Enhanced Lip Reading Using Deep Model Feature Fusion: A Study on the MIRACL-VC1 Dataset

Susmitha Vekkot[1*], Doradla Kaushik[1†], Konduru Praveen Karthik[1†], Taduvai Satvik Gupta[1†]

[1]Department of Electronics and Communication Engineering, Amrita School of Engineering, Bengaluru, Amrita Vishwa Vidyapeetham, India.

*Corresponding author(s). E-mail(s): v_susmitha@blr.amrita.edu;
Contributing authors: kaushikdoradla@gmail.com; praveenkarthik2290@gmail.com; satviktaduvai@gmail.com;
†These authors contributed equally to this work.

Abstract

This paper presents a lip reading strategy by uniquely combining the feature learning capabilities of different deep models. Training models like Resnet, Inception-V3, VGG-16 and Mobilenet architectures are utilised as lip feature extractors. The extracted features are combined and utilised for word prediction on the MIRACL-VC1 dataset using various classifiers. The best-performing model used a combination of Resnet and random forest followed by dimensionality reduction using PCA. The best model's performance evaluation yielded results of 75% accuracy, 74% precision, 75% recall, and 74% F-score. It is found that this model is performing significantly better than the SOTA models, thereby demonstrating its potential for real-world applications in noisy environments, security, and human-computer interaction.

Keywords: Inception-v3, Mobilenet, PCA, Random Forest, Resnet, SVM, VGG-16

1 Introduction

Visual Speech Recognition (VSR) interprets dialogue using analysis of visual cues i.e. movement of the speaker's lips and tongue. It plays a vital role where audio is not available, as a means of communication for hearing impaired [1]. It can enhance human-computer interaction by enabling voice commands and interactions without

sound and is used in quiet environments like libraries or during meetings [2]. Above 5% population of the world requires rehabilitation to help with hearing loss with lip reading technology integrated into supportive devices [3]. A variety of human-centric technologies emerged in recent years using deep learning paradigms [4, 5], with focus on audio-visual speech recognition [6, 7]. Current research mainly focuses on using deep architectures like Convolutional Neural Networks (CNNs), its variants such as Alexnet, Inception-V3 [8], VGG-16 [9], Viterbi algorithm and combinations thereof (e.g., CNN-LSTM) to extract and interpret visual cues from lip movements. They have achieved varying levels of accuracy depending on the dataset and context.

This research focuses on the integration of learning models to develop a pipeline that can be used with different datasets to provide better lip reading performance. The usage of only DL models on the image-based MIRACL-VC1 dataset to develop a lip reading model results in lesser accuracy. This research combines the feature extraction capability of deep models and classify using shallow models. The major contributions of this research are:

- A versatile pipeline for lip reading combining feature extraction using various pre-trained models like Resnet, Inception-V3, VGG-16 and Mobilenet.
- Utilisation of features learnt from the above models for performing ML-based classification and word prediction.
- Performance analysis to determine the best model for lip reading using benchmark evaluation metrics.
- Comparison of results of this research with current SOTA models.

The research is set up as follows: Section 2 discusses recent research in lip reading. Section 3 details the lip reading architecture. Section 4 discusses important findings while Section 5 concludes the work with future insights.

2 Related Work

Similar pipelines, which extract spatiotemporal components surrounding the lips (based on motion, geometric characteristics, or both), have been employed in most of lip reading studies. Initially, authors in [10] used the LSTM-5 model for predicting visual ambiguities between words, but failed to generalize samples at different accents. Following this, [11] employed CAE as feature extractors following which this data was given to LSTM to predict a word. The system in [12] records and analyzes lip movements to generate various lip forms and patterns that correspond to different phonemes and words to facilitate effective communication for the voice impaired. A video's frame sequence is processed to determine the weights based on a lip's form in relation to time sequence in [13]. Researchers in [14] combines speaker recognition and feature fusion with supervised pre-training by using a shared visual encoder with CNNs and BGRUs. Using the AVLetters dataset, the suggested model had a classification accuracy of 70.77%. Authors in [15] created a mobile app that runs using CNN model but lacks accurate generalization leading to loss of scalability.

Subsequent studies focused on minimizing variations in the speaker's accent, lighting, image frame quality, speaking posture, and tempo. Researchers in [16] applied

batch normalization to obtain test accuracies of 52.9% and 38.5%. Model developed in [17] recognises words based on visual cues, using Inception-V3 model and achieved 64.6% accuracy. However, the model had constraints in terms of data dependency, computational power and speaker independence. Following this, [18] employed 3D DenseNet for greyscale input together with conv-3D and LSTM which improved the accuracy of the model. Recently, a non-autoregressive transformer-based speech synthesis model was proposed using an encoder and a GAN-based vocoder. In addition, the LipSound2 model [19] studies how to map facial picture sequences to large-scale spectrograms by pre-training an encoder-decoder architecture using cross-modal and a location-aware system. A lexicon-free system was developed in [2] that purely uses visual cues, achieving a 15% lower word error rate. Additionally, they were successful in getting an accuracy of 64.6% with a complex LRS2 dataset. Researchers in [20] followed a similar approach with multi-lingual dataset and experimented with English and Chinese languages by optimizing the produced audio for a speech recognition system that has already been trained.

Popular datasets in the literature are LRW, LRS2 and MIRACL-VC1 of which LRW and LRS2 are video-based datasets whereas MIRACL-VC1 is an image-based dataset [21]. SOTA models for lip reading are CNN variants and sequence models. Research in this paper optimizes the categorization of words in the MIRACL-VC1 dataset by extracting the lip features using deep extractors followed by word prediction using ML classifiers.

3 Proposed Methodology

The workflow of the suggested model, as seen in Fig. 1, is discussed in this section. The subsections provide various steps involved in the lip reading process.

3.1 Dataset

This research uses Multimedia Information Systems and Advanced Computing Laboratory (MIRACL-VC1) [21]. In this research, we use data from 15 speakers (10 female and 5 male). The dataset consists of 10 key frames for each instance, totalling 1500. A synchronous flow of 640x480 pixel color and depth images form each utterance. Each speaker utters the word 10 times and there are 10 words viz. Begin, Start, Stop, Choose, Connection, Next, Previous, Navigation, Hello and Web. The color images are used to locate the lip region and the depth images give the actual values of the images, aligned using sensor calibrator.

3.2 Data Pre-Processing

The steps involved in the preprocessing are detection of facial landmarks followed by lip region localisation and image concatenation, as explained below.

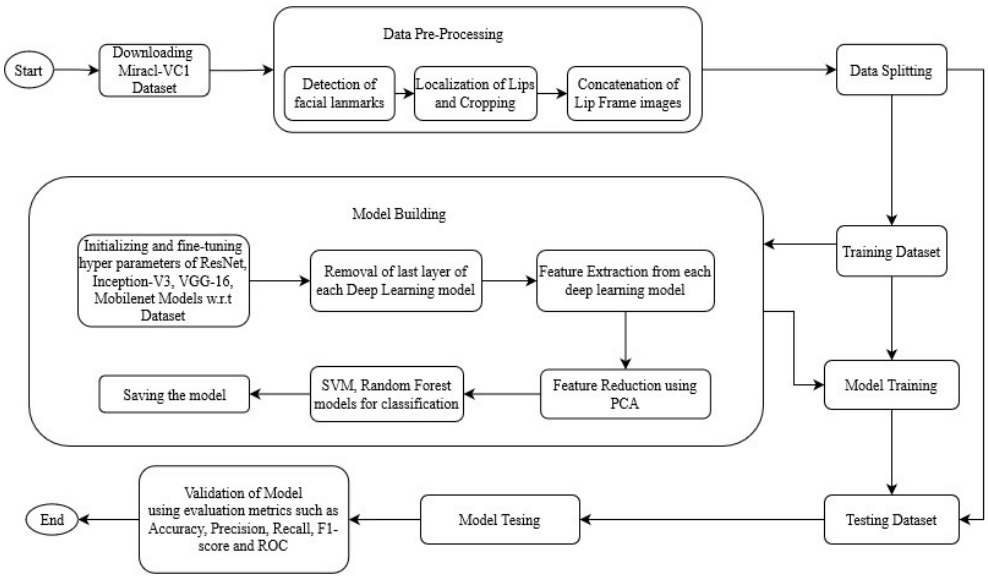


Fig. 1 Methodology of the Proposed Model

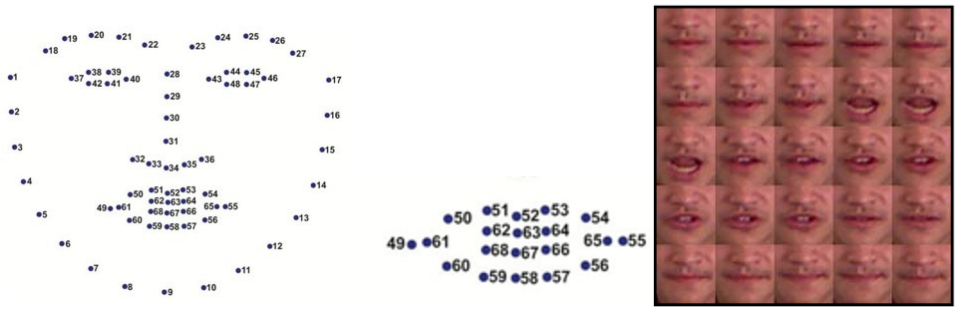
3.2.1 Detection of Facial Landmarks

Initially, we apply a facial landmark detection algorithm in Alg.1 to each image in the dataset and store the detected landmarks for further processing. The landmark detection is illustrated in Fig. 2(a).

Algorithm 1 Facial Landmark Detection

Require: Image I, face box B, model M
Ensure: Landmark positions L

```
Initialize L using mean shape S scaled to B
for t = 1 to T do
    Initialize feature vector F
    for each landmark l in L do
        Extract local features around l
        Append features to F
    end for
    for each tree R in ensemble E_t do
        Predict update ΔL_R using R and F
        Update landmarks: L ⇐ L + ΔL_R
    end for
end for
return L
```



(a) (b) (c)

Fig. 2 (a)Location of the landmarks detected: Jawline: Points 1-17, Eyebrow: Right: 18-22, Left: 23-27, Nose bridge: 28-31, Lower nose: 32-36, Eye: Right: 37-42, Left: Points 43-48, Lip: Outer: 49-60, Inner: 61-68 (b)Landmarks ROI (c)Image Concatenation

3.2.2 Localization of lips and Image Concatenation

From the landmarks detected previously, the landmarks of the lip region are our region of interest (ROI). ROI is cropped to make all the images into the same dimension ie. 40 X 40. Following this, all the frames are brought into a single image matrix using the batch normalization by Eqn. 1.

$$concatenated\_seq[i] = Frame[\frac{i * frame.no}{25}] \tag{1}$$

where $concatenated\_seq[i]$ represents the frame number in the new sequence; Frame[i] indicates the initial frame no. in the dataset; with 25 frames in the concatenated image. The cropped image landmarks and the image concatenation process are illustrated in Fig. 2(b) and Fig. 2(c).

3.3 Model Building & Validation

This research uses a unique architecture of combining learning models as in Fig. 1. A feature extractor is built using pre-trained models like Resnet, VGG-16, MoblieNet, and Inception-V3, with the architecture as described in Table 1. We eliminate the last layer of the deep-learning model and flatten the previous layer which in turn forms the features extracted from the images. Further, PCA-based dimensionality reduction is optionally performed with the parameters (PCA Components, N, Depth, C) calculated using a grid search algorithm as in Alg. 2. The PCA-reduced features are visualised in Fig. 3. Features are then given to models like SVM (Support Vector Machine) and Random Forest. The data is split in 80-20 train-test split format. 8 women and 4 men speakers are used for training while 2 women and 1 men speakers are used for testing. This model is implemented on a PC with an Intel i5 processor and an Nvidia GTX 1650 graphics card. The following gives a brief description of feature extractors:

- Resnet : A deep neural network architecture to address the degradation problem using residual learning and skip connections [14].

- Inception-V3 : A variant of CNN to optimize performance, incorporated with batch normalization, regularisation, efficient grid size reduction etc. [8]

- VGG-16 : A variant of CNN that is a popular choice for image recognition tasks. [9]

- MobileNet : CNN with depthwise separable convolutions with parameter reduction and computing complexity. [22]

Table 1 Architecture of the deep-learning models

Parameter	Resnet	Inception-V3	VGG-16	Mobilenet
Optimizer	Adagrad	Adagrad	Adagrad	Adagrad
Loss Function	Log loss	Log loss	Log loss	Log loss
Input Size	100x100	100x100	100x100	100x100
Initial Layer	Conv2D (64 filters, 7x7)	Conv2D (64 filters, 7x7)	Conv2D (64 filters, 3x3)	Conv2D (32 filters, 3x3)
Activation Function	ReLU	ReLU	ReLU	ReLU
Blocks	4 residual	2 Inception	5 Convolution	3 MobileNet
Fully Connected Layers	4096, 2048, 1024, 512	1024, 512	4096 units, Dropout (0.5)	1024, 512, 256 units
Output Layer	Dense (10 units, Softmax)	Dense (10 units, Softmax)	Dense (10 units, Softmax)	Dense (10 units, Softmax)

Algorithm 2 Grid Search for Hyperparameter Tuning

```
function GRID_SEARCH(Model, Hyperparameters, Grid)
    bestscore ← −∞                          ▷ use ∞ for minimization problems
    bestparams ← None
    for params in Grid do
        Model.set_params(params)
        scores ← cross_validation(Model)
        score ← average(scores)
        if score > bestscore then
            bestscore ← score
            bestparams ← params
        end if
    end for
    return bestparams
end function
```

Feature dimension of (1200,8192) is reduced to (100,8192) after applying PCA. The model hyperparameters are listed in Table 2.

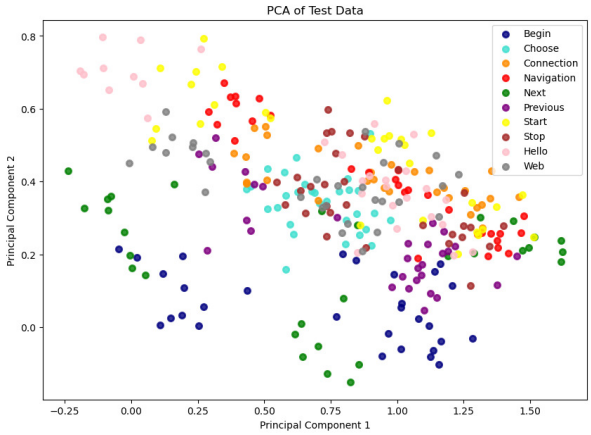


Fig. 3 PCA reduced features visualisation

Table 2 Hyperparameters of Models

Model No	Deep Learning Model	Machine Learning Model	Specifications	PCA([1])
Model-1	ResNet	SVM	Linear Kernel	No
Model-2	ResNet	RandomForest	N=1000, Depth=90	No
Model-3	ResNet	SVM	Linear Kernel	Yes(100)
Model-4	ResNet	RandomForest	N=200, Depth=5	Yes(50)
Model-5	Inception-V3	RandomForest	N=400,Depth=30	No
Model-6	Inception-V3	RandomForest	N=400, Depth=30	Yes(100)
Model-7	Inception-V3	SVM	Linear Kernel	No
Model-8	VGG-16	SVM	Linear Kernel	No
Model-9	VGG-16	RandomForest	N=500,Depth=50	No
Model-10	MobileNet	SVM	Linear Kernel	No
Model-11	MobileNet	RandomForest	N=2000,Depth=30	No

[1]No.of Components for PCA

The model's prediction is validated based on the following performance measures as given by Eqns. 2, 3, 4 and 5.

$$Accuracy = \frac{No.\ of\ correct\ prediction}{Total\ no.\ of\ predictions} \tag{2}$$

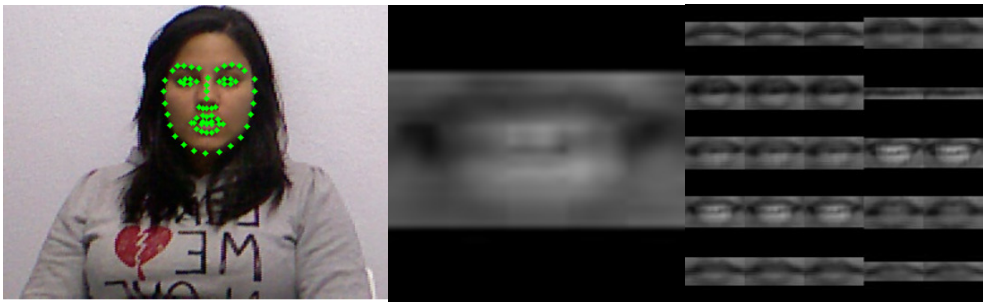
$$Precision = \frac{True\ Positives}{True\ Positive + False\ Positives} \tag{3}$$

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives} \tag{4}$$

$$F1 - Score = \frac{2 * Precision * Recall}{Precision\ + Recall} \tag{5}$$

4 Results and Discussion

Benchmark metrics for assessment are employed to judge the proposed lip reading strategy and the results are compared with SOTA. The results of the pre-processing operations performed on the input images are illustrated in Fig. 4. Fig. 4(a) shows the landmarks extracted from the face in the dataset. Fig. 4(b) provides the ROI extraction and converting it into 40X40 dimension. Fig 4(c) gives the concatenated image.



(a) (b) (c)

Fig. 4 Pre-processing results : (a)Detection of facial landmarks (b)Separation of lip ROI (c)Image concatenation using batch normalization.

The models in Table. 2 were trained using test data, and the evaluation was performed based on the measures described in Eqns. 2-5 in Fig. 5, alongwith ROC and confusion matrix as described in Fig. 7. The training scores of all the models are almost 100%. The test data is used for Fig. 5. As shown in Fig. 5, it is evident that Resnet pipelined with random forest followed by PCA gives the highest testing accuracy of 75%, precision of 74%, recall of 75% and F1-Score of 74%. The test results for 4 different inputs provided to Model 4 are illustrated in Fig. 6.

As observed in Fig. 7(a) the AUC-ROC for the model-4 is greater than 0.75 for every class. Fig. 7(b) depicts the confusion matrix from which we can interpret that only for class 4, Model-4 performs average while for all other classes, it is performing well. Random forest particularly is useful as it prevents overfitting where noise and variations in lips are present. The features extracted can be handled by RF as they are high-dimensional data. The dimensionality reduction using PCA further enhances the result leading to better performance.

Finally, the suggested methodology is compared with SOTA models, as listed in Table. 3. It is observed that the performance of the best model has increased by almost 5% compared to the AV Letters dataset which is larger in size compared to the MIRACL-VC1. Compared to the studies involving MIRACL-VC1 dataset, the proposed model performance has increased by almost 12%.

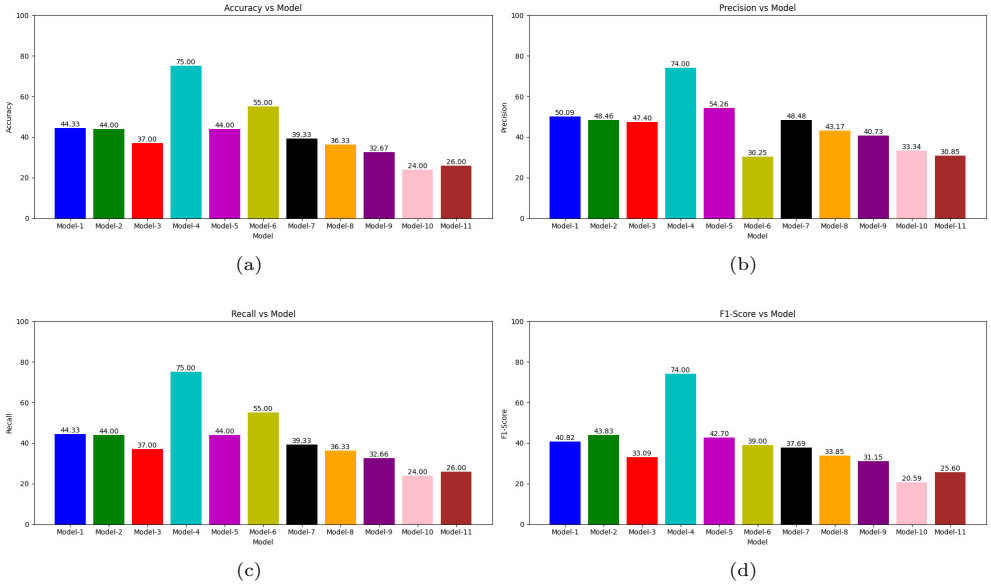


Fig. 5 Performance of the models (a)Accuracy (b)Precision (c)Recall (d)F1-Score

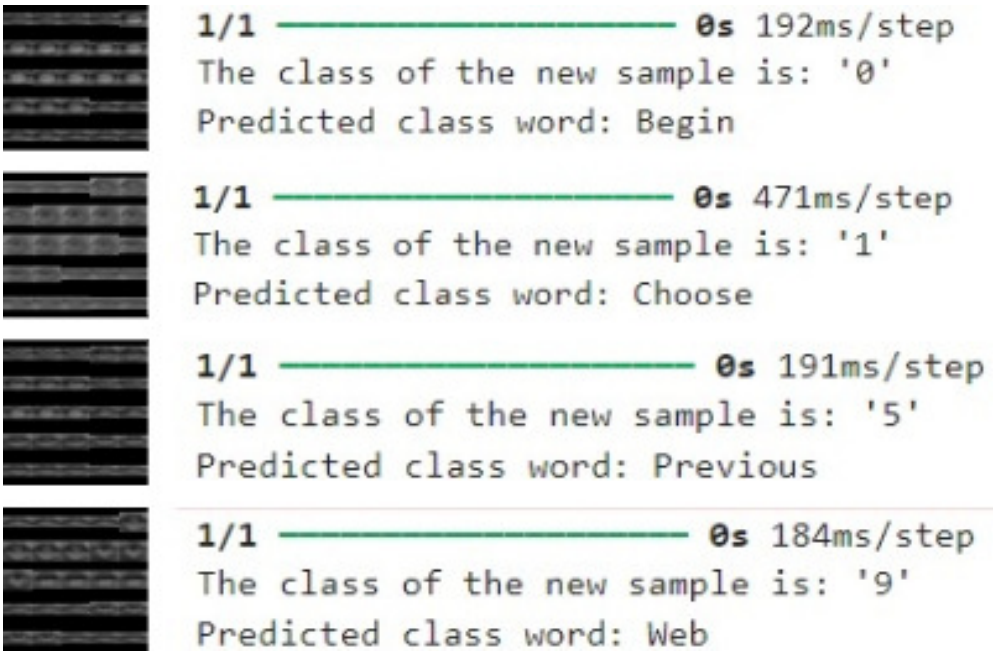
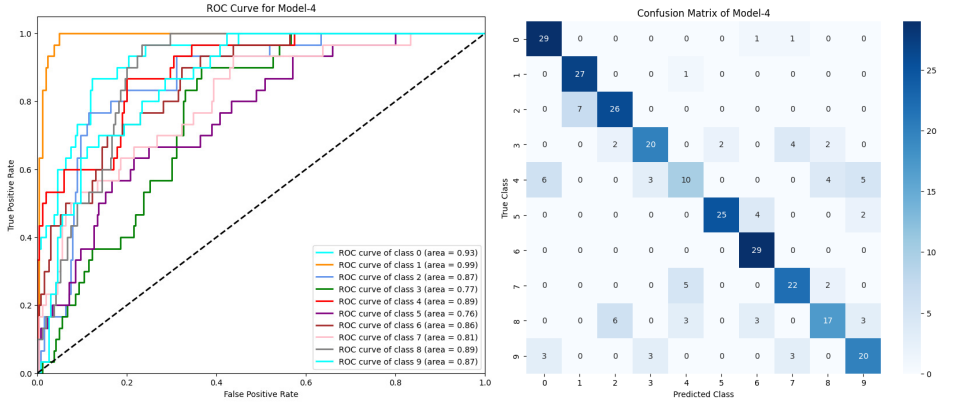
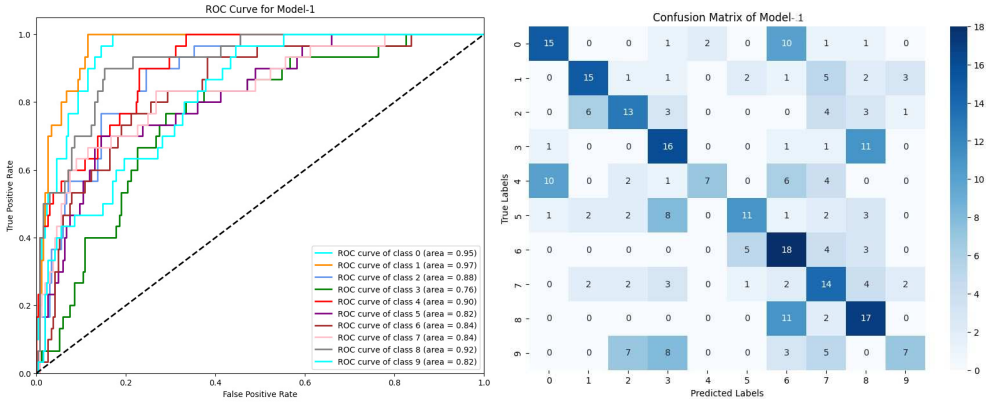


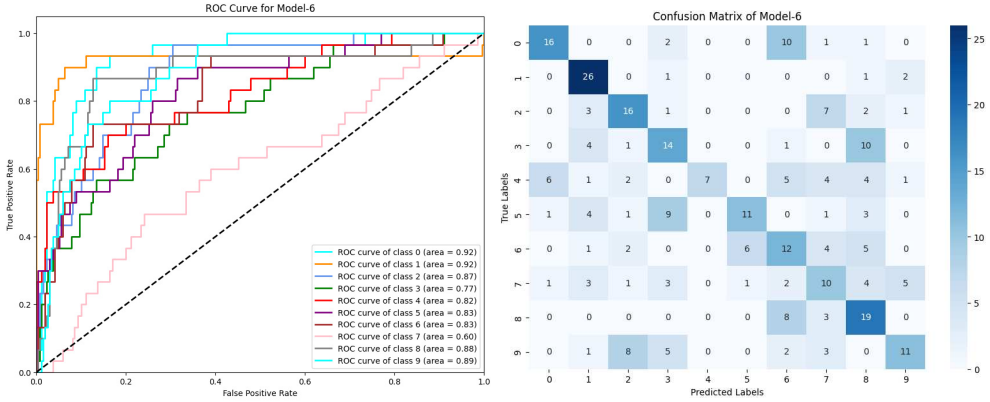
Fig. 6 Prediction of the word for new sample (a) Begin (b) Web (c) Previous (d) Choose



(a) (b)



(c) (d)



(e) (f)

Fig. 7 Performance of Models:(a)ROC Curves of the Model-4 (b) Confusion Matrix of Model-4 (c)ROC Curves of the Model-1 (d) Confusion Matrix of Model-1 (e)ROC Curves of the Model-6 (f) Confusion Matrix of Model-6

Table 3 Comparative study with state of the art

Author(Year)	Model Used	Accuracy	Dataset	Size of Data (Type)
Hashmi et.al(2018)[17]	CNN	56	MIRACL-VC1	1500 Instances(Image Based)
Sindhura et.al(2018)[16]	Inception-V3	37.1	MIRACL-VC1	1500 Instances(Image Based)
Abrar et.al(2019)[23]	CNN	60	MIRACL-VC1	1500 Instances(Images Based)
Nandini et.al(2019)[13]	Deep Weighted Feature Representation	68.46	Kannada dataset	Not known
Parekh et.al(2019)[11]	CAE+LSTM	63.22	MIRACL-VC1	1500 Instances(Image Based)
Bi et.al(2019)[18]	DenseNet	37.92	LRW-1000	500000 Instances(Video Based)
Fenghour et.al(2020)[24]	CNN	65.5	LRS2	145000 Instances(Video Based)
Muhamad et.al(2022)[25]	CNN+ResNet-18	44.6	LRW-1000	500000 Instances(Video Based)
Wu et.al(2023)[14]	CNN+ResNet18	70.77	AVLetters	4700 hours(Video based)
Qu et.al(2024)[19]	Multitask CNN	43.53	TCD-TIMIT	6300 Instances(Video based)
2024	**Proposed Model-4**	**75**	**MIRACL-VC1**	**1500 Instances(Image Based)**

5 Conclusion and Future scope

A lip-reading system has been developed and tested for English words based on the MIRACL-VC1 dataset. The proposed model involves detecting facial landmarks, from which the lips would be localized, cropped and concatenated for efficient processing. Different deep feature extracters, such as ResNet, VGG-16, MobileNet, and Inception-V3, were tested in combination with machine learning classifiers such as SVM, and Random Forest for optimized word classification and provided an accuracy of 75%. Integration of these models significantly enhances lip-reading accuracy. This approach effectively generalises across multiple speakers, which is critical for practical applications.

In future, the results can be extended to bigger and more diverse datasets which can aid in enhancing model generalisation capabilities. Finetuning the models with a versatile dataset can reduce the need for excessive training of the dataset. Text classification models can be further integrated into the architecture for transcribing the recognised speech as text and converting the text into speech.

Compliance with Ethical Standards

Conflict of Interest

On behalf of all authors, the corresponding author states that there is no conflict of interest.

Informed Consent

Not applicable.

Ethical approval

This article does not contain any studies with human participants or animals.

References

[1] Li, Y., Xue, F., Wu, L., Xie, Y., & Li, S. Generalizing sentence-level lipreading to unseen speakers: a two-stream end-to-end approach. Multimedia Systems, 30(1), 42, (2024).

[2] Hao, M., Mamut, M., & Ubul, K. "A survey of lipreading methods based on deep learning", In Proc. Int. Conf. Image Proc. Machine Vision (pp. 31-39), (2020).

[3] World Health Organization (WHO 2021). Deafness and hearing loss. Retrieved : https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss.

[4] Chen, J., Vekkot, S., & Shukla, P. Music Source Separation Based on a Lightweight Deep Learning Framework (DTTNET: DUAL-PATH TFC-TDF UNET). In Proc. ICASSP (pp. 656-660), (2024).

[5] Senthil, B., Sundaram, V., Chauhan, A. S., & Vekkot, S. Empowering Facial Analytics: A Unified Approach for Emotion, Age, Gender and Object Identification. In Proc. SPIN, (pp. 145-150), (2024).

[6] Kumar, L. A., Renuka, D. K., & Priya, M. S. Towards robust speech recognition model using Deep Learning. InProc. ICISCoIS, (pp. 253-256), (2023).

[7] Kumar, L. A., Renuka, D. K., Rose, S. L., & Shunmugapriya, M. C. Attention-based multi-modal learning for audio visual speech recognition. Proc. AIST, (pp. 1-4), (2022).

[8] Szegedy, C., Vanhoucke, V, Ioffe, S, Shlens, J & Wojna, Z. "Rethinking the Inception Architecture for Computer Vision," in Proc. IEEE Conf. Computer Vision and Pattern Recognition, (2016).

[9] Simonyan, K. & Zisserman, A. "Very deep convolutional networks for large-scale image recognition," in Proc. ICLR, (2015).

[10] Joon Son C., Andrew Z. "Learning to lip read words by watching videos", Computer Vision and Image Understanding, Volume 173, Pages 76-85, (2018).

[11] Parekh, D., Gupta, A., Chhatpar, S., Yash, A. & Kulkarni, M. "Lip Reading Using Convolutional Auto Encoders as Feature Extractor," Proc. I2CT, pp. 1-6, (2019)

[12] K. Neeraja, K. Srinivas R., & Praneeth, G. "Deep Learning based Lip Movement Technique for Mute," Proc. ICCES, Coimbatore, India, (2021).

[13] Nandini, M. S., Nagavi, T. C. & Bhajantri, N. U. "Deep Weighted Feature Descriptors for Lip Reading of Kannada Language," Proc. SPIN, (2019).

[14] Wu, J., Zhang, Y., Zhang, X., Zheng, C., Yan, Y., & Yin, E. "Improving Visual Speech Recognition for Small-Scale Datasets via Speaker Embedding", In Proc. CSECS (pp. 01-06), (2023)

[15] Afouras, T., Chung, J. S., Senior, A., Vinyals, O.. Zisserman, A. "Deep Audio-Visual Speech Recognition",(2018) arXiv:1809.02108

[16] Sindhura, P, Preethi, S. J. & Niranjana, K. B. "Convolutional Neural Networks for Predicting Words: A Lip Reading System," Proc. ICEECCOT, Msyuru, India, (2018).

[17] Nadeem Hashmi, S., Nanda, A. & Gupta, S. "A Lip Reading Model Using CNN with Batch Normalization," Proc. IC3, Noida, India, (2018)

[18] Bi, C., Zhang, D., Yang, L., & Chen, P. "A Lipreading Modle with DenseNet and E3D-LSTM," Proc. ICSAI, Shanghai, China. (2019)

[19] Qu, L., Weber, C. Wermter, S. Lipsound2: Self-supervised pre-training for lip-to-speech reconstruction and lip reading. IEEE Trans. Neural Networks and learning systems, 35(2), 2772-2782, (2024)

[20] Chung, J. S., & Zisserman, A. Lip reading in the wild. In Proc. Computer Vision–ACCV 2016, Taiwan, (pp. 87-103), (2016).

[21] Ahmed R, Achraf B., & Walid M. A new visual speech recognition approach for RGB-D cameras. In Proc. ICIAR2014, Vilamoura, Portugal, October 22-24 (2014).

[22] Bhargav, G. P., Reddy, K. S., Viswanath, A., Teja, B., & Byju, A. P. An Integrated Facemask Detection with Face Recognition and Alert System Using MobileNetV2, In Proc. ICICC (pp. 77-87), (2022).

[23] Abrar, M. A., Islam, A. N., Hassan, M. M., Islam, M. T., Shahnaz, C., & Fattah, S. A. Deep lip reading-a deep learning based lip-reading software for the hearing impaired. In Proc. R10-HTC (47129) (pp. 40-44). (2019).

[24] Fenghour, S., Chen, D., Guo, K., & Xiao, P. Lip reading sentences using deep learning with only visual cues, IEEE Access, 8, 215516-215530, (2020).

[25] Haq, M.A., Ruan, S.J., Cai, W.J., & Li, L.P.H. Using lip-reading recognition to predict daily Mandarin conversation. IEEE Access, 10, 53481-53489., pp. 53481-53489, (2022).