

# Crime Investigation Using Lip Reading\*

Mansimran Kaur <sup>1,\*, †</sup>, Dhruv Rastogi <sup>2, †</sup>, Anjali Sharma <sup>3, †</sup>, Anushka Dahiya <sup>4, †</sup>,  
Preeti Nagrath <sup>5, †</sup>

<sup>1,2,3,4,5</sup> Department of Computer Science and Engineering, Bharati Vidyapeeth's College of Engineering, New Delhi, India

## Abstract

The use and development of lip-reading techniques have been revolutionized through the incorporation of deep learning architectures and visual speech analysis which has greatly helped crime investigations whereby investigators can understand conversations using surveillance videos without sound. This capability has been found to be very useful in solving many cases with a lot of complications. In contrast to prior studies using the Grid Corpus dataset that relied on known models such as LipNet that have had high accuracy, we aim to train a new lip-reading model from scratch using a mix of Gated Recurrent Unit (GRU) networks and Convolutional Neural Networks (CNN). This innovative methodology has made 86.17% of accuracy in the given problem. The conclusion that can be drawn from this work points to the significance of this project in playing a positive role in criminal investigations, specifically within the areas of enhanced lip-reading analysis using deep learning technology and helping the law enforcement agencies in improving their ability to understand visual speech from security cameras.

## Keywords

Lip-reading, Deep Learning, Convolutional Neural Networks, GRU, Crime Investigation, Forensic Analysis, Evidence Analysis

## 1. Introduction

Computer vision and language understanding technology have for instance been used in lip reading to enhance the interaction between man and computer. This technology has brought a great change in the lives of the people with hearing impairments since it offers them a device that enhances better means of communicating. Besides its usage in personal life, lip reading also brings additional enhancements to security systems and the quality of surveillance and evidence. Also, it has a significant function in the creation of the assistive technologies which assist in filling communication gaps in different areas, including police-community-police interaction, patient-doctor interaction, and so on; thereby making such interactions more effective and efficient.


Lip reading is important not only in facilitating communication but also useful for people with hearing impairments. It lets them understand what has been said through oral communication by the gestures, lip movements, facial expressions, etc. Thus, in the conditions of acoustic interference or when all audible signals are distorted, lip reading acts as useful signal in addition to auditory information. It also leads to better interaction with people as the clients can be made to attend to conversations better than before. Apart from its use to the deaf and the hard of hearing, lip reading is important in areas such as law enforcement and security, where speech without sound is important in several practical scenarios.

---

SCCTT-2024: International Symposium on Smart Cities, Challenges, Technologies and Trends, 29th Nov 2024, Delhi, India  
\* Corresponding author.

† These authors contributed equally.

✉ mansimran2703@gmail.com (Mansimran Kaur); dhruvrastogi43@gmail.com (Dhruv Rastogi);  
anjali2609@gmail.com (Anjali Sharma); dahiya0502@gmail.com (Anushka Dahiya);  
preeti.nagrath@bharatvidyapeeth.edu (Preeti Nagrath);

 0000-0002-5863-3113 (Preeti Nagrath)

© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

The progress in the last few years has been significantly boosted by the large datasets such as the LRW (Lip Reading in the Wild) and the Grid Corpus. Specifically, the research activities related to the LRW dataset have been boosted in the past years. Many researches have shown good performances and high hit rates, for instance, LipNet model that has been proven to be efficient in recognizing basic sentence by detecting lips movement with an accuracy of more than 95%. These achievements highlight a lot of progress that has been made in the application of existing paradigms for the accurate interpretation of visual speech gestures.

However, our approach to solving the problems in this area is significantly different from the typical practice of using ready-made models. The main aim of this study is to investigate the dynamic relationship between visual speech input and the factors associated with sentence prediction. As such, we have started with the challenging goal of building a new lip-reading model from the ground up. This new model is designed based on both Convolutional Neural Networks (CNNs) and Bidirectional Gated Recurrent Unit (GRU) networks, so that we can explore the variations in the architecture and layers to better understand the working of lip reading.

In this paper, the major emphasis is placed on the Grid Corpus dataset, and while our goal goes beyond simply comprehending the nuances of visual speech, we strive to decipher the intricacies of the most advanced sentence construction. Over the course of this project, we have been able to hit an accuracy rate of 86.17%. This success is due to our efforts to respond to the peculiarities of the Grid Corpus dataset, with a keen emphasis on the nuances that matter for lip reading.

By building a new model tailored to the unique characteristics of this dataset, we aspire to enhance the capabilities of lip-reading systems and further the understanding of visual speech processing. Our research not only highlights the potential for improved accuracy in lip-reading applications but also sets the stage for future advancements in this exciting area of study. As we continue to delve into the nuances of visual speech recognition, we hope to contribute to the broader goal of developing more robust and reliable communication technologies.

Lip-reading technologies have the potential of improving the investigation of crime processes through the following reasons. When the video recording has no clear audio, or the audio is missing, low quality, or has been altered by someone with a malicious intent, or contains background noise, these technologies come handy. Lip-reading systems also work by detecting vocal gestures and kinematics involving mouth movements, facial expressions, and gestures to translate spoken words and discover important information that may be lost. This capability is especially useful in situations where it is impossible to take an audio capture, for instance, in espionage or important protection surveillance. In other words, lip-reading technology provides law enforcement with a tool to get valuable dialogue from mute videos that in turn helps them to assemble the evidence, comprehend the perpetrators' motives, and solve crimes more efficiently. As the technology progresses, this integration into investigations could change how LEA's deal with cases where traditional audio doesn't suffice.

The research work of this paper is intended to solve one of the biggest problems in the area of criminal investigation to help make a change by improving the efficiency of the police. To this end, we have carefully developed a novel model that employs deep learning architectures, including both CNN and GRU models. The aim of this work is to identify speech from the video data that is captured in surveillance videos where audio may be missing or of poor quality.

This revolutionary method has the potential of greatly boosting the efficiency of investigation in police work since visual data, which is usually silent, can be turned into useful information. As a result of translating visual signals into comprehensible language, our model can reveal important information, find suspects, and provide other information that is essential for solving crimes. The significance of this study is thus significant to law enforcement organisations as it provides a state-of-art solution that can identify key conversational and exchange occurrences in video footage and enhance crime solving. By way of this work, we hope to provide investigators with the means to do their work better and, in so doing, bring more safety and security into society.

## 2. Related Work

This section gives an outline of all the different forms of automatic lip reading that are in practice in the present times. In the past, there has been limited application of deep learning methods to automated lip reading; most of the previous work employed different strategies. Such approaches were often accompanied by the need to extract image features from frames or video features from video sequences. Some of the previous techniques used were optical flow, movement detection and other manually designed vision pipelines [7], [8], [9], [10], [11], [13].

Due to the vast amount of literature available on automated lip reading, it is impossible to discuss all the areas in detail in this context. For the more detailed information, the literature review of the current state of art in lip reading technologies will help to understand the development and the current state of the technologies.

Goldschen was the first to work on visual sentence-level speech recognition using HMMs on a small hand segmented phones dataset [1]. This was an important step in beginning to investigate the possibility of visual speech. On this basis, Neti further extended the work by developing the first audio-visual speech recognition system that operates at the sentence level using HMMs with specific features that were trained on the IBM ViaVoice dataset [2]. Their revolutionary work combined both, visual and audio data which enhanced the speech recognition in noisy environment. However, it is necessary to point out that the dataset used in this research, including 17,111 utterances from 261 people, or approximately 34.9 hours, is not publicly accessible. Another important point is that their visual-only results do not contain individual visual recognition outputs, but are used to rescore noisy audio-only predictions.

Potamianos continued this work, obtaining WER of 82.31% and 91.62 % for speaker adapted and speaker independent models respectively using the same IBM ViaVoice dataset [3]. For speaker adapted models, on the connected DIGIT corpus – a dataset consisting of phrases with digits, he got WERs of 38.53% and for speaker independent models, WERs of 16.77%. These improvements demonstrate the advancements in the methodologies used for AVSR, where specific advancements have been made in speaker adapted and speaker independent cases across different datasets.

Furthermore, Gergen used a speaker-dependent training with the GMM/HMM system on the mouth regions only after applying LDA on the DCTs. This approach resulted in outstanding speaker-dependent accuracy of 86.4% which set a new benchmark for GRID corpus dataset. However, some issues were still present, such as the ability to generalize performance to other speakers and in extracting motion features [16]. To overcome these limitations, LipNet was considered as the best solution that provides improvements in the generalization of the speaker and in the extraction of motion related features in lip reading tasks.

In recent years, deep learning has gained increased attention in the lip reading area, but most of the work has still been centered on word or phoneme recognition. LipNet, however, is different from other methods as it goes further to predict the whole phrases in the sequence rather than

predicting single words. Current approaches in this area tend to focus on learning multimodal audio-visual features [18] [20] [21] and incorporating video into conventional speech recognition models including GMM-HMM and HMM for word and character recognition [24] [26]. Sometimes, these approaches use one or more than one technique [4].

Malek Miled and his research team document a holistic algorithm for lip-reading that adequately combines advanced image processing methods with deep learning methodologies. With this end, the study advocates an innovative hybrid model in which an edge-based filter is utilized to separate the mouth region, thus enhancing the accuracy of lip movement detection. The combination of CNN with Bi-GRU ends up in a robust model that is highly sensitive to the spatio-temporal mechanisms of lip movement. The algorithm thus achieved an excellent accuracy of 90.38% in testing, which indicates significance improvement in efficiency compared with traditional methods. This research not only pushes forward the domain of lip reading but has also opened up potential applications in silent communication as well as speech recognition technologies [34].

Mini-3DCvT is a newly designed lipreading technique focusing on the complexity needed to adequately extract visual spatial characteristics, temporal dynamics, and at the same time maintain a lightweight model structure [33]. The technique combines visual transformers with 3D convolution for apt capturing of spatiotemporal local and global attributes in a sequence of continuous images. Weight transformation and distillation strategies come into play within the architectures of both convolution and transformer that make the model more streamlined to drastically improve its efficiency. The method manifests itself in a high recognition accuracy, scoring 88.3% on the LRW dataset and 57.1% on LRW-1000, with good computational complexity and a minimal number of parameters.

Co-adaptation of feature detectors in neural networks is a problem that Hinton et al. address in their paper. They propose a method known as dropout which during the training phase, certain neurons are switched off at random. This approach also helps the network to create better feature representations in a way that it doesn't overly depend on one neuron. The authors provide convincing proof of dropout improving accuracy for different architectures and tasks of neural networks. Furthermore, they explain the theory behind dropout and, as a result, improve the understanding of how neural networks work and become popular in deep learning areas such as computer vision and natural language processing [27].

In their paper, Gergen et al. introduce a dynamic stream weighting technique for turbo-decoding-based audiovisual automatic speech recognition (ASR) [25]. This method enhances the integration of audio and visual information, allowing for more accurate speech recognition in challenging environments. They demonstrate that adapting the weighting of audio and visual streams dynamically can improve performance significantly. Meanwhile, Haliassos et al. (2020) present a robust approach to face forgery detection, emphasizing that "lips don't lie." Their work focuses on developing a generalizable model capable of accurately identifying manipulated facial features, contributing to advancements in security and integrity verification in visual media [26].

In another paper published in 2024, Robin Anburaj B contributes to the field of lip reading by introducing a vision-based system that effectively combines a convolutional neural network (CNN) with an attention-based Long Short-Term Memory (LSTM) architecture. By leveraging pre-trained CNN models, the study enhances feature extraction from processed video frames, which are crucial for understanding temporal characteristics of lip movements. The system achieves an impressive 80% accuracy using TensorFlow and ensemble learning techniques, demonstrating its potential for practical applications. Furthermore, the research highlights the

importance of integrating machine learning with visual speech analysis to advance communication accessibility, particularly for individuals with hearing impairments [28].

This paper contributes a lot towards lip reading since it filled up a gap of similar such Turkish-language datasets [30]. For that reason, video data were recorded from 72 different people pronouncing 71 different words involving audio streams as much as possible to the greatest extent visual information was focused. The replication was done through the Camtasia application for increasing the dataset size and diversity. The percentages used to test this proposed model are as follows: adjectives with 71.8%, nouns with 71.88%, and verbs with 79.69%. This work is of help to fill gaps in Turkish lip reading resource while achieving an enhancement on communication aids for the hearing impaired.

Chung and Zisserman were able to contribute by using spatial and spatiotemporal convolutional neural networks based on the VGG architecture for classifying words [5]. The authors in their studies, on a dataset from BBC TV with 333 and 500 classes, showed that ST models are less accurate than S models by an average of 14%. Furthermore, their models were restricted in modeling sequences of different lengths, and they did not consider the sequence prediction at the sentence level, which could be explored in more detail in other tasks.

In a different approach, Garg used a pre-trained VGG model that targets faces to classify words and phrases, although their study was carried out on a small MIRACL-VC1 dataset, which has only 10 words and expressions each [6]. Garg's strongest recurrent model used a training method that froze the VGGNet parameters while training RNN, a method that deviated from joint training methods used in other models. Even when working with a very simple dataset and the classification tasks were restricted to only 10 classes, the authors' model obtained fairly low, yet still reasonable, accuracies of 56.0% for word classification and 44.5% for phrase classification.

In contrast to these efforts, LipNet represents a major breakthrough in optical speech recognition by offering a fully end-to-end model capable of predicting sequences at the sentence level. LipNet's approach is distinct in that it generates sequences of tokens directly from an input series of images, eliminating the need for explicit alignments. This is achieved through its use of Connectionist Temporal Classification (CTC) during training, which allows the model to learn sequences without requiring precise frame-by-frame labeling. LipNet's end-to-end structure marks a significant advancement in the field, showcasing the potential for more robust, comprehensive, and context-aware visual speech recognition systems.

A novel paper makes important contributions to the literature of lip reading because it proposes a novel deep learning model that maps directly a video sequence of lip movements into text transcriptions [29]. The system works pretty well using this end-to-end architecture combining 3D convolutional neural networks with bidirectional Long Short-Term Memory networks, effectively interpreting visual cues within motions. For benchmark datasets, the model shows remarkable performance with character error rate being 1.54% and the word error rate being 7.96%. Such headways not only offer the hearing-impaired more precise lip-reading technologies but also promote accessibility for them to have unobstructed communication in challenging auditory environments.

A 2024 paper contributes quite a lot to lip-reading research as the first large-scale Korean dataset on lip reading is incorporated, which comprises over 120,000 utterances derived from diverse TV broadcasts, including news, documentaries, and dramas [31]. The article is designed to help fill the gap in existing resources for Korean lip reading, which had previously been substantially underexplored compared to English. However, the authors suggest a strong

preprocessing method to extract a consistent region of interest for facial parts and introduce a transformer-based model concentrated on grapheme units for efficient sentence-level analysis. Experimental results validate the effectiveness of this dataset and model and hence pave the way for further possible developments in Korean lip-reading technology.

Another recent research makes major contributions in the area of lip reading by proposing a comprehensive Cantonese sentence-level lip-reading dataset with over 500 unique speakers and more than 30,000 samples [32]. It especially appeals to the relatively low number of Cantonese datasets compared to the mushrooming Mandarin ones. The research boasts a rich pipeline of dataset collection and construction, including a new visual frontend: the 3D-visual attention net, a combination of convolutional and self-attention mechanisms for the detailed representation of lip-region features. Coupled with an effective backend conformer in modeling temporal sequences, this laid foundation for highly valuable future research into dialect-specific lip reading.

Analysis of some publicly available datasets, such as LRW, OuluVS, CUAVE, and SSSD, is conducted in the study [35]. Advanced deep learning models are elaborately examined for lip reading at the word level. Observations made on various state-of-art architectures during this study resulted in achieving new accuracy while lip-reading, significantly on the LRW dataset, a surprise from 66.1% to 94.1%. The conducted research combines well-established models, using which the effectiveness of ResNet, WideResNet, EfficientNet, MS-TCN, and ViViT were improved by using alternative modified variants of feature extractors and classifiers. According to the results, for feature extraction, settings of 3D-Conv + ResNet18 as well as the MS-TCN model selection for inference enable generalization over various datasets and lead to better performance in tasks of lip reading.

The evaluation of LipNet leverages the GRID corpus, chosen for its sentence-level structure and large dataset. The phrases within this corpus follow a well-defined grammatical pattern, consisting of six distinct word categories: command (4 options), color (4 options), preposition (4 options), letter (25 options), digit (10 options), and adverb (4 options). Each category contains a specific set of possible words, such as {bin, lay, place, set} for commands, {at, by, in, with} for prepositions, and {blue, green, red, white} for colors. The letter category ranges from A to Z, with W included separately, while digits span from zero to nine, and the adverbs are drawn from {again, now, please, soon}. This structured combination results in a total of 64,000 possible sentence configurations. Example sentences from the dataset include statements like "place red at C zero again" and "set blue by A four please," illustrating the variety and complexity of potential phrases.

### **3. Methodology**

The main objective of this study is to improve lip reading through the use of deep learning approach with the use of Conv3D and GRU networks.

The study uses GRID Corpus dataset not only because it is available but also because of the variety of speakers and the variety of lighting conditions. These characteristics make it ideal for the research objective which is to construct a generalized lip-reading model that is effective in all conditions.

In the case of video data, pre-processing is done in the following steps to ensure that the input data is ready for model training. First, frames are captured from the video stream then the frames are converted to the right color space. The face region of interest, particularly the lip region, is then extracted from each frame, and then the illumination is corrected for any variation. The frames are also uniformly resized to a standard resolution. Next, label encoding is performed on the dataset followed by the tokenization and converting into numerical form which is suitable for the model output layer format.

The proposed model architecture integrates several key components: Conv3D layers to learn spatial and temporal features, Activation layers to inject non-linearity and MaxPooling3D layers to down sample the feature maps. Time Distributed and Flatten layers are used to pass the data for sequential model and Bidirectional GRU layers are employed to capture both past and future contexts in lip movements. To reduce overfitting, Dropout layers are used and the final layer is an output layer with Dense layer to give the final predictions. All of them are important in the process of making the final decision on lip movements and the corresponding sentence in order to have the best performance of the model.

This combination of methods and the systematic pre-processing of the GRID corpus is intended to generate a sound and general lip-reading model. role in capturing spatial and temporal features crucial for lip reading. The Connectionist Temporal Classification (CTC) loss function is applied with the Adam optimizer and 0.01 as the learning rate to train the model, and early stopping mechanisms to prevent overfitting. The dataset is divided into training and validation sets, besides accuracy is chosen as the evaluation metric.

This research addresses the gap in lip-reading methodologies by proposing a deep learning model that combines Conv3D and GRU networks. The study employs the Grid Corpus dataset for its diversity and aligns with ethical considerations.

### **3.1 Data Collection:**

This work also benefits from the Grid Corpus created by Oxford University as a substitute for the datasets that are not available to the public, including LR2 and LRW. This feature-rich dataset contains a diverse set of linguistic material, which includes speakers with various accents and speaking styles, which is particularly important for training effective lip-reading models.

Furthermore, the Grid Corpus is intended to cover different lighting conditions, so the model will be able to learn lip movements under different visual situations. Notably, the dataset is accompanied by fine-grained annotations associating spoken words with lip movements, which are instrumental in training and testing models intended for analyzing VSC information. In this way, the Grid Corpus greatly improves the functionality of the automated lip reading tools for researchers and developers in the field, and thus helps to advance the use of lip reading in assisting technologies and criminal investigations.

**Table 1:**  
Structure of Grid Corpus Dataset

<b>Data</b>	<b>Audio</b>	<b>Front</b>	<b>Side</b>	<b>Alignment</b>	<b>Meta</b>
		<b>Video</b>	<b>Video</b>		<b>Data</b>
Size (MB)	651.4	837.1	870	2	0.062

### 3.2 Video Pre-processing:

Only frames are extracted evenly from the video sequences, which synchronously and effectively samples the dynamic lip movements necessary for lip reading. Once extracted, these frames are converted to grayscale since the analysis does not require all the colors, only the necessary information in order to carry out the analysis. Next, lip regions are extracted by face landmark detection which locates a set of important facial landmarks and allows the model to concentrate on the areas of interest only.

In addition to reduce the effect of illumination on the model, lighting normalization is performed on the dataset. This step assists in making the dataset more uniform since issues such as variations in lighting will be minimized from the actual video clips. Further, the frames are scaled to a fixed size to standardize the format of the data set and to ease the computational processing during the training of the models. All these preprocessing steps are very important in the preprocessing of the dataset and fine tuning of the dataset to enhance the efficiency of training and enhancing the lip-reading model.

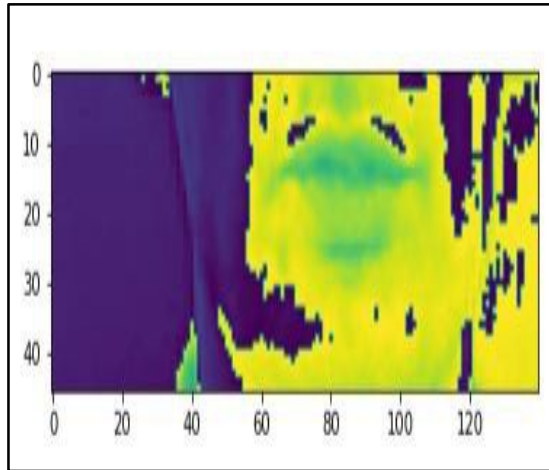
### 3.3 Label Encoding:

Since text processing is always pronounced on text, it is very important to divide spoken phrases in preparation for processing, by converting these phrases into handy tokens which are referred to as text tokens. Such approach can let the model analyze the components of the speech separately, which, in turn, makes the task easier.

After tokenization, a mapping from characters to numbers takes place to reduce these tokens to numbers. This encoding process maps each of them to a specific integer so that the model can understand and comprehend it well.

The adoption of numerical representations is critical due to the fact that the result of most machine learning models, specifically the output layer, frequently employs numerical data for processing. First and foremost, encoding helps to define a better mapping from characters to numbers which helps the learning step better in finding patterns and relations in the data. It also creates a solid base for the further work of the model, which can predict the desired accuracy based on the set of encoded spoken phrases.

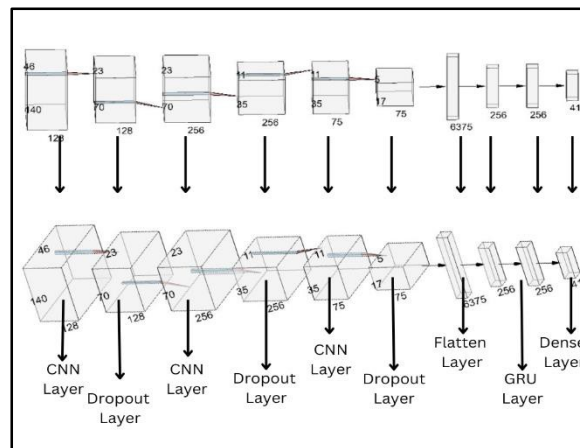




**Figure 1:** Grey-scaled and normalized frame extracted from video

### 3.4 Model Architecture:

The proposed architecture integrates Conv3D layers for spatial-temporal features, Activation layers for non-linearity, MaxPooling3D layers for down sampling, Time Distributed and Flatten layers for temporal processing, Bidirectional GRU layers for sequential modelling, Dropout layers for regularization, and a Dense output layer for predictions.



**Figure 2:** Lip Reading Model Architecture

### 3.5 Model Training:

The Grid Corpus data set is divided systematically by the `tf.data` module of TensorFlow into training and validation. In training, it engages 10-epoch training with the training set and further configures several important parameters aiming to optimize performance. Specifically, it is evident that the loss function Connectionist Temporal Classification aptly fits its usage on sequence-based tasks like lip reading due to its flexibility with respect to aligning input sequences and output labels. As for optimization, the Adam optimizer is also used with a learning rate of 0.01, which is a good compromise between convergence speed and stability.

The mechanisms for early stopping added in the training procedure monitor how the model is performing and stops when improvements become stagnant. This is an efficient technique for preventing overfitting, whereby the model would generalize well for data yet to be seen.

Generally, these strategies will make a potent training framework that will contribute favorably to boost the ability of the model in learning and predicting visual speech patterns based on the Grid Corpus dataset.

## 4.Result and Analysis

### 4.1. Result

The Connectionist Temporal Classification (CTC) loss function used plays a crucial role in the training process of the introduced model. This particular loss function is specifically tailored to cope with the difficulties of the task, which is the sequence based tasks like lip reading. Another advantage of using CTC is wanting in flexibility in managing different output lengths, which is essential in lip reading since the duration by which words are spoken may not necessarily correspond to the frame found in the video. CTC helps the model to learn the alignments between the input sequences, which are the frames of lip movement video and the related target labels while the timing for each phoneme does not need to be annotated accurately. This flexibility is especially helpful in those cases where it is challenging to provide accurate labels for speech data.

After the training process carried out for 10 epochs, the desirable accuracy of the model was reached and it constituted 86.17%. This outstanding performance goes on confirming the efficiency of the architecture in perceiving the spatial and temporal dynamics related to lip movements. The capacity to perceive these subtleties is important so that the real picture of the information contained in speech can be seen. The CTC loss function was particularly significant in this success since it allowed the model to match predictions with target label values.

CTC effectively deals with the problems connected with the temporal relation of phoneme occurrences, which in natural speech may occur at different time points. The fact is that traditional methods that rely on the strict correspondence of time points to certain events may fail in this case, while the CTC approach enables the model to deal with the shift between the lip movements and phoneme production. CTC in a way improves the generalization of the model by allowing the model to predict the most probable sequence of outputs given the sequence of inputs, hence the lip-reading predictions will be accurate even if the speaking style, speed etc. changes.

The effectiveness of the model at the level of accuracy is quite gratifying, yielding a high rate of 86.17%. The reason for such success of the network can be explained by the very structure of the network, which is a combination of convolutional layers and recurrent layers. This architectural design allows the model to learn local spatial features which are necessary for identifying lip motion within a particular temporal frame and, at the same time, learn long range temporal dependencies which are also important for recognizing lip movements.

Also, it is necessary to draw attention to the fact that the convolutional layers play the most essential part in the model for feature extraction. These layers perform well in their role by applying successive convolutional filters on the input matrix which makes it easier to extract features from the video frames. Consequently, they can gather details such as the shape of lips and their movements during the entire video, which will help the model to distinguish small peculiarities connected with definite phonemes. This capability greatly improves the model's performance and is a major contribution to the efficiency of the lip-reading process. With this, it is easier for the model to learn and interpret visual features that are essential for speech recognition.

On the other hand, Gated Recurrent Unit (GRU) makes the model more resistant because prediction is done using the previous and the subsequent frames. This capability is especially relevant when the information may be sporadic or intermittent as is the case with lips movement of the speaker as well as variations that may be noted with time. Therefore, the GRU improves

the temporal information from both directions and expands the scope of where the lip movements occur. This temporal awareness is helpful for the sequential aspects related to visual speech in a way that even in unfavorable conditions, better prediction is provided.

In totality, the proposed model is a comprehensive and efficient model that incorporates CNNs and GRUs for the purpose of extracting local spatial features and for managing long-term dependencies, respectively. This dual use is beneficial to the model's operation in various circumstances and indicates a high level of resilience. This shows that the architecture is viable in practice where the observer may not have all the information necessary to perform the task as this study has shown that it is capable of performing well even under these circumstances will without doubt make it a worthwhile tool not just in lip reading but in many other related applications. With further advancements in research and development, this model could potentially lead to even more significant improvements in visual speech recognition systems.

```

print('*100, 'REAL TEXT')
[tf.strings.reduce_join([num_to_char(word) for word in sentence]) for sentence in [sample[1]]]
-----
[ctf.tensor: shape=(), dtype=string, numpy=b'bin red at s nine again']
-----
yhat = model.predict(tf.expand_dims(sample[0], axis=0))
1/1 [*****] - 1s 1s/step
-----
decoded = tf.keras.backend.ctc_decode(yhat, input_lengths=[75], greedy=True)[0][0].numpy()
print('*100, 'PREDICTIONS')
[tf.strings.reduce_join([num_to_char(word) for word in sentence]) for sentence in decoded]
-----
[ctf.tensor: shape=(), dtype=string, numpy=b'bin red at s nine again']
-----

```

**Figure 4:** Predicted Result of the Model

**Table 2:** Performance of the model compared to BiLSTM

Criteria	BiLSTM	GRU
Accuracy	75%	86.17%
Validation Accuracy	34%	77.5%
Loss	0.69	0.5

## 4.2 Analysis and Future Prospects

The visual speech model presented in this paper is a stack of 3D convolutional and bidirectional GRU layers, thereby boosting the results using the CTC loss function. The results produce a magnificent accuracy of 86.17% in predicting visual speech cues.

## 5. Conclusion

A review of the literature has been done where more focus has been placed on the importance of lip reading systems in enhancing crime investigation. This paper emphasises that the above systems can significantly contribute to improving the investigative process.

The combination of CNNs and GRUs has been vital to the creation of a vigorous deep learning model for lip reading. All these architectures play an independent part in enhancing the accuracy and efficiency of the given model. CNNs are especially useful for the lip image to extract the facial features which help the model to understand and learn the complicated visual patterns and details implicitly linked with speech. On the other hand, GRU networks perform well in capturing temporal dependencies that are crucial for lip reading since the model learns the temporal dependencies of lips movements.

This framework using the best of both CNNs and GRUs provides a good starting point to future development of lip reading. This integration not only improves the credibility of the model but also paves way for further advancements in the relation between human and computer interaction and especially in the field of accessibility. Therefore, this research creates a foundation for new developments that can enhance knowledge sharing and reduce misunderstanding in different fields.

The findings of this project prove that the CNN-GRU model can detect visual speech through lip reading with considerable accuracy across the various data sets. The high performance of this model can be considered as its applicability in real-world scenarios, especially in crime investigation to collect a large amount of evidence. It also has potential for the purpose of communication between the hearing impaired people and provides them with a tool to improve interaction in their environment. However, this paper has demonstrated some significant improvements in lip reading technology although the researcher wishes to acknowledge that there is so much more that has not been explored yet in this field.

To sum up, the relevance of deep learning in the field of lip reading – with CNNs and GRUs in particular – means enormous potential in the sphere of changing communication systems. Indeed, as we progress further in this area, future research and development will not only improve the efficiency of the current models but also increase the number of possible uses. Its focus will help shape a better world in which more people have equal opportunities due to elimination and or reduction in barriers with emphasis on communication, and the general use of technology that acts as a key to unearthing commonalities of the disabled to the rest of society.

## **Acknowledgements**

This research would not have been possible without the unwavering support and encouragement of numerous individuals. A deep appreciation is extended to Dr. Preeti Nagrath and Dr. Dharmender Saini, whose expertise and thoughtful guidance have been invaluable throughout this entire journey. Their insightful suggestions, constructive feedback, and steadfast support played a crucial role in shaping the direction and quality of this research, making it a truly enriching experience.

Gratitude is also owed to the colleagues at Bharati Vidyapeeth's College of Engineering, whose stimulating discussions and collaborative efforts greatly enriched the research process. Their diverse perspectives and shared knowledge fostered an environment of innovation and creativity that contributed significantly to the outcomes of this study. Special thanks are extended to [specific individuals, if applicable] for their technical assistance and generous advice, which were instrumental in navigating various challenges encountered during the research.

Finally, heartfelt thanks are offered to family and friends for their constant understanding and encouragement. Their unwavering support and belief in this endeavor provided invaluable motivation throughout this journey. This research is not just a culmination of academic effort

but also a reflection of the collective support and inspiration received from all those involved. Their contributions have made this achievement possible, and for that, deep appreciation is expressed.

## 6. References

- [1] J. Goldschen, O. N. Garcia, and E. D. Petajan, "Continuous automatic speech recognition by lipreading," in *Motion-Based recognition*, pp. 321–343, Springer, 1997.
- [2] C. Neti, G. Potamianos, J. Luetttin et al., "Audio visual speech recognition," Technical report, IDIAP, 2000.
- [3] G. Potamianos, C. Neti, G. Gravier et al., "Recent advances in the automatic recognition of audiovisual speech," *Proceedings of the IEEE*, vol. 91, no. 9, pp. 1306–1326, 2003.
- [4] H. Ninomiya, N. Kitaoka, S. Tamura et al., "Integration of deep bottleneck features for audio-visual speech recognition," in *International Speech Communication Association*, 2015.
- [5] J. S. Chung and A. Zisserman, "Lip reading in the wild," in *Asian Conference on Computer Vision*, 2016a.
- [6] A. Garg, J. Noyola, and S. Bagadia, "Lip reading using CNN and GRU," Technical report, Stanford University, CS231n project report, 2016.
- [7] Matthews, G., Zeidner, M., & Roberts, R. D. Boston "Review. Emotional intelligence: Science and myth", 2002
- [8] G. Zhao, M. Barnard, and M. Pietikäinen, "Lipreading With Local Spatiotemporal Descriptors," *IEEE TRANSACTIONS ON MULTIMEDIA*, 2009.
- [9] M. Gurban and J. -P. Thiran, "Information Theoretic Feature Extraction for Audio-Visual Speech Recognition," in *IEEE Transactions on Signal Processing*, Dec. 2009.
- [10] G. Papandreou, A. Katsamanis, V. Pitsikalis et al., "Adaptive multimodal fusion by uncertainty compensation with application to audiovisual speech recognition," in *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 3, pp. 423–435, 2009.
- [11] V. Pitsikalis, A. Katsamanis, G. Papandreou., "Adaptive multimodal fusion by uncertainty compensation," in *Interspeech*, 2006.
- [12] Z. Zhou, G. Zhao, X. Hong , "A review of recent advances in visual speech decoding," in *Image and vision computing*, vol. 32, no. 9, pp. 590–605, 2014.
- [13] P. Lucey, S. Lucey, and S. Sridharan, "Using a Free-Parts Representation for Visual Speech Recognition," in *Digital Image Computing: Techniques and Applications (DICTA'05)*, Queensland, Australia, 2005.
- [14] Goldschen, Alan & Garcia, Oscar & Petajan, Eric.. "Continuous Automatic Speech Recognition by Lipreading", 1997.
- [15] J. S. Chung and A. Zisserman, "Out of time: automated lip sync in the wild," in *Workshop on Multi-view Lip-reading, ACCV*, 2016b.
- [16] Z. Zhou, G. Zhao, X. Hong et al., "A review of recent advances in visual speech decoding," in *Image and Vision Computing*, vol. 32, no. 9, pp. 590–605, 2014.
- [17] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in *Proc. CVPR*, 2016.
- [18] J. Ngiam, A. Khosla, M. Kim et al., "Multimodal Deep Learning," in *Proceedings of the 28th International Conference on Machine Learning, ICML 2011*, pp. 689-696, 2011.
- [19] Y. Assael, B. Shillingford, S. Whiteson et al., "LipNet: Sentence-level Lipreading," 2016.
- [20] S. Petridis and M. Pantic, "Deep complementary bottleneck features for visual speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, 2016, pp. 2304-2308, doi: 10.1109/ICASSP.2016.7472088, 2016.
- [21] Y. Fu, S. Yan, and T. S. Huang, "Classification and feature extraction by simplexization," in *IEEE Transactions on Information Forensics and Security*, vol. 3, no. 1, pp. 91–100, 2008.

- [22] I. Almajai, S. Cox, R. Harvey et al., "Improved Speaker Independent Lipreading using Speaker Adaptive Training and Deep Neural Networks," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2016.
- [23] D. Hu, X. Li et al., "Temporal multimodal learning in audiovisual speech recognition," in IEEE Conference on Computer Vision and Pattern Recognition, pp. 3574–3582, 2016.
- [24] A. Takashima, I. Bakker, J. G. van Hell et al., "Interaction between episodic and semantic memory networks in the acquisition and consolidation of novel spoken words," in *Brain Lang.*, vol. 167, pp. 44-60, 2017.
- [25] S. Gergen, S. Zeiler, A. H. Abdelaziz et al., "Dynamic stream weighting for turbo-decoding-based audiovisual ASR," in *Interspeech*, pp. 2135–2139, 2016.
- [26] A. Haliassos, K. Vougioukas, S. Petridis, and M. Pantic, "Lips Don't Lie: A Generalisable and Robust Approach to Face Forgery Detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
- [27] Hinton, Geoffrey & Srivastava, Nitish & Krizhevsky, Alex & Sutskever, Ilya & Salakhutdinov, Ruslan. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint. arXiv*, 2019.
- [28] Bs, Prashanth & M V, Manoj & Puneetha, B. & Lohith, R & Gowda, V & Chandan, V & Sneha, H.. (2024). Lip Reading with 3D Convolutional and Bidirectional LSTM Networks on the GRID Corpus. 1-8. 10.1109/NMITCON62075.2024.10699241
- [29] Pourmousa, Hadi & Özen, Üstün. (2024). Lip reading using deep learning in Turkish language. *IAES International Journal of Artificial Intelligence (IJ-AI)*.
- [30] Cho, Sunyoung & Yoon, Soosung. (2024). Korean Lip-Reading: Data Construction and Sentence-Level Lip-Reading. *Journal of the Korea Institute of Military Science and Technology*.
- [31] Xiao, Yewei & Liu, Xuanming & Teng, Lianwei & Zhu, Aosu & Tian, Picheng & Huang, Jian. (2024). Cantonese sentence dataset for lip-reading. *IET Image Processing*.
- [32] Wang, Huijuan & Cui, Boyan & Yuan, Quanbo & Pu, Gangqiang & Liu, Xueli & Zhu, Jie. (2024). Mini-3DCvT: a lightweight lip-reading method based on 3D convolution visual transformer. *The Visual Computer*.
- [33] Miled, Malek & Messaoud, Mohammed & Bouzid, Aicha. (2022). Lip reading of words with lip segmentation and deep learning. *Multimedia Tools and Applications*.
- [34] Arakane, Taiki & Saitoh, Takeshi. (2023). Efficient DNN Model for Word Lip-Reading. *Algorithms*.