction is performed, followed by feature extraction using landmark points and classification using dynamic time warping is applied. The author has achieved 63% recognition rate for this experiment [62].

Nandini et al. have proposed deep weighted feature algorithm for lip-reading of Kannada language. They achieved 84.82% of accuracy with a dataset was collected from a deaf and dumb organization [175].

Patil et al. have given LSTM based lip-reading approach for Devanagari script. They have created a dataset by recording videos of 50 people. A content of a dataset is a paragraph with 58 unique words of Devanagari script. After applying Face landmark algorithm of Dlib, they have applied Lip height calculator which calculates gap between lips for each frame. In lip movement, series of lip heights have specific pattern. To capture this pattern, they have applied two methods: one height feature extraction and two height extraction method. Output of this method is a scaled height of spoken words, which is used for training the model with LSTM. In training with one height feature extraction, above model gives accuracy of 77% for 3 words spoken by 9 speakers. When number of words and speakers increased, accuracy decreased to 8%. In two height feature extraction method, accuracy achieved is 35.60% for 10 words spoken by 30 speakers. Again if words are increased up to 58, accuracy decreased to 12.32%. The author justify this accuracy fall with small dataset size and low frame rate. They also justify that LSTM model used gives higher accuracy for large dataset. Dlib library algorithm face landmark provide only 20 coordinates for lip detection, which are not sufficient, specifically when speakers have moustaches and beards. To improve results they have divided the paragraph in sentences and achieved accuracy up to 20% [176].

Rudregowda et al. [177] worked on a deep learning approach for Automatic Lip Reading (ALR). They developed their own dataset consisting of 5 words in the Kannada language. Using the VGG16 convolutional neural network with the Rectified Linear Unit (ReLU) as the activation function, their model achieved an accuracy of 91.9% [177]. Table 3 summarizes the ALR work done for Indian languages.

Sr. No.	**Title**	**Dataset & Language**	**Approach**	**Accuracy**
1	Visual Speech Recognition Based on Lip Movement	0-9 digits in English,	Traditional (Image processing	90% [174]

	for Indian Languages (2017)	Hindi, and Telugu	Techniques) + Machine Learning	
2	Deep Weighted Feature Descriptors for Lip Reading of Kannada Language (2019)	Kannada words	Traditional (Image processing Techniques) + Deep Learning	84.82% [175]
3	Marathi digit recognition using lip geometric shape features and dynamic time warping (2017)	First three Marathi digits	Traditional (Image processing Techniques) + Machine Learning	63% [62]
4	LSTM Based Lip Reading Approach for Devanagari Script. (2019)	Paragraph, Sentences in Hindi	Traditional (Image processing Techniques) + Deep Learning	77.02% (9 SP+3 W) 8.10% (30 SP +58 W) 20% (18 SP + 10 S) [176]
5	Visual Speech Recognition for Kannada Language Using VGG16 Convolutional Neural Network (2023)	5 Words in Kannada	Deep Learning	91.9% [177]
6	A viseme recognition system using lip curvature and neural networks to detect Bangla vowels	-	-	[178]

Table 3 ALR work for Indian languages

2.6.2 About Indian Languages

India is one of the most linguistically diverse nations in the world. According to most recent census of India, 2011 total 1369 are rationalized mother tongues, which are grouped in 121 languages. In these 121 languages 22 are major languages. Among these languages, majority of them are Indo-Aryan languages spoken by 78.05% of Indians. Gujarati is one of the most widely spoken Indo-Aryan languages [179], [180].

Majority of the Indian languages are derived from Devanagari scripts. Languages derived from Devanagari scripts have some special characteristics as compared to the English language and other non-Indian languages. They have a scientific way of speaking wherein the alphabets are categorized based on how they are spoken. The arrangements of letters in the Gujarati language are called "***Mulakshar***" which means basic letters. In English alphabets, the arrangement of letters is not logical. There is no reason why vowels are scattered around in the alphabets' set or why the letter G comes before the letter H. In Devanagari script and hence all languages derived from it, the consonants, and the vowels

are categorized separately. The alphabets (vowels and consonants) are arranged based on where and how the sound of that letter is produced inside the mouth. As the work which is being carried out is on Gujarati language, the next section discusses for Gujarati language only.

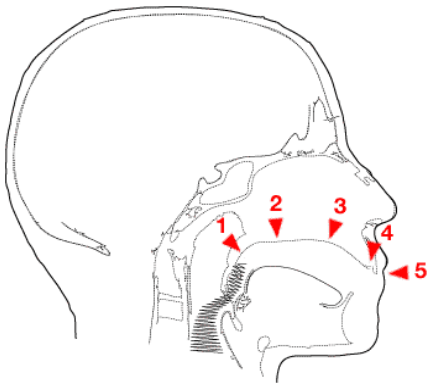
2.6.3 The Gujarati Language and Lip-Reading Work

In Gujarati language, there are 36 consonants and 12 vowels. Unlike the English language, vowels are not scattered in between consonants. They are separated and arranged based on where and how the sound is produced while speaking.

The classification of consonants based on spoken style for the Gujarati language is shown in table 6 and with figure 5. The first five consonants as shown in the table are called guttural as the sound of these consonants comes from the throat. Similarly, palatal group consonants are articulated when the tongue touches the hard palate. Retroflex group consonants are articulated when the tongue curls back a bit and touches the roof of the palate. A dental group of consonants is produced when the tongue touches the upper teeth and labial is produced using lips.

Name of the class	**Alphabets of the class**	**Spoken by**
Guttural	‘ક’, ‘ખ’, ‘ગ’, ‘ઘ’, ‘ઙ’, ‘હ’	back of the tongue touches the velum
Palatal	‘ચ’, ‘છ’, ‘જ’, ‘ઝ’, ‘ઞ’, ‘ય’, ‘શ’	the tongue touches the hard palate
Retroflex	‘ટ’, ‘ઠ’, ‘ડ’, ‘ઢ’, ‘ણ’, ‘ર’, ‘ષ’	the tongue curls back a bit and touches the alveolar ridge
Dental	‘ત’, ‘થ’, ‘દ’, ‘ધ’, ‘ન’, ‘લ’, ‘સ’	the tongue touches the back of the teeth
Labial	‘પ’, ‘ફ’, ‘બ’, ‘ભ’, ‘મ’, ‘વ’	rounded lips

Table 4 Classification of Devanagari Alphabets



1. **Guttural**
2. **Palatal**
3. **Retroflex**
4. **Dental**
5. **Labial**

Figure 6 Spoken style of alphabets

Gujarati language has special alphabets like 'ણ', 'ળ', 'છ', 'ક્ષ', 'જ્ઞ', 'ત્ર', 'શ્ર' and these alphabets have no specific equivalent unique alphabet in The English language. People, who have a native language as English, cannot pronounce these alphabets easily. Combinations of alphabets in these languages have the same spellings in English. If we want to differentiate them, we need to check their phonics.

Gujarati Alphabet	English spelling	Phonics	Gujarati Alphabet	English spelling	Phonics
ટ, ત	Ta	ṭa, ta	ષ, સ	Sa	ṣe, se
થ, ઠ	Tha	tha, ṭha	ળ, લ	La	ḷe, la
ડ, દ	Da	ḍa , da	ન, ણ	Na	ne, ṇe
ઢ, ધ	Dha	ḍhe, dhe			

Table 5 Alphabets and their corresponding phonics

With the technological developments, researchers would like to explore challenging areas to work. Many researchers have worked with ASR in Gujarati language in last decade [181], but no significant work is done for ALR in Gujarati language.

In 2015, Patel and Nandurbarkar [182] implemented speaker recognition system based on MFCC (Mel Frequency Cepstral Coefficients) and GMM (Gaussian Mixture Model) with 30 speakers. In 2016, Vijayendra and Thakar [183] implemented similar work with MFCC-RC (Real Cepstral Coefficients) for feature extraction and neural network for classification. Valaki and Jethva [184] gave a hybrid approach based on HMM and ANN for Gujarati speech recognition. Tailor and Shah [185] developed an HMM based lightweight speech recognition system. Raval et al. in [186], Pandit et al. in [187], and Tailor et al. [188] also worked with speech recognition system in Gujarati.

All above work is implemented for acoustic signals only. To the best of my knowledge, this is the first comprehensive research work on ALR specifically focused on Gujarati language. This research work addresses the critical need for developing ALR systems tailored to the linguistic diversity of India, covering languages like Gujarati and all languages which are derived from Devanagari script.

2.7 CHALLENGES IN AUTOMATIC LIP READING

Automatic Lip Reading (ALR) presents significant challenges due to the nature of its inputs, which primarily consist of video or image sequences. These sequences often feature image contents that may appear similar or unchanged across frames. The primary reason for this distinction lies in the subtle yet crucial changes in lip movements that occur during speech articulation.

When individuals articulate different phonemes, such as the alphabets, the changes in lip movements are exceedingly minute. These subtle shifts in lip configuration play a pivotal role in distinguishing between various phonemes. However, capturing and interpreting these nuanced variations pose formidable challenges for ALR systems. These challenges are described below:

1. **External Factors:**
    a. **Environmental Factors:** Different Environmental factors affects ALR accuracy. Like poor or changing lighting can affect the visibility of lip movements, variation in camera angles and resolutions can lead to inconsistent lip movement captures.
    b. **Physiological Factors:** Different factors like illumination, skin colour, beards, wrinkles on the skin etc. affect the process of feature extraction. To overcome this problem, traditional lip-reading methods use shape-based methods in which shape-based methods extracted features include the shape of the lips only [38] , [58] and other external factors like illumination, skin colour, and beard are discarded. In deep learning-based methods, various methods are used to extract spatial and temporal features of lip movement.
2. **Visual Factor:**
    a. **Co-articulation Effects**: Lip movements for a specific phoneme can vary depending on the preceding and following phonemes, making it difficult to isolate individual sounds accurately.

b. **Ambiguities in speech:** In alphabet pronunciations, different phonemes have same mouth shape. Such visemes are difficult to distinguish without context. Speakers' accent will add more complexity to the feature extraction task.

   Phoneme-to-viseme mappings given in [51], [189], [190] and adjacent character/words phenomena in [109], [161], [192] solve the problem of visible ambiguity up to some extent.

c. **Facial expressions and emotions:** Different speakers have unique ways of moving their lips, even speaking the same words. In some cases same speaker might produce variations in lip movements due to mood, health, fatigue, or speaking context. Multiple shots of same speaker may resolve the problem of variation in lip movements [109].

d. **Occlusions and obstructions:** This refers anything that blocks the view of speakers' lips and mouth and making it difficult to read lip movement accurately. That may include hand movements, facial hair (beards), wearing mask or other facial coverings.

e. **Speakers' pose:** It refers to the orientation and positioning of speakers' head and face relative to the camera. In ALR, variations in speakers' poses present a significant challenge, particularly when data are collected from sources such as TV shows or online sites. In such sources, speakers may exhibit a wide range of poses, leading to changes in the angle of the speaker in every frame of the video sequence. These variations make tasks such as Region of Interest (ROI) detection and feature extraction difficult, as the lip region may not be consistently positioned or aligned across frames. For ALR systems, accurate ROI detection and feature extraction are crucial for effectively capturing and interpreting lip movements. However, the dynamic nature of speakers' poses complicates these tasks, as the position and orientation of the lips may vary significantly from frame to frame.

   By leveraging multi-view datasets, ALR systems can improve their robustness to changes in speakers' poses and enhance their performance in accurately detecting and interpreting lip movements across diverse contexts and viewing angles. To address this challenge, researchers have developed multi-view datasets such as LRW[159], LRW-1000[193], LRS2-BBC[162], and OuluVS2[157]. These datasets provide video sequences captured from multiple viewpoints, allowing ALR systems to learn and adapt to variations in speakers' poses more effectively.

3. **Dataset Factors:** Datasets with limited number of speakers, corpus and samples also affect the performance of lip-reading task. For the datasets collected from TV shows, the background, illumination, environment and other parameters would be almost similar and the language contents are also very limited. A large-scale of datasets with a more number of speakers from different regions and different postural background can give more fruitful results for lip-reading task, and influence of speakers' dependency can be reduced.

SUMMARY

This chapter goes through a detailed study of the work that has been carried out for Automatic Lip Reading across the world. For Gujarati is our mother tongue and no ALR work is done yet, this research work is a small footstep in lip reading for Gujarati language.