IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 26, Issue 4, Ser. 3 (July. – August. 2024), PP 10-18
www.iosrjournals.org

Speech Recognition Based On Lip Movement Using Deep Learning Models - A Review

Meena Ugale, Aditya Pole, Sumit Desai, Harshita Gupta, Saquib Khan
[1,2,3,4,5](Department Of Information Technology, Xavier Institute Of Engineering/ University Of Mumbai, India)

Abstract:
Beyond spy movies, lip-reading technology whispers the promise of a quieter world. From assisting the hearing-impaired to navigating noisy environments and silent settings, it unlocks communication avenues like lip into text translation and biometrics. But challenges murmur concerns: limited data, visual ambiguity, and complex models. Researchers, however, shout back with diverse approaches, from feature-based analysis to powerful deep learning. Datasets like MIRACL-VC1 fuel the voice, while custom collections refine and personalize models. Performance speaks volumes, reaching a word recognition accuracy of 91.9%. The future whispers even more: robust, adaptable models, integrated with other modalities like audio and facial expressions, pave the way for a responsible and inclusive world where silent words find their voice.

Key Word*: Speech recognition, Lip movement, CNN, LSTM, Feature Extraction.*

Date Of Submission: 11-07-2024 Date Of Acceptance: 21-07-2024

I. Introduction

Lip reading recognition is a method of understanding spoken language by watching the movements of the speaker's lips. It has many uses, such as helping people with hearing problems, improving human-computer interaction, giving forensic proof, confirming biometric identity, and allowing silent dictation. Lip reading is a hard skill that needs knowledge and context, as many lip movements are invisible or unclear, involving the tongue and teeth. Lip reading is also influenced by many factors, such as the size and color of the lips, the position and expression of the face, the existence of facial hair or hand movements, the speed and quality of speech, the light and background conditions, and the accent and pronunciation of the speaker. Therefore, lip reading recognition is not an easy task, and it needs advanced methods and models to achieve high accuracy. Future research in lip reading recognition can explore the use of deep learning, multimodal fusion, and real-time systems.

One of the main challenges of lip-reading recognition is to get the important features from the lip region that can show speech information. Traditional methods depend on hand-made features, such as geometric shapes, contours, histograms, and optical flow, that capture the space and time changes of the lip movements. However, these features may not be strong to noise, occlusion, and illumination changes, and they may not capture the fine and complex patterns of the lip movements. Therefore, recent methods use deep learning methods, such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), and attention mechanisms, that can automatically learn the features from the raw pixel values of the lip images or videos. These methods can also deal with large-scale and high-dimensional data, and they can achieve state-of-the-art performance on lip reading recognition tasks.

An additional obstacle in lip-reading recognition involves integrating data from various sources like audio, video, and text to enhance accuracy. This integration, termed multimodal fusion, demands addressing synchronization, alignment, and fusion challenges. Contemporary approaches employ techniques such as encoder-decoder structures, cross-modal attention mechanisms, and multimodal fusion methods to effectively amalgamate information and enhance recognition systems' precision and resilience. Moreover, creating real-time applications for diverse scenarios poses another challenge. These applications aid individuals with hearing impairments in understanding speech in noisy environments or when the speaker's face is obscured. They also enhance human-computer interaction through silent dictation, voice control, and speech recognition. Additionally, they contribute to forensic analysis by scrutinizing suspects' or witnesses' speech in videos and verifying biometric identity through lip movement matching with voice or facial features. To address this challenge, recent methods employ lightweight models, edge computing, and online learning to reduce computational costs and latency, thus rendering lip-reading recognition systems more practical and adaptable across various situations and languages.

CNN, or convolutional neural network, is a type of deep learning technique that can extract features from images or videos by applying multiple filters or kernels. CNN can learn the features automatically from the raw pixel values, without relying on hand-crafted features, such as geometric shapes, contours, histograms, and optical flow. CNN can also handle large-scale and high-dimensional data, and achieve state- of-the-art performance on

lip reading tasks.CNN can be used for lip reading in different ways. One way is to use CNN to extract the features from the lip region of each frame in a video sequence, and then feed the features to another network, such as RNN, LSTM, or GRU, to process the temporal information and generate the output words or sentences. Another way is to use CNN to concatenate the lip images from a video sequence into a single image, and then use CNN to classify the image into a word or phrase. A third way is to use CNN to encode the lip images into a latent vector, and then use a decoder, such as attention mechanism, to decode the vector into a word or sentence. CNN has shown promising results for lip reading in various datasets, such as LRW, LRS2, LRS3-TED, CAS-VSR-W1k, GRID, and LRW-1000. However, there are still many challenges and opportunities for future research, such as improving the robustness and generalization of the models, integrating the information from other modalities, such as audio and text, and developing real-time and practical applications for lip reading.

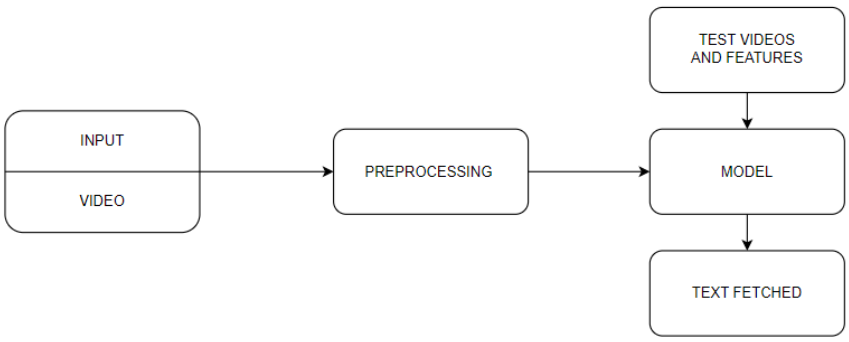

Figure 2. Generic Block Diagram

Figure 1 outlines the process of extracting text from a video. It begins with the input of a video, followed by preprocessing, then it goes through a model that tests videos and features, and finally, the text is fetched.

II. Literature Review

Hendrik Laux et al. in [1] presented a method to improve the communication skills of patients who have difficulty speaking in a critical care setting by using lip-reading. They developed a method to estimate the silent speech by using visual speech recognition, i.e., lip-reading. In a two-stage architecture, they used the patient's face images to infer audio features as an intermediate prediction target, which were then used to estimate the spoken text. The dataset they used was an audio-visual dataset recorded in the University Hospital of Aachen's ICU with a language corpus chosen by experienced clinicians.

Amit Garg et al. in [2] is about some new methods for visual speech recognition, which is the task of recognizing what someone is saying by looking at their lips. It also explains how the methods work and how they compare with each other. First method used was a pre-trained VGGNet model to classify a grid of images that represent the first k frames of a lip sequence. One method uses interpolation to create a grid that is invariant to speaking speed and uses more input data. Other method tries to train a smaller model from scratch on the grid images but fails due to the small dataset size. Final method used was LSTM layers to handle variable-length sequences, but does not perform well because it does not use temporal information during feature extraction. The best-performing method was the interpolated grid method, which is followed by the original grid method. The LSTM and the scratch methods perform poorly.

Jyotsna Uday Swami et al. in [3] review different methods for lip reading recognition, which is the task of turning speech (lip movements) into text without sound. The paper evaluates five methods: Dynamic Time Warping, Shape Template, CNN, Snake's approach and Hidden Markov model. The paper uses a part of the GRID dataset to measure the performance of each method in terms of word accuracy rate. The paper also uses a new LSTM to get the CNN features and track the lip shape points. It also discovers that the CNN method with different appearance parameters works better than other methods, and that the appearance information is more useful than the shape information for lip reading. The paper also notes that pixel-based methods and LSTM can be used to normalize the images and remove the shape variations. It proposes some possible enhancements for future work, such as using visemes as classifiers, and using the inner appearance of the mouth area as features.

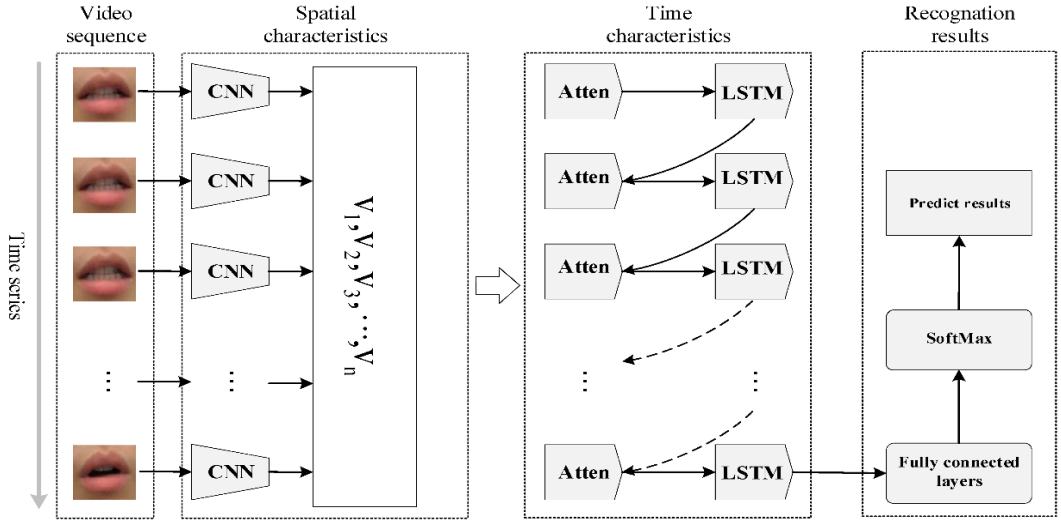

Figure 2. Architecture Diagram for Lip Reading Recognition [3].

Figure 2 outlines how a machine learning model can recognize and interpret video sequences of lips. It uses two types of neural networks: one to analyse the shape of the lips in each frame, and another to analyse the changes over time. The model then predicts the results, such as what words are being spoken or what emotions are being expressed.

Akshay S. Nambeesan et al. in [4] introduced a system that could analyze the shape and movement of lips from video and depth images, and convert them to text using deep learning. The system aimed to help hearing impaired people communicate with others without learning sign language or depending on human assistance. They applied OpenCV to identify and separate the face and the lip region from the input images. They employed faculties to obtain the coordinates of the mouth and other facial features. They resized and merged the lip frames into a mega-image for each word or phrase spoken by the target speaker. They built a convolutional neural network (CNN) and a long short-term memory (LSTM) network for classification and prediction using Keras and TensorFlow. The CNN extracted features from the mega- image, and the LSTM captured the temporal dependencies of the lip frames. The output layer of the network estimated one of the 20 possible words or phrases from the MIRACL-VC1 dataset. They achieved an accuracy rate of 85% for their system, and contrasted it with a CNN-only model. They demonstrated that the LSTM enhanced the performance of the system by handling the coarticulation and homophones problems. They also examined the advantages and limitations of their system, and proposed future improvements.

Gerald Schwiebert et al. in [5] presents GLips, a collection of 250,000 videos of German speakers from the Hessian Parliament, who say 500 different words. The dataset is compatible with the English LRW dataset, which is a common standard for lip reading. The authors used an automated process to extract the videos, align the subtitles, and detect the faces of the speakers. The dataset also has metadata and phonetic information for each word. The authors trained a deep neural network (X3D) on GLips and LRW, and examined whether transfer learning between the two languages could improve the word recognition rate. They showed that transferring knowledge from one language to another could speed up the learning process and achieve better performance, especially when the target dataset was small or noisy. They also discussed the factors that affected the quality and compatibility of the datasets.

Srikanth, G. N. et al. in [6] presented a technique to identify words using both the sound and the lip movements of the speaker. They argued that this technique could increase the accuracy of word recognition, especially when the sound was noisy or the speaker had an accent. They used Mel Frequency Cepstral Coefficients (MFCC) to capture the sound features, and geometrical parameters of the lip region to capture the visual features. They merged these features into a single vector, and used the variance of each feature as a measure of its importance. They selected the most relevant features to reduce the dimensionality of the feature vector. They used Multilayer Perceptron (MLP) and K- Nearest Neighbor (KNN) to classify the words based on the fused features. They compared the performance of these algorithms on a dataset of 460-word utterances from the VidTIMIT database. They reported that MLP achieved a higher accuracy than KNN, and that using both sound and visual features improved the accuracy over using only one modality.

Hameed Hira et al. in [7] introduced a technique to identify words from the lip shape and movement using radio frequency (RF) signals, even with a face mask on the speaker. The technique used Wi-Fi and radar technologies to detect the lip movements and translate them to text using deep learning. The paper aimed to overcome the challenges of camera-based lip-reading systems, such as occlusion, lighting, and privacy issues. The

paper gathered a dataset of vowels and empty (closed lips) utterances using both Wi-Fi and radar, with a face mask. The paper-trained machine learning and deep learning models on the dataset, and achieved a high accuracy of 95% using neural networks. The paper also examined the difficulties and future directions of RF-based lip reading.

Ziad Thabet et al. in [8] presented LipDrive, a system that uses machine learning to read lips and enable communication between humans and autonomous vehicles in noisy environments. They described how they used DLib, a library, and Z-order, a technique, to extract features from the speakers' lip movements. They also explained how they processed the images and created feature sequences for each word. They then compared the performance of nine different linear classifiers on a large dataset of 500 words and reported the accuracy and the confusion matrix of each one. They discussed the challenges and opportunities of lipreading in different scenarios and concluded that Gradient Boosting, Support Vector Machine, and Logistic Regression were the best classifiers for this task, with accuracy rates of 64.7%, 63.5%, and 59.4%, respectively.

Lap Poomhiran et al. in [9] presented a new technique for lip reading that enhanced the recognition performance by using the concatenated three-sequence keyframe image technique. This technique chose three keyframes from each word utterance and merged them into a single image. The image was then assigned a label by a convolutional neural network (CNN). The authors employed the MIRACL-VC1 dataset, which consisted of 10 words and 10 phrases uttered by 15 speakers. They contrasted their technique with two other techniques: one that used only the first keyframe of each word and one that used all the frames of each word. They also examined the effect of different CNN architectures and input resolutions on the recognition performance. The authors revealed that their technique attained an accuracy of 86.67% for words and 83.33% for phrases, which was significantly higher than the other techniques. They also demonstrated that using a deeper CNN architecture and a higher input resolution could boost the accuracy. The authors claimed that their technique was simple, robust, and efficient for lip reading.

Shashidhar Rudregowda et al. in [10] introduced an enhanced lip-reading method employing concatenated three sequence keyframe images. This approach selected three keyframes per word utterance, merging them into a single image for labeling by a convolutional neural network (CNN). Utilizing the MIRACL-VC1 dataset, containing 10 words and 10 phrases spoken by 15 speakers, they compared their method with alternatives using either the first keyframe or all frames per word. Experimenting with different CNN architectures and input resolutions, they achieved significantly higher accuracies of 86.67% for words and 83.33% for phrases. Their method's simplicity, robustness, and efficiency were underscored, further validated by applying it to a newly created Kannada language dataset, yielding an accuracy of 91.9%.

Dmitry Ryumin et al. in [11] introduced a method to detect words and gestures using signals captured by mobile device sensors like phones and tablets. They employed deep neural networks for real-time and offline analysis, gathering data from 20 volunteers through mobile devices and extracting features like MFCC and optical flow. Their approach comprised two models: one for word recognition utilizing RNN with LSTM and CTC, and another for gesture recognition using CNN with SoftMax. Testing on their dataset and comparing with existing methods, they demonstrated high accuracy and low latency. Potential applications included human-computer interaction for smart home control, gaming, and language learning. They acknowledged challenges such as noise, occlusion, and device orientation, asserting their method as innovative, effective, and user-friendly for mobile device sound-image speech and gesture recognition.

L. Ashok Kumar et al. in [12] presented an AVSR system that could help hearing impaired students by converting speech to text using both sound and image information. The system used DLib library and Z-order technique to extract features from the speaker's voice and lip movements. The system then combined the sound and image features into a single feature vector that represented the spoken word. The system used a deep learning-based speech recognition model that had a RNN with LSTM and CTC. The model took the feature vector as input and generated the text. The authors evaluated their model on a large-scale dataset of 500 words spoken by different speakers. They showed that their model had a high accuracy of 95% and a low word error rate of 6.59%. The authors claimed that their system was a novel, efficient, and user- friendly assistive technology for hearing impaired people. They also discussed the challenges and limitations of their system, such as the need for more data, the variability of lip shapes, and the synchronization of sound and image signals.

Timothy Israel Santos et al. in [13] introduced a visual speech recognition technique using the Inception V3 deep learning model. Unlike speaker-dependent methods, their approach was speaker-independent, adaptable to various individuals without prior training. They utilized the GRID corpus, containing audio-video samples from 34 speakers, processing lip images by resizing and converting them to black and white. Training and testing on this data, they achieved a notable accuracy of 93.8% with a 6.2% error rate, surpassing other methods. Their method's robustness was evaluated against factors like image quality, speaker diversity, and background noise, affirming its novelty, efficiency, and reliability for visual speech recognition.

Kiran Surywanshi et al. in [14] presented a new system for recognizing words from lip movements using Marathi digits. They apply deep learning methods such as CNN, VGG16, and VGG19 to learn and classify

features from silent videos of Marathi speakers. They also create their own database of Marathi digits and use various methods to process and augment the data to enhance the performance of their system. The paper shows high accuracy rates for their system and contrasts it with some existing methods. The paper aims to advance the field of visual speech recognition, especially for regional languages and noisy settings.

Pooventhiran G. et al. in [15] proposed a technique for identifying what words are spoken by looking at the speaker's lips. They did not use any sound information, which could be helpful for interacting with computers in noisy situations. They also used a new method of combining several lip movements for each word, instead of linking each lip movement to a sound. This could make the identification process more accurate and less confusing. They used a 3D CNN to analyze and classify the sequence of lip movements into text. The 3D CNN could learn both simple and complex features of the lip movements, such as edges, lines, movements, and patterns. The paper tested the technique on the MIRACL-VC1 dataset, which had color and depth images of 10 words and 10 phrases spoken by 15 speakers. The paper showed that the technique achieved an accuracy of 76.89%, which was much better than the existing techniques.

Zhi-Ming Chan et al. in [16] suggested a way to recognize words from pictures of lips without sound. They had recorded videos of 15 people speaking numbers from 0 to 9 and had taken out the lip part from each video frame. They applied a kind of neural network named VGG-M, which was a change of VGG-16, a famous neural network model for image recognition. They combined each video frame into a single image and had fed it to the neural network model to estimate the spoken word. They evaluated their model with another model named EF3 and had changed various parameters such as kernel size, learning rate, and optimizer. They achieved a validation accuracy of 87% for the seen test and 30% for the unseen test.

Pingchuan Ma et al. in [17] had presented a way to transcribe speech from lip movements in six languages, without using the audio stream. The way had used a neural network model named VGG-M, which was enhanced by adding extra objectives that had predicted audio features or phonetic labels from the lip images. The way had also used fine-tuning and data augmentations to make the model more flexible and resilient. The way had achieved state-of-the-art results on a large-scale multilingual dataset named LRS3-TEDx, and had shown that using more data, even in other languages or with automatic transcriptions, could further improve the performance.

Triantafyllos Afouras et al. in [18] proposed a way of using deep learning to perform lip reading, which was the task of recognizing what a person was saying from the visual information of their lip movements. The authors had proposed two models based on the transformer self-attention architecture, and had compared two different types of losses for training them: connectionist temporal classification (CTC) and sequence-to-sequence (seq2seq). They had shown that seq2seq loss had performed better than CTC loss, especially for longer sentences and unseen speakers. They had also investigated how lip reading had complemented audio speech recognition, especially when the audio signal was noisy or corrupted. They had shown that combining the visual and audio features had improved the recognition accuracy, and that lip reading had also helped to resolve ambiguities caused by homophones (words that sounded the same but had different meanings). They had introduced and publicly released a new dataset for audio- visual speech recognition, called LRS2- BBC, which had consisted of thousands of natural sentences from British television. They had used this dataset to train and evaluate their models, and had shown that they had surpassed the performance of previous work on a lip-reading benchmark dataset.

Souheil Fenghour et al. in [19] outlined the complexities and elements of lip-reading, including lip shape variability, visual speech ambiguity, and the scarcity of large and diverse datasets. They addressed challenges such as the computational demands of deep learning models, audio-visual database integration, and feature extraction methods. The paper compared various features and networks like raw pixels, geometric features, CNNs, RNNs, attention-transformers, and TCNs. Evaluation across datasets and tasks encompassed word recognition, sentence-level recognition, and speaker identification. Factors impacting performance, such as dataset and vocabulary size, network depth, and architecture, were analyzed. The paper explored lip-reading applications in biometric authentication, speech enhancement, and human-computer interaction, while suggesting future research directions to enhance model robustness, generalization, interpretability, and integration with other modalities like audio and facial expressions.

Stavros Petridis et al. in [20] introduces an innovative system designed to recognize speech by analyzing only the movements of the speaker's lips, without using any audio input. It utilizes a deep neural network that processes both images of the mouth and their differences to produce the spoken words as an output. Specifically tailored for small-scale datasets, which pose more realistic and challenging scenarios, the paper presents several key contributions. These include a sophisticated two-stream end-to-end model that extracts features directly from pixel data and incorporates LSTM networks to capture the temporal dynamics of lip movements. Additionally, the research thoroughly compares the effectiveness of 2D versus 3D convolutions and the performance of different optimization methods such as SGD and Adam in the realm of visual speech recognition. To validate its effectiveness, the proposed model undergoes extensive evaluation across four publicly available datasets (OuluVS2, CUAVE, AVLetters, and AVLetters2, consistently demonstrating superior performance when compared to existing state-of-the-art approaches.

Kanagala Srilakshmi et al. in [21] introduced a deep learning-based method for speech detection solely from lip movements, without relying on audio data. They employed EfficientNet B0, a variant of the ResNet-50 architecture, within a deep neural network framework. Utilizing the MIRACL-VC1 dataset, featuring videos of 12 speakers uttering 10 digits and 7 English sentences, they trained and evaluated various deep learning models. Their proposed model integrated EfficientNet B0 with an attention mechanism and LSTM network to enhance performance. Achieving a 91.13% accuracy on MIRACL-VC1, their approach surpassed existing methods.

Brendan Shillingford et al. in [22] proposed an innovative approach to visual speech recognition, focusing exclusively on lip movements to comprehend speech. They prioritize scalability for managing vast vocabularies and diverse content. The study introduces the Lip-Reading Sentences 3 (LRS3) dataset, the largest in visual speech recognition, containing 3,886 hours of video and 127,055 unique words. They develop a novel lipreading system comprising a video processing pipeline, deep neural network, and speech decoder. Evaluation on LRS3 demonstrates a significantly improved word error rate (WER) of 40.9%, surpassing existing methods like LipNet and Watch, Attend, and Spell (WAS). Comparative analysis with professional lipreaders highlights the system's robust performance, showcasing deep learning's potential in visual speech recognition complexities. The paper underscores both the feasibility and challenges of this technology, suggesting practical applications in assisting speech-impaired individuals and enhancing audio speech recognition in noisy environments.

Marzieh Oghbaie et al. in [23] conducted an review of recent developments and challenges in deep lip reading, focusing on recognizing speech through visual mouth movements. The survey addressed various topics including the motivation and applications of visual speech recognition (VSR), such as aiding audio speech recognition, enabling silent communication, and enhancing biometric security. It explored obstacles like limited and diverse datasets, suggesting solutions like data augmentation while considering ethical implications. It examined task-specific complexities like lip shape variability, head poses, and modal fusion of audio-visual cues. The paper analyzed VSR pipeline components, comparing different approaches and discussing datasets, evaluation metrics, and future research directions.

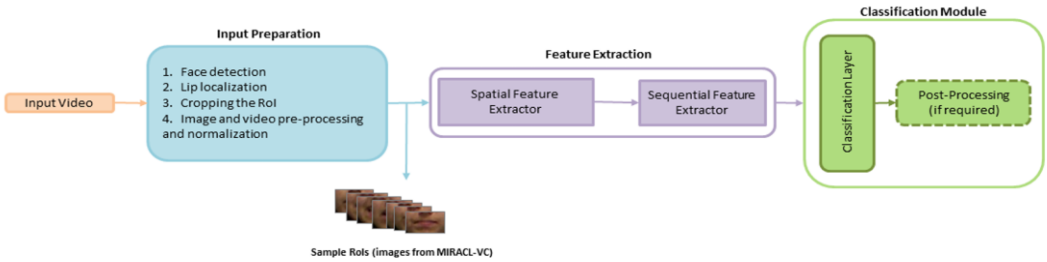

Figure 4. Baseline VSR Pipeline [23].

Figure 4 shows how to analyse and classify facial features from an input video. It involves steps like face detection, lip localization, and cropping in the input preparation stage; spatial and sequential feature extraction; and finally, classification with optional post-processing. The flowchart uses sample images from MIRACL-VC, a dataset for visual speech recognition.

Priyanka P. Kapkar et al. in [24] encompassed a comprehensive survey that focused on the pivotal tasks of lip feature extraction and lip movement recognition, crucial for biometric applications like speaker identification and verification. It delved into various facets of lip-reading systems, commencing with preprocessing steps to refine input images or videos through noise reduction, illumination normalization, mouth region cropping, and alignment of lip positions. Transitioning to feature extraction, it extensively compared diverse feature types—geometric, appearance-based, hybrid, and deep learning-based—discussing their respective strengths and limitations. Addressing the database aspect, the paper highlighted existing lip-reading databases such as M2VTS, XM2VTS, VidTIMIT, CUAVE, and GRID, outlining their characteristics and challenges. It further explored classifiers used in lip feature recognition— hidden Markov models (HMMs), support vector machines (SVMs), artificial neural networks (ANNs), and convolutional neural networks (CNNs)—assessing their performance and accuracy. Additionally, the paper delved into applications and future trajectories of lip-reading systems, spanning audio-visual speech recognition, lip synchronization, and lip animation.

Shashidhar R et al. in [25] describes a method to identify what someone is saying by watching their mouth movements in a video that has no sound. The authors made their own dataset of videos of people speaking different words and used a deep learning model that was already trained called VGG16 to sort the words by the shapes of the lips. They tested their method with another model called HCNN and saw that their method was more accurate.

The paper also talks about how visual speech recognition can help people who cannot hear and in solving crimes and finding evidence.

Tabel 1. Summary of existing work on Speech Recognition using Lip Movement

Author name	Algorithm / model used	Database used	Performance Metrics	Remarks
Hendrik Laux et al. [1] (2023)	AFE (Audio Feature estimator) and STT (Speech to text) Model	VSRICU and GRID	Error rate = 6.3%	No assessment of different speech impairment cases
Amit Garg et al. [2] (2016)	CNN AND LSTM	MIRACL-VC1	Accuracy = 76%	Real-world difficulties of lip reading not considered
Jyotsna Uday Swami et al. [3] (2021)	Snake approach, CNN, LSTM and HMM	GRID	Accuracy = 95.2 %	No attention mechanism to fuse visual features
Akshay S. Nambeesan et al. [4] (2021)	CNN AND LSTM	MIRACL-VC1	Accuracy = 85 %	Dataset restricted to Korean digits
Gerald Schwiebert et al. [5] (2022)	Feedforward and X3D CNN Models	GLips and LRW	--	Real-world difficulties of lip reading not considered
Srikanth, G. N. et al. [6] (2022)	MLP and KNN classifier	TIMID	Accuracy = 91% and 61%	No assessment of different speech impairment cases
Hameed Hira et al. [7] (2022)	Neural Network and VGG16 deep learning model	Own dataset	Accuracy = 95%	No attention mechanism to fuse visual features
Ziad Thabet et al. [8] (2018)	Gradient Boosting, SVM and logistic regression	Own dataset	Accuracy = 64.7, 63.5 and 59.4%	Real-world difficulties of lip reading not considered
Lap Poomhiran et al. [9] (2021)	CNN	THDigits, AVDigits	Accuracy = 95.06 and 85.62 %	Fixed keyframes neglect lip movements
Shashidhar Rudregowda et al. [10] (2023)	VGG16 CNN	Own dataset	Accuracy = 91.90 %	Simple concatenation of keyframe images as CNN input
Dmitry Ryumin et al. [11] (2023)	ResNet, VGG and PANN model	LRW, AUTSL	Accuracy = 98.76 %	No benchmarking with other approaches
L Ashok Kumar et al. [12] (2022)	ASM and AAM model	LibriSpeech and GRID	Accuracy = 95 % , Error rate = 6.59 %	Limited and unrealistic dataset
Timothy Israel Santos et al. [13] (2021)	Inception v3 CNN model	GRID corpus	Precision = 0.61, Recall = 0.53, F1-score = 0.51, Accuracy = 79.6%	Visual input issues concerning quality
Kiran Surywanshi et al. [14] (2023)	CNN, VGG16 and VGG19	Own dataset	Accuracy = 97.12 %	Absence of attentional visualization or analysis
Pooventhiran G. et al. [15] (2020)	3D CNN	MIRACL-VC1	Precision = 80.24, Recall = 76.89, F-measure = 77.40, Accuracy = 76.89 %	Keyframe photos combined during the CNN input process
Zhi-Ming Chan et al. [16] (2020)	VGG-M Model	CUAVE	Accuracy = 87 % (seen test) and 30 % (unseen test)	No evaluation on real-world scenarios
Pingchuan Ma et al. [17] (2022)	ASR and VSR Model	LRS2, LRS3, CMLR, CMU-MOSEAS, Multilingual TEDx, AVSpeech	--	Untested scenarios in noisy environments
Triantafyllos Afouras et al. [18] (2022)	CNN	LRS2-BBC	--	Not always available are lip characteristics
Souheil Fenghour et al. [19] (2021)	SVM, CNN, VGG and LSTM	AVLetters, CUAVE, GRID, LRW, OuluVS2, BBC-LRS2	Accuracy = 95.20 - 98.70 %	Problems with diversity and lack of data
Stavros Petridis et al. [20] (2020)	BLSTM	OuluVS2, CUAVE, AVLetters and AVLetters2	Accuracy = 93.6, 87.3, 66.3 and 36.8 %	Large data set needed for optimal performance
Kanagala Srilakshmi et al. [21] (2022)	CNN, EfficientNet and LSTM,	MIRACL-VC1	Accuracy = 91.13%.	No comparison with state-of-the-art methods

Brendan Shillingford et al. [22] (2018)	V2P and LSTM	LSVSR	Error rate = 40.9 %, Accuracy = 81.2 %	No variations in language or accent, and no scenes that are noisy
Marzieh Oghbaie et al. [23] (2021)	VSR Model	Multiple Dataset	--	High heterogeneity in lip forms and minimal data availability
Priyanka P. Kapkar et al. [24] (2019)	CNN, KNN, HMM, LSTM	M2VTS, AVletters, PKU-AV, XM2VTS	--	Costly and intricate feature extraction
Shashidhar R et al. [25] (2021)	VGG16 and Hahn CNN	Own dataset	Accuracy = 76 %	No analogy with human lipreading

III. Dataset Used

Table 2 represents the summary of the datasets used by the existing literature on Speech Recognition using Lip Movement.

Table 2. Dataset Used in existing reference papers for Speech Recognition using Lip Movement.

Dataset name	Specification
VSRICU [1]	Video and audio data of 20 patients who were in the ICU of a German hospital.
GRID [1][3][12][19]	Video and audio data of 34 people.
MIRACL-VC1[2][4][15][21]	Data of 15 people who spoke 10 words and 10 sentences each.
GLips [5]	Video and audio data of 250,000 words that were spoken by people from the Hessian Parliament in Germany.
LRW [5][11][19]	Video and audio data of 500,000 words.
TIMID [6]	Video and audio data of 630 people who speak 8 different kinds of American English.
AVDigits [9]	Data of 53 people who said numbers and sentences in three ways: normal, quiet, and without sound.
AUTSL [11]	Videos of Turkish signs. It has 226 signs of 43 people and there are 38,336 videos in all.
LibriSpeech [12]	The collection has about 1000 hours of speech in all.
CUAVE [16][19][20]	Video and audio data of 34 speakers who spoke 1000 sentences each.
LRS2[17]	1000 hours of speech in total spoken by various speakers from BBC television.
LRS3[17]	1000 hours of speech in total spoken by various speakers from BBC television.
CMLR [17]	Video and audio data of 102,072 sentences spoken by 11 speakers in Mandarin Chinese.
CMU-MOSEAS [17]	Video and audio data of 40,000 sentences in Spanish, Portuguese, German, and French.
Multilingual TEDx [17]	Audio and video data of sentences spoken by various speakers from TEDx talks in 8 languages: Spanish, French, Portuguese, Italian, Russian, Greek, Arabic, and German.
AVSpeech [17]	Data of 4700 hours of video segments with approximately 150,000 distinct speakers
LRS2-BBC [18][19]	Has 1000 hours of speech in total
AVLetters [19][20][24]	Video and audio data of 26 speakers who spoke 30 sentences each. The sentences are single letters from A to Z.

IV. Conclusion

Lip movement recognition is an interesting and difficult field of research that has many uses in different areas, such as security, deafness, and noise. In this paper, we surveyed the current literature on lip movement recognition methods, paying attention to the visual features, the deep learning models, and the datasets used for testing. We also talked about the benefits and drawbacks of lip movement recognition, as well as the future trends and challenges in this field. We found that lip movement recognition can add to the information from audio signals and enhance the accuracy and reliability of speaker recognition systems. However, there are still many problems

to be solved, such as the variation of lip movements among speakers, languages, and situations, the scarcity of large and diverse datasets, and the demand for more effective and precise deep learning models. We hope that this paper can be a helpful guide for researchers and practitioners who are curious about speech recognition using lip movement.

References

[1]. H. Laux, A. Hallawa, J. C. S. Assis, A. Schmeink, L. Martin And A. Peine, "Two-Stage Visual Speech Recognition For Intensive Care Patients", Scientific Reports, 13(1), 928, January 2023. Available: Https://Www.Nature.Com/Articles/S41598-022-26155-5.

[2]. A. Garg, J. Noyola And S. Bagadia, "Lip Reading Using Cnn And Lstm", Technical Report, Stanford University, Cs231 N Project Report, 2016. Available: Http://Vision.Stanford.Edu/Teaching/Cs231n/Reports/2016/Pdfs/217_Report.Pdf.

[3]. J. U. Swami And J. S. R., "Lip Reading Recognition", Jetir, Volume 8, Issue 5, May 2021. Available: Https://Www.Jetir.Org/Papers/Jetir2105723.Pdf.

[4]. A. S. Nambeesan, C. Payyappilly, E. J. C., J. John P, S. Alex, "Lip Reading Using Facial Feature Extraction And Deep Learning", International Journal Of Innovative Science And Research Technology, Volume 6, Issue 7, July – 2021. Available: Https://Ijisrt.Com/Assets/Upload/Files/Ijisrt21jul035.Pdf.

[5]. G. Schwiebert, C. Weber, L. Qu, H. Siqueira And S. Wermter, "A Multimodal German Dataset For Automatic Lip Reading Systems And Transfer Learning", Arxiv Preprint Arxiv:2202.13403, May 2022. Available: Https://Arxiv.Org/Abs/2202.13403.

[6]. G. N. Srikanth And M. K. Venkatesha, "Word Recognition Through Mapping Of Lip Movements From Speech Utterance Using Audiovisual Fusion And Mlp", International Journal Of Health Sciences, 6(S2), 4533-4545, April 2022. Available: Https://Www.Neliti.Com/Publications/429450/Word-Recognition-Through-Mapping-Of-Lip-Movements-From-Speech-Utterance-Using-Au.

[7]. H. Hameed, M. Usman, A. Tahir, A. Hussain, H Abbas, T. J. Cui, M. A. Imran And Q. H. Abbasi, "Pushing The Limits Of Remote Rf Sensing By Reading Lips Under The Face Mask", Nature Communications, 13(1), 5168, September 2022. Available: Https://Www.Nature.Com/Articles/S41467-022-32231-1.

[8]. Z. Thabet, A. Nabih, K. Azmi, Y. Samy, G. Khoriba And M. Elshehaly, "Lipreading Using A Comparative Machine Learning Approach", In 2018 First International Workshop On Deep And Representation Learning (Iwdrl), Pp. 19-25. Ieee, May 2018. Available: Https://Ieeexplore.Ieee.Org/Abstract/Document/8358210/.

[9]. L. Poomhiran, P. Meesad And S. Nuanmeesri, "Improving The Recognition Performance Of Lip Reading Using The Concatenated Three Sequence Keyframe Image Technique", Engineering, Technology & Applied Science Research, 11(2), 6986-6992, April 2021. Available: Http://Www.Etasr.Com/Index.Php/Etasr/Article/View/4102.

[10]. S. Rudregowda, S. P. Kulkarni, G. H. L., V. Ravi And M. Krichen, "Visual Speech Recognition For Kannada Language Using Vgg16 Convolutional Neural Network", In Acoustics (Vol. 5, No. 1, Pp. 343-353). Mdpi, March 2023. Available: Https://Www.Mdpi.Com/2624-599x/5/1/20.

[11]. D. Ryumin, D. Ivanko And E. Ryumina, "Audio-Visual Speech And Gesture Recognition By Sensors Of Mobile Devices", Sensors, 23(4), 2284, February 2023. Available: Https://Www.Mdpi.Com/1424-8220/23/4/2284.

[12]. L. A. Kumar, D. K. Renuka, S. L. Rose And I. M. Wartana, "Deep Learning Based Assistive Technology On Audio Visual Speech Recognition For Hearing Impaired", International Journal Of Cognitive Computing In Engineering, 3, 24-30, June 2022. Available: Https://Www.Sciencedirect.Com/Science/Article/Pii/S2666307422000031.

[13]. T. I. Santos, A. Abel, N. Wilson And Y. Xu, "Speaker-Independent Visual Speech Recognition With The Inception V3 Model", In 2021 Ieee Spoken Language Technology Workshop (Slt) (Pp. 613-620) Ieee, March 2021. Available: Https://Ieeexplore.Ieee.Org/Abstract/Document/9383540/.

[14]. K. Surywanshi, K. Shinde And C. Kayte, "Deep Learning-Based Visual Speech Recognition System Using Marathi Digit", In Biogecko, Vol 12 Issue 02 2023 Issn No: 2230-5807, February 2023. Available: Https://Www.Biogecko.Co.Nz/Admin/Uploads/Nc-Sco-104.Pdf.

[15]. G. Pooventhiran, A. Sandeep, K. Manthiravalli, D. Harish, And R. D. Karthika, "Speaker-Independent Speech Recognition Using Visual Features", International Journal Of Advanced Computer Science And Applications, Vol. 11, No. 11, 2020. Available: Https://Www.Academia.Edu/Download/79780798/Paper_75-Speaker_Independent_Speech_Recognition.Pdf.

[16]. Z. M. Chan, C. Y. Lau And K. F. Thang, "Visual Speech Recognition Of Lips Images Using Convolutional Neural Network In Vgg-M Model", Journal Of Information Hiding And Multimedia Signal Processing, Volume 11, Number 3, Issn 2073-4212, September 2020. Available: Https://Bit.Nkust.Edu.Tw/~Jihmsp/2020/Vol1/2_Jihmsp-1522_Vol3.Pdf.

[17]. P. Ma, S. Petridis And M. Pantic, "Visual Speech Recognition For Multiple Languages In The Wild", Nature Machine Intelligence, 4(11), 930-939, October 2022. Available: Https://Www.Nature.Com/Articles/S42256-022-00550-Z.

[18]. T. Afouras, J. S. Chung, A. Senior, O. Vinyals And A. Zisserman, "Deep Audio-Visual Speech Recognition", Ieee Transactions On Pattern Analysis And Machine Intelligence, 44(12), 8717-8727, December 2022. Available: Https://Ieeexplore.Ieee.Org/Abstract/Document/8585066/.

[19]. S. Fenghour, D. Chen, K. Guo, B. Li And P. Xiao, "Deep Learning-Based Automated Lip-Reading: A Survey", Ieee Access, 9, 121184-121205, August 2021. Available: Https://Ieeexplore.Ieee.Org/Abstract/Document/9522117/.

[20]. S. Petridis, Y. Wang, P. Ma, Z. Li And M. Pantic, "End-To-End Visual Speech Recognition For Small-Scale Datasets", Pattern Recognition Letters, 131, 421-427, January 2020. Available: Https://Www.Sciencedirect.Com/Science/Article/Pii/S0167865520300349.

[21]. K. Srilakshmi And R. Karthik, "A Novel Method For Lip Movement Detection Using Deep Neural Network", Journal Of Scientific & Industrial Research, 81(06), 643-650, June 2022. Available: Http://Op.Niscpr.Res.In/Index.Php/Jsir/Article/View/53898.

[22]. B. Shillingford, Y. Assael, M. W. Hoffman, T. Paine, C. Hughes, U. Prabhu, H. Liao, H. Sak, K. Rao, L. Bennett, M. Mulville, B. Coppin, B. Laurie, A. Senior, N. D. Freitas, "Large-Scale Visual Speech Recognition", Arxiv Preprint Arxiv:1807.05162, October 2018. Available: Https://Arxiv.Org/Abs/1807.05162.

[23]. M. Oghbaie, A. Sabaghi, K. Hashemifard And M. Akbari, "Advances And Challenges In Deep Lip Reading", Arxiv Preprint Arxiv:2110.07879, October 2021. Available: Https://Arxiv.Org/Abs/2110.07879.

[24]. B. S. Priyanka And S. D. Bharkad," Lip Feature Extraction And Movement Recognition Methods: A Review", International Journal Of Scientific Technology Research, (8), 50-55, August 2019. Available: Https://Www.Semanticscholar.Org/Paper/Lip-Feature-Extraction-And-Movement-Recognition-A-Kapkar-S.D.Bharkad/71ccf23123f868e708097e1190c6971fa743e1f6.

[25]. S. R. And S. Patilkulkarni, "Visual Speech Recognition For Small Scale Dataset Using Vgg16 Convolution Neural Network", Multimedia Tools And Applications, 80(19), 28941-28952, June 2021. Available: Https://Link.Springer.Com/Article/10.1007/S11042-021-11119-0.