Analysis of Lip Reading of Assamese Digits using Deep Learning

Rabinder Kumar Prasad[1], Dhiraj Kalita[2*], Zakariya Momin Mondal[3], M. Tiken Singh[4], S Md S Askari[5], and Chandan Kalita[6]
[1,2,3,4]Department of Computer Science and Engineering, DUIET,
Dibrugarh University, Dibrugarh 786 004, Assam, India
[5] Department of Computer Science and Engineering, Rajiv
Gandhi University, Doimukh-791112, AP, India
[6]Department of Information Technology, Gauhati
University, Gauhati 781014, Assam, India
E-Mail: rkp@dibru.ac.in, rs_dhirajkalita@dibru.ac.in*, zakariya14112001@gmail.com,
tiken.m@dibru.ac.in, sikdar.askari@rgu.ac.in, kalitachandan@gauhati.ac.in,
*Corresponding Author

Abstract—Effective communication in noisy environments, such as aviation, construction, and manufacturing, is often hindered due to auditory challenges, making oral communication difficult. To address this issue, we propose an automatic lip-reading system specifically designed for recognizing Assamese digits in high-noise settings. This study introduces a deep learning-based approach that extracts the geometric features of lip movements from video data to accurately predict spoken digits. Traditional lip-reading models struggle with language-specific nuances due to reliance on generic datasets. To overcome this limitation, we construct a custom dataset of video recordings featuring diverse speakers varying in age, gender, and accent, ensuring a more robust and adaptable model. We employ a CNN+LSTM architecture, where Convolutional Neural Networks (CNNs) capture spatial features, and Long Short-Term Memory (LSTM) networks learn temporal dependencies. Experimental results demonstrate that our CNN+LSTM model outperforms conventional architectures like RNN+LSTM and RNN+CNN, achieving an accuracy of 83%. The findings highlight the effectiveness of deep learning in enhancing accessibility for the deaf and hard-of-hearing and enabling voice-free human-computer interaction.

Keywords: Long-Short Term Memory (LSTM), Lip Region Recognition, Color Imaging, Deep Learning. Lip reading, Assamese language, Custom dataset, Digit recognition.

INTRODUCTION

The importance of lip reading has been increasingly acknowledged, particularly with communication accessibility and technology advancements. Lip reading, the capacity to comprehend speech by analysing the motions of the lips, face, and tongue, has shown to be very beneficial for individuals who are deaf or have hearing difficulties. As technology advances, the potential to enhance lip reading is also evolving via creative methods such as artificial intelligence and augmented reality. In [1] shown that visual cues, such as lip movements, significantly enhance speech comprehension in noisy settings. Their study highlighted the need of integrating visual and auditory input to improve communication, influencing later studies on multimodal speech comprehension. Moreover, [2]

https://doi.org/10.22232/stj.2025.13.01.18

noted the challenges in distinguishing consonant sounds alone by concentrating on lip motions during visual perception. The study provided significant insights into the limitations of identifying analogous sounds in visual speech perception, contributing to a broader understanding of challenges in visual communication. In [3], investigated the significance of beginning with simpler tasks when training neural networks, emphasising that a progressive escalation in complexity results in improved learning outcomes. The study demonstrated that neural networks exhibit superior performance when initiated with simpler inputs before progressing to more complex patterns, hence enhancing comprehension of cognitive development and the formulation of artificial intelligence systems. [4] were the pioneers in examining the influence of context on the orthography of English vowels, discovering that adjacent letters significantly affected vowel spelling.

Subsequently, [5] demonstrated the enhancement of speech recognition through audio-visual information by comparing three hidden Markov model methodologies: two top-down approaches employing lip contour fitting and PCA, and a bottom-up method grounded in nonlinear scale-space analysis of pixel intensities. [6] suggested a robust model for mouth forms in sign language identification, which is particularly pertinent for the era of lip reading, using deep convolutional neural networks. Mouth forms in sign language are challenging to annotate, and there is a scarcity of publicly accessible annotations. This work uses interconnected information sources to provide inadequate supervision. Subsequently, several writers have conducted study on oral morphology and lip movement [7]. In [8], suggested an innovative deep learning architecture for voice enhancement using lip-reading. It encompasses speech improvement derived from a lip-reading model. The advancement of deep learning and multimodal methodologies has resulted in significant progress in the interconnected domains of lip reading and voice improvement.

Lip reading is essential for tasks such as speech recognition, biometric authentication, and assisting those with hearing impairments, as it involves understanding speech through visual lip movements. Lip reading can be difficult for humans, especially without context, but it helps bridge communication gaps in loud environments and improves comprehension [9]. Recent developments in deep learning have changed the way lip reading is done. Initial attempts, such as the work of [10], brought in deep models, whereas [11] created a complete architecture merging CNNs and LSTMs, boosting accuracy by 6.8%. In this field the progress has been advanced by the presence of extensive datasets like LRW, LRW-1000, LRS2, and LRS3. Nevertheless, the majority of research is centered on English datasets, underscoring the necessity for studies in local languages [12].

The contribution of this research are given below:

(i) A dataset for lip-reading for Assamese digits were generated.

(ii) Pre-process the dataset with different augmentation techniques.

(iii) Build a Deep Neural Network Architecture by comparing different architectures through which the model can predict the digits using the dataset with best accuracy.

The remaining paper consists of the following sections. Literature review on lip-reading identification models are discussed in Section 2. The proposed works and the methodology used are presented in Section 3, the results are discussed in Section 4, Finally, the findings of this work are summarized in Section 5 along with future directions of the work.

LITERATURE REVIEW

In this section, we define the existing work that has been accomplished in the subject modern-day lip analyzing. maximum tactics contain system ultra-modern techniques. After the emergence ultra-modern deep cutting-edge methods, it produced state-of-the-art consequences. Recent progress in lip reading utilizes different neural network structures to improve precision, durability, and practicality in real-life situations.

The recognition of vowels using KNN algorithms was introduced by [13], which transformed the video of the subject articulating vowels into pictures, thereafter selecting images manually for further processing. Nevertheless, other problems such as rapid speech, inadequate pronunciation, insufficient lighting, facial movement, and the presence of beards-moustaches complicate reading of lip. Contour tracking methods and template matching are employed to identify the lips on a face. The K Nearest Neighbor algorithm is utilized to classify images as either “speaking” or “quiet.” The sequence of images is subsequently transformed into segments of speech. A feature vector is calculated for each frame across all segments and stored in the database along with correctly labeled classes. Character recognition is executed via a modified KNN algorithm that allocates more weight to proximate neighbors in [14], that describes techniques for predicting words and sentences from video without audio. We study alternative strategies to use a

VGGNet pretrained on celebrity faces from IMDB and Google Images to handle these picture sequences. The VGGNet is trained on pictures concatenated from numerous frames in each sequence and utilised with LSTMs to extract temporal information. LSTM models fail to outperform other approaches for various reasons, however the concatenated image model with nearest-neighbor interpolation achieved 76% validation accuracy. A lipreading model featuring a well-structured integration and arrangement of processing components to capture highly unique visual features has been detailed in [15]. This highlights the use of a well-organized Deep Belief Network (DBN)-based recognition system. Multi-speaker (MS) and speaker-independent (SI) tasks were performed utilizing the CUAVE database, resulting in phone recognition rates (PRRs) of 77.65% and 73.40%, respectedly. The best word recognition rates (WRRs) achieved in the MS and SI tasks are 80.25% and 76.91%, respectively. The results demonstrate that the proposed method outperforms conventional Hidden Markov Models (HMM) and holds its own against top visual speech recognition methods.

In [16], two distinct deep-learning models for lip-reading were developed: the first model utilised a spatiotemporal convolutional neural network, a bi-gated recurrent neural network, and Connectionist Temporal Classification Loss for video sequences; the second model processed audio by inputting MFCC features into a layer of LSTM cells to produce the output sequence. They furthermore produced a little audio-visual dataset to train and evaluate our algorithm. Their objective was to amalgamate both models to enhance speech recognition in a noisy setting.

An algorithm utilizing a lip deep learning model was proposed in [17] to enhance the accuracy of lip-reading recognition. A binary representation of the lip contour motion sequence was displayed on the spatio-temporal energy, while a dynamic gray-scale of the lips was utilized to reduce noise interference during the recognition phase, significantly enhancing the results of lip-reading recognition thanks to the advanced abilities of deep learning. The results of the experiments suggest that deep learning can extract beneficial features from the dynamic changes in lip gray-scale, leading to better identification results.

A deep learning voice augmentation system based on lip reading was shown [8]. In contrast to benchmarks, the filtering-based approach employs deep learning and analytical audio modelling. Audio-visual (AV) speech augmentation is provided at two levels. Level one uses an innovative deep learning regression model for lip reading. An improved, visually based Wiener filter predicts the second-level clean audio power spectrum for lip-reading. A stacked long-short-term memory (LSTM) regression model for lip reading generates distinct sounds by using temporal visual data from many frames. Utilising inferred speech characteristics, the distinctive filter bank-domain EVWF forecasts intelligible speech spectrums. We evaluate the EVWF against spectral subtraction and log-minimum mean-square error using both ideal and LSTM-driven AV mapping. The suggested AV speech augmentation system is assessed in four dynamic real-world contexts (café, street intersection, public transportation, and pedestrian zone) over a range of low to high signal-to-noise ratios using benchmark grid and ChiME3 datasets. Assessment of restored speech using perceptual quality. Subjective testing utilizes the standard mean opinion score along with inferential statistics. In such simulations, lip-reading and voice enhancement improve speech quality and comprehensibility. A convolutional neural network model is presented in [18] for predicting words based on movies without utilizing any audio. This model is built upon the VGG Net architecture, which has been pretrained on the ImageNet Dataset, with modifications made for the MIRACL-VCl Dataset, which includes 10 words. The model reached 94.86% accuracy in training, 93.82% accuracy in validation, and 60% accuracy in testing. A mobile application has been developed based on this framework, allowing for real-time performance through cloud computing on any smartphone, aiding individuals with hearing impairments in their everyday activities and promoting more natural and spontaneous conversations in a budget-friendly way.

The study [19] presents a way to include facial expression characteristics, namely expression-based and action unit-based data, into the lip-reading technique. Evaluation tests are performed using three public databases: OuluVS, CUAVE, and CENSREC-1-AV. The integration of face expression features has been demonstrated to enhance recognition accuracy across all databases.

In [12], a lip-reading system using neural networks has been introduced to forecast phrases from silent video recordings of individuals conversing. The system is devoid of a language, depending only on visual indicators shown via a restricted set of visemes (unique lip motions). It is designed to accommodate an extensive vocabulary, including hitherto unencountered terms, and exhibits resilience to fluctuating illumination conditions. The technique, validated on the BBC LRS2 dataset, demonstrates markedly enhanced word categorization accuracy relative to leading methodologies.

Subsequently, in [20], the emphasis was placed on the construction of the collection, processing, and data recognition network structure for lip reading. They created a precise and resilient system for lip reading in the study. Initially, they isolated the oral region and segmented it utilising a novel hybrid model incorporating a newly introduced edge and filter. Subsequently, they trained their spatiotemporal model by integrating Convolutional Neural Networks (i.e. CNN) with Bi-directional Gated Recurrent Units (i.e. Bi-GRU). Ultimately, they assessed their algorithm and achieved an accuracy assessment of 90.38%. The outcome demonstrates the efficacy of our system via the use of lip segmentation as inputs to the suggested spatio-temporal model. Our succinct study indicates the want for a suitable model proficient in lip reading Assamese digits. This study introduces a method for identifying lip motions corresponding to Assamese numerals. This section will concisely delineate our intended work and technique.

PROPOSED WORK AND METHODOLOGY

Identifying the Digits uttered by speakers in the Assamese Language. We considered three Architectures and compared their accuracy and training times: (i) A recurrent model using Long Short-Term Memory (i.e. LSTM) (ii) A recurrent model using CNN, and (iii) proposed model i.e. combining CNN and LSTM model. The flow diagram is described in Fig. 1

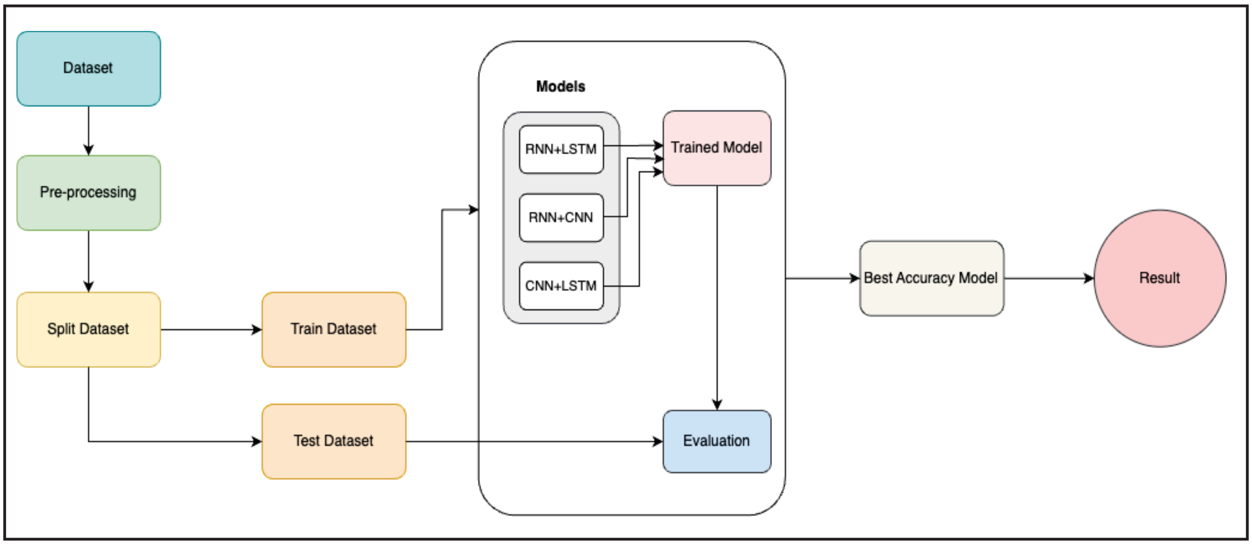

Figure 1: Flow diagram of the Proposed Model

Data Collection

Deep learning models are trained with a self-compiled dataset because appropriate online datasets or videos are not available. The speakers' excessive movement in online videos made it difficult to detect their lips. As a result, a fresh dataset was formed by 30 volunteers, with speakers saying Assamese numbers from 0 to 10. The dataset contains individuals of different age groups, genders, and accents, ensuring appropriate lighting and stable camera positioning for optimal lip clarity. Methods such as OpenCV and dlib were utilized for extracting and standardizing the lip area in videos, while the dataset comprises 10 categories, with each one corresponding to a numeral. The detailed information on the specifications of the dataset and the number of samples for each digit are reported in Table 1 and in Table 2. The size of the dataset is specified in Table 3.

Table 1: Assamese digits, their transliteration and the number of samples

Digit	Transliteration	No. of Samples
0	xuinno	41
1	ek	80
2	dui	80
3	tini	80
4	sari	80
5	pas	80
6	soy	80
7	xat	80
8	ath	80
9	no	80
10	doh	80

Table 2: Specification of Dataset

Aspect	Specification
Pixels	1920 × 1080
Contents	Digits from '0' to '10'
Number of Speakers	30
FPS (Frames Per Second)	15
Frames Per Video	30
Categories	Facial Expression and Accent
No. of Repetitions	3 times for each digit

Table 3: Size of Dataset

Data	No. of Images
Before Augmentation	2230
After Augmentation	4500
After Data Gain	6200

Data Preprocessing

The acquired dataset is preprocessed to make it suitable for input. It consists of five steps: (a) Video to Frame Conversion; (b) Mouth Localization; (c) Data Cleaning; (d) Image Augmentation; and (e) Concatenating the frames. Using cv2, the video is broken down into frames at a rate of 15 frames per second with a resolution of 1920x1080. The face and lip regions of each frame are identified using dlib's facial landmark detector during processing. The process begins by changing the frame to grayscale, with dlib then detecting 68 facial landmarks. Points 49-68 are related to the area around the mouth, from which the measurements of the lip's height and width are taken, shown in Fig. 2. These measurements serve as attributes for teaching the lip-reading algorithm. The frame-wise information of each digit are shown in Fig. 3.

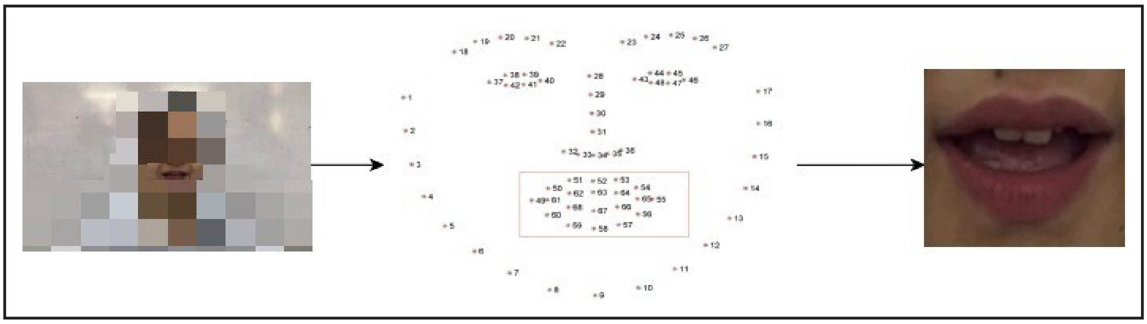

Fig. 2: Lip Detection

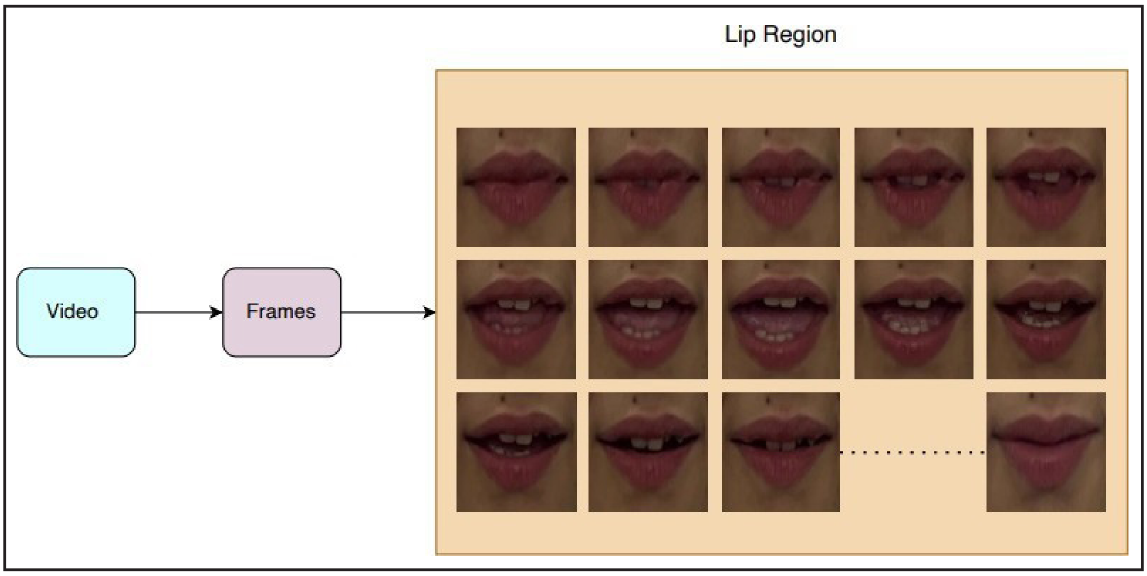

Fig. 3: Detected Lips

Data cleaning is necessary in machine learning in order to eliminate inconsistencies, inaccuracies, and errors that may impact the performance of the model. This involves activities such as standardization, rescaling, identification of anomalies, and filling in missing values. Real-life data can be filled with noise and gaps, and the process of cleaning it guarantees accuracy and dependability. Image augmentation creates synthetic training images using methods such as rotating, flipping, resizing, and adding noise. This increases the volume and variety of the dataset, helping to mitigate overfitting and improve the model's generalization. The enhanced lips are combined into one 224x224 image, in the order of frames. This picture serves as the model’s input, containing utterances for every digit that are grouped together and divided into training, validation, and testing groups. The augmentation and concatenated frame representations are shown in Fig. 4, Fig. 5, and Fig. 6.

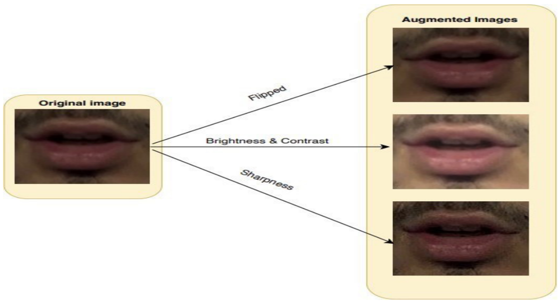

Fig. 4: Augmentation

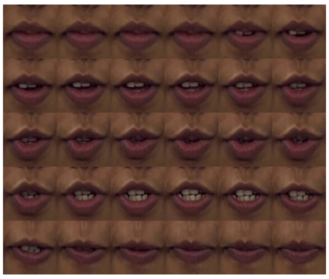

Fig. 5: Concatenated Frame 1

Fig. 6: Concatenated Frame 2

NEURAL NETWORK MODELS

Neural network models are artificial systems that mimic the structure of the human brain, created to identify patterns and forecast outcomes based on data.

A. RNN

A type of neural network, the Recurrent Neural Network (RNN), is specifically created for handling sequential data like time series, natural language, and speech recognition. RNNs differ from traditional neural networks in that they utilize the output of the previous step as input for the current step, enabling them to retain historical data. This is crucial for activities such as anticipating the following word in a sentence, which requires context from earlier words. RNNs’ main characteristic is the Hidden State or Memory State, which stores details of the sequence.

Types of RNN Structures Comprise:

- Recurrent Neural Networks (RNN) that go in both directions, known as Bidirectional Recurrent Neural Networks (BRNN)
- GRU stands for Gated Recurrent Units.
- Long Short Term Memory (LSTM) is a type of neural network architecture.

B. LSTM

Long-Short Term Memory (i.e. LSTM), created by [21], is a recurrent neural network (rnn) that aims to capture long-term relationships in sequence prediction tasks such as machine translation, time series forecasting, and speech recognition [22]. In contrast to traditional RNNs, LSTMs utilize a memory cell containing three gates—input, forget, and output gates— that regulate the information flow, enabling the model to choose what information to keep or discard and to efficiently grasp long-term dependencies [23].

C. CNN

CNNs are created for handling data in grid formats, such as images and videos. Three primary layers make up their composition.

- **Convolutional layers:** Convolutional layers utilize filters on input data to identify particular features like edges. The filter moves across the data to create a map displaying the existence of these features.
- **Pooling layers:** Down sampling through pooling helps decrease the dimensionality of feature maps. This assists in decreasing parameters and computations, enhancing efficiency and reducing overfitting. Max-pooling and average-pooling are frequently employed methods.

- **Fully connected layers:** Fully connected layers link each neuron in one layer to every neuron in the following layer, commonly employed in classification scenarios. The feature maps are flattened and moved through these layers to produce probabilities for each class.

- CNNs may consist of activation, dropout, and normalization layers. Parameters are adjusted using gradient descent during their training via backpropagation. CNNs are very successful in tasks related to processing images and videos, such as classification, object detection, and segmentation [24].

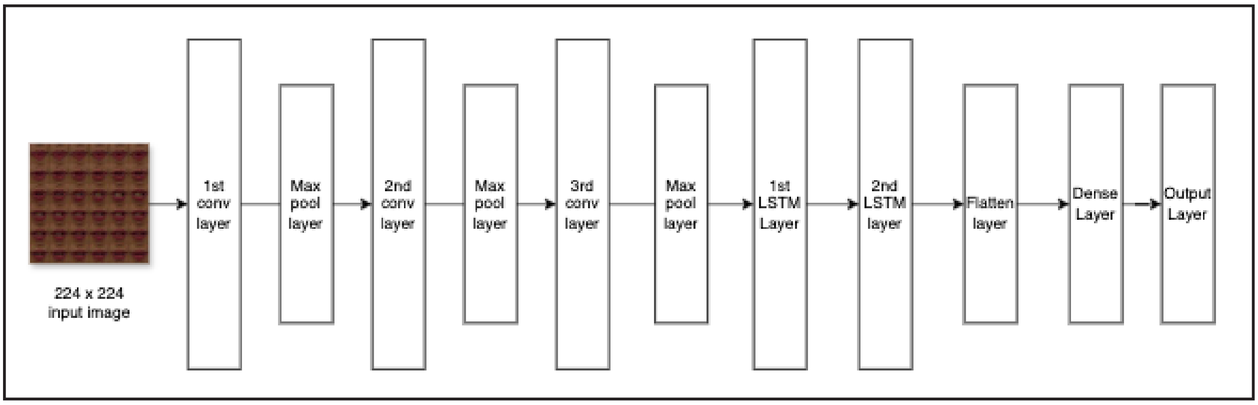

Fig. 7: Classification Process with Model Architecture

D. Model Architecture

Here combining Convolutional Neural Networks (CNN) with Long Short-Term Memory (LSTM) networks we have created the model. The CNN layers focused on extracting spatial features, whereas the LSTM layers were used to capture the temporal relationships in the sequences of lip movements. In this implementation, we included a max pooling layer and applied a dropout rate of 0.2 following the dense layer which contained 374 neurons. For the training process, we configured a batch size of 64 and established a learning rate of 0.001. The model underwent training for 100 epochs while utilizing the RMSprop optimizer. The model's performance on the validation set was monitored after each epoch. The classification process is shown in Fig. 7. The detailed information on Model Architecture with Layers and Output Shapes is reported in Table 4.

Table 4: Proposed model Architecture with Layers and Output Shapes

Layer (Type)	Output Shape
Conv3D	(None, 28, 24, 32, 64)
MaxPooling3D	(None, 9, 8, 10, 64)
Conv3D	(None, 9, 8, 10, 128)
MaxPooling3D	(None, 9, 8, 10, 64)
Conv3D	(None, 9, 8, 10, 128)
MaxPooling3D	(None, 9, 8, 10, 64)
Conv3D	(None, 9, 8, 10, 128)
LSTM (Layer 1)	(None, 9, 8, 10, 128)
LSTM (Layer 2)	(None, 9, 8, 10, 128)
Flatten	(None, 2304)
Dense	(None, 2304)

E. Activation Function

We have used two activation function namely ReLU and softmax. ReLU (Rectified Linear Unit) and Softmax are both commonly used activation functions in neural networks, each with a specific purpose.

Rectified Linear Unit (ReLU): ReLU, also known as Rectified Linear Unit, is an activation function that outputs the input value or zero, whichever is greater. It turns on with positive input and gives out zero for negative input. ReLU allows the model to learn complex relationships by adding non-linearity, which is why it is useful in deep learning. It is efficient in terms of computation and commonly employed in hidden layers to enhance the model’s ability to represent information. The function is described as:

$$f(x) = \max(0, x)$$

Softmax: Softmax serves as an activation function employed in the output layer specifically for classification assignments. It converts the final hidden layer's output into a probability distribution, ensuring that the total probabilities add up to 1. The predicted output is chosen from the class with the highest likelihood. Softmax is frequently combined with cross-entropy loss to assist the model in reducing the gap between predicted and actual class distributions. The function is described as:

$$\text{Softmax}(x_1) = \frac{e^{x_i}}{\sum_j e^{x_j}}$$

The overall performance of the model is then evaluated by comparing the results and accuracy from each of the validation and testing units.

RESULTS AND DISCUSSION

Once lip sequences are extracted, the videos are divided into segments, with each segment representing a spoken digit. The lip area is separated using methods in image processing, eliminating unimportant surroundings, and the lip pictures are adjusted for uniformity. After that, every lip sequence is assigned a label matching its digit. In convolutional neural networks, the convolutional layers generate image features using a 3x3 filter size and are often paired with max pooling layers to reduce dimensionality. Next, a feature vector is created for classification by a fully connected layer, where the weights are trained using backpropagation. LSTM networks handle input sequences in a step-by-step manner when dealing with sequential data. Every LSTM cell employs three gates: the forget gate (for discarding data), the input gate (for incorporating new data), and the output gate (for determining output according to the cell state). These gates aid in capturing extensive dependencies by tuning weights through backpropagation during training. The data was into an 80% training set and a 20% testing set to validate. The model was trained using the training set. The combined images were fed into CNN layers to extract features and then passed through max pooling layers to reduce dimensionality. Following three convolutional and max pooling layers, two LSTM layers were used to capture temporal patterns in the output. In the end, the result was inputted into two fully connected layers prior to reaching the output layer.

Validation and Testing

To prevent overfitting and enhance model performance, 10% of the dataset was reserved for validation during training. Validation accuracy was monitored by evaluating the softmax activation function after each batch, allowing for necessary hyperparameter adjustments. This approach ensured that the model remained robust, avoiding excessive specialization and enabling it to generalize effectively to new data. Following the training phase, 20% of the dataset was allocated for testing to assess the model's final performance. Maintaining a consistent batch size throughout training was crucial for ensuring stable and effective learning. The model's ability to generalize to fresh data was evaluated using validation accuracy tracked through the softmax function. The study compares the performance of three different models RNN+LSTM, RNN+CNN, and the proposed CNN+LSTM model-based on test accuracy, class count, kernel size, and the number of training epochs, as reported in Table 5. The CNN+LSTM model achieved the highest accuracy (83%), outperforming both RNN+LSTM and RNN+CNN. This indicates that integrating CNN for feature extraction with LSTM for sequence learning provides a more robust solution for automatic lip-reading. This hybrid approach proves particularly effective in handling the unique challenges of recognizing Assamese digits in noisy environments. With an accuracy of 83%, the model is now ready for real-time deployment, offering a reliable solution for interpreting lip movements in challenging acoustic conditions.

Confusion Matrix

A confusion matrix evaluates a classifier's performance by organizing predictions into False Positives (FP), True Positives (TP), False Negatives (FN), and True Negatives (TN) [14]. Figure 8 represents the Confusion Matrix for LSTM (Lip Reading of Digits (0-9)).

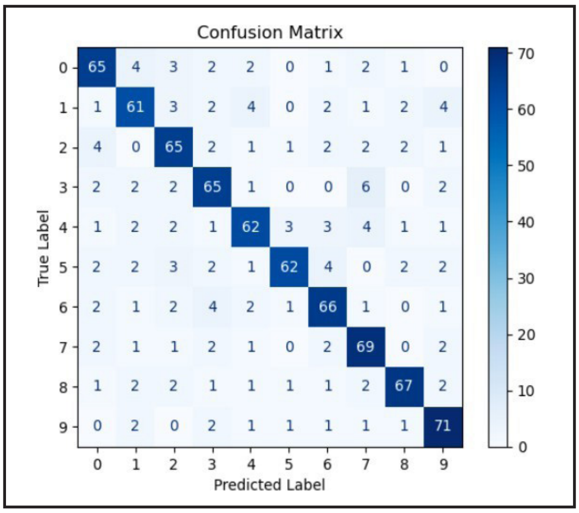

Fig. 8: Confusion Matrix for LSTM (Lip Reading of Digits (0–9))

Table 5: Model with Accuracy

Model	Test Accuracy	Class	Kernel Size	Epochs
RNN+LSTM	76%	10	3x3x3	100
RNN+CNN	68%	10	3x3x3	100
Proposed (CNN+LSTM)	83%	10	3x3x3	100

Analysis Between Accuracy Curve and Loss Curve

The loss curve and accuracy curve of proposed method are shown in Figs. 9, and 10. Form Fig. 9, the training and validation loss decreasing steadily over epochs, indicating effective model learning. The small gap between the curves suggests minimal overfitting and good generalization to unseen data. The consistent downward trend implies the model is improving with more training. Similarly, from Fig. 10 shows training and validation accuracy steadily increasing over epochs, indicating improved model performance. The training accuracy remains slightly higher than validation accuracy, but the small gap suggests minimal overfitting and good generalization. The upward trend demonstrates effective learning.

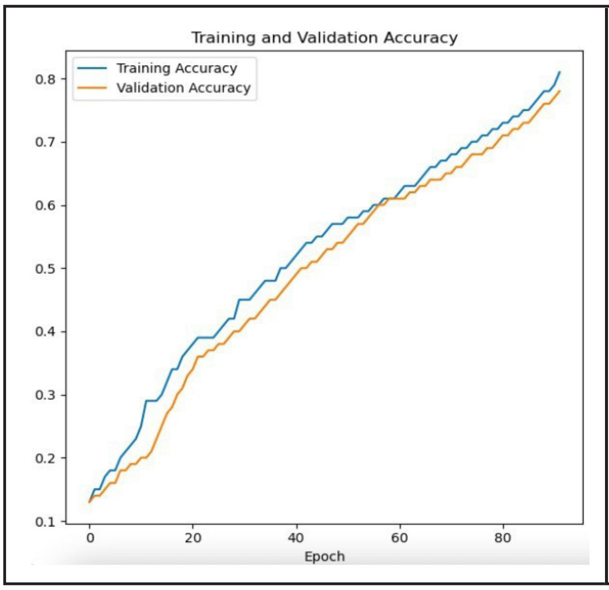

Fig. 9: Accuracy Curve

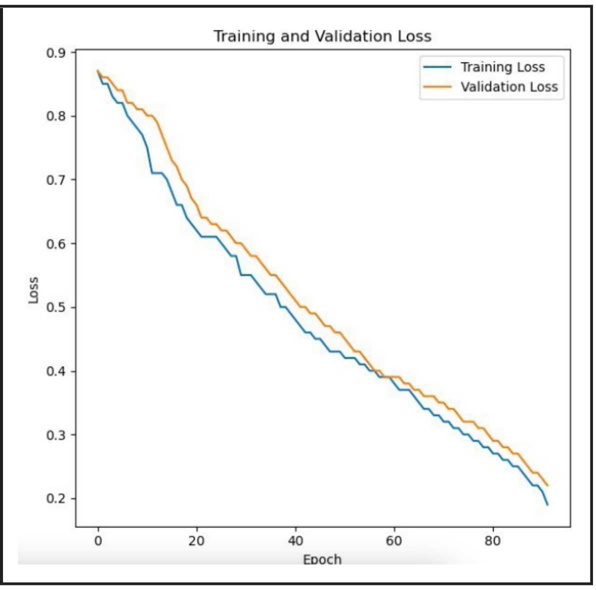

Fig. 10: Loss Curve

CONCLUSION AND FUTURE DIRECTIONS

Our research on lip-reading Assamese numbers demonstrates significant progress in understanding language signals within a regional context. By leveraging a custom CNN+LSTM model and a self-collected dataset, we establish the feasibility of applying lip-reading technology to regional languages. A comparative analysis of different models RNN+LSTM, RNN+CNN, and CNN+LSTM reveals that the proposed CNN+LSTM model outperforms the others with an accuracy of 83%. The integration of CNN for spatial feature extraction and LSTM for temporal dependency learning proves to be the most effective approach for recognizing Assamese digits in noisy environments. This study holds promising implications for improving accessibility, particularly for deaf and hard-of-hearing individuals, enabling them to understand spoken Assamese digits without relying on sound. Additionally, it enhances human-computer interactions, paving the way for voice-free command execution in Assamese. These advancements contribute to the broader goal of inclusive communication technologies, making digital interactions more accessible across diverse linguistic and auditory needs.

This study establishes a strong foundation for future advancements in lip-reading technology. Future research can explore pre-trained models or develop customized neural network architectures to further enhance accuracy. Expanding the dataset in both size and diversity will improve the model's generalization, making it more robust for wider applications and diverse users. Our findings underscore the crucial role of deep learning in advancing lip-reading technology and highlight the importance of transfer learning for multilingual adaptability. Extending these advancements to other languages will contribute to a more inclusive and accessible technological landscape, broadening the impact of lip-reading solutions in real-world scenarios.

CONFLICTS OF INTEREST

No conflict of interest was declared by the authors.

REFERENCES

Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. The journal of the acoustical society of america, 26(2), 212-215.

Fisher, C. G. (1968). Confusions among visually perceived consonants. Journal of speech and hearing research, 11(4), 796-804.

Elman, J. L. (1993). Learning and development in neural networks: The importance of starting small. Cognition, 48(1), 71-99.

Treiman, R., Kessler, B., & Bick, S. (2002). Context sensitivity in the spelling of English vowels. Journal of Memory and Language, 47(3), 448-468.

Matthews, I., Cootes, T. F., Bangham, J. A., Cox, S., & Harvey, R. (2002). Extraction of visual features for lipreading. IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(2), 198-213.

Koller, O., Ney, H., & Bowden, R. (2015). Deep learning of mouth shapes for sign language. In Proceedings of the IEEE International Conference on Computer Vision Workshops (pp. 85-91).

Pawar, D., Borde, P., & Yannawar, P. (2024). Generating dynamic lip-syncing using target audio in a multimedia environment. Natural Language Processing Journal, 100084.

Adeel, A., Gogate, M., Hussain, A., & Whitmer, W. M. (2019). Lip-reading driven deep learning approach for speech enhancement. IEEE Transactions on Emerging Topics in Computational Intelligence, 5(3), 481-490.

Chickerur, S., Patil, M. S., Anand, M. E. T. I., Nabapure, P. M., Mahindrakar, S., Sonali, N. A. I. K., & Kanyal, S. (2019). LSTM Based Lip Reading Approach for Devanagiri Script. ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal, 8(3), 13.

Assael, Y. M., Shillingford, B., Whiteson, S., & De Freitas, N. (2016). Lipnet: End-to-end sentence-level lipreading. arXiv preprint arXiv:1611.01599.

Stafylakis, T., & Tzimiropoulos, G. (2017). Combining residual networks with LSTMs for lipreading. arXiv preprint arXiv:1703.04105.

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

Kalbande, D., & Patil, S. (2011, September). Lip reading using neural networks. In International Conference on Graphic and Image Processing (ICGIP 2011) (Vol. 8285, pp. 310-316). SPIE.

Garg, A., Noyola, J., & Bagadia, S. (2016). Lip reading using CNN and LSTM. Technical report, Stanford University, CS231 n project report.

Vakhshiteh, F., Almasganj, F., & Nickabadi, A. (2018). Lip-reading via deep neural networks using hybrid visual features. Image Analysis and Stereology, 37(2), 159-171.

Faisal, M., & Manzoor, S. (2018). Deep learning for lip reading using audio-visual information for urdu language. arXiv preprint arXiv:1802.05521.

Zhu, M.-l., Wang, Q.-q., and Luo, J.-l. (2019). Lip-reading based on deep learning model. Transactions on Edutainment XV, pages 32–43.

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

Shirakata, T., & Saitoh, T. (2020). Lip reading using facial expression features. Int. J. Comput. Vis. Signal Process, 1 (1), 9-15.

Miled, M., Messaoud, M. A. B., & Bouzid, A. (2023). Lip reading of words with lip segmentation and deep learning. Multimedia Tools and Applications, 82(1), 551-571.

Hochreiter, S. (1997). Long Short-term Memory. Neural Computation MIT-Press.

Staudemeyer, R. C., & Morris, E. R. (2019). Understanding LSTM--a tutorial into long short-term memory recurrent neural networks. arXiv preprint arXiv:1909.09586.

Yu, W., Zhang, J., and Li, Y. (2019). A review of lstm networks and their applications in speech recognition. IEEE Transactions on Speech and Audio Processing, 27(6):987–1000.

Wagle, A., Sharma, B., and Singh, C. (2021). An overview of convolutional neural networks (cnns) and their applications. International Journal of Artificial Intelligence Research, 45(3):123–134.

Townsend, J. T. (1971). Theoretical analysis of an alphabetic confusion matrix. Perception & Psychophysics, 9, 40-50.