International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 04 | April - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Lip Reading Using Visual Cues

Sandesh Shinde*1, Ishita Shinde*2, Pranita Shinde*3, Mandar Shitole*4
*1,2,3,4 Student, Department of Information Technology, SVPM's College of Engineering, Malegaon(Bk), Maharashtra, India

---***---

Abstract - Lip reading has emerged as a challenging and impactful area of research within speech recognition, with potential applications ranging from assistive technologies for the hearing impaired to communication in noisy environments. This project proposes a deep learning-based system designed to recognize and decode sentences from videos using only visual information. The system is lexicon-free and focuses on the classification of visemes visual representations of phonemes by leveraging a spatial-temporal convolutional neural network (CNN) in combination with a transformer-based architecture for sequence processing. A novel perplexity analysis approach is employed to convert recognized visemes into words. The model has been trained and evaluated on the real time videos, demonstrating significant improvements in word error rate (WER) and viseme classification accuracy. Furthermore, the system is robust to varying lighting conditions, making suitable for practical applications. The results indicate a promising direction for real-time automated lip reading, offering new possibilities in fields such as human-computer interaction, surveillance, and assistive communication technologies.

Key Words: Lip Reading, Deep Learning, CNN, Transformer, Perplexity Analysis, Visual Speech Recognition.

1. INTRODUCTION

Lip reading has long been a challenging task in the field of speech recognition. The ability to decipher speech purely from visual inputs can enable communication in noisy environments or assist individuals with hearing impairments. Lip reading, also referred to as visual speech recognition, involves interpreting lip movements to understand spoken words in the absence of audio. In recent years, advances in deep learning have led to significant improvements in the field of automated lip reading. Most approaches rely on recognizing words or sentences based on visible lip movements, either by classifying entire words or sequences of phonemes. Despite advancements, lip reading remains an unsolved problem due to several factors :

- **Ambiguity in Visemes:** Many phonemes (basic units of sound) share the same visual lip patterns, known as visemes, leading to difficulty in distinguishing between certain sounds, such as /p/ and /b/.

Generalization: Most systems fail to generalize to words or sentences that were not seen during training.

- This limits their applicability in real-world scenarios with a wide range of vocabulary.
- **Environmental Variations:** Lighting conditions, facial orientations, and individual differences in lip movements further complicate the task of accurate lip reading. This paper introduces a deep learning-based approach to lip reading that overcomes many of these challenges. The system utilizes spatial-temporal CNNs for feature extraction from video frames and a transformer-based classifier to predict visemes. A key contribution of this work is the introduction of a perplexity analysis module that converts visemes into words. The model is evaluated on the videos, which includes a wide variety of speakers, lighting conditions, and sentence structures, making it one of the most important benchmarks for lip reading.

2. METHODOLOGY

A. Data Preprocessing

To enable lip reading, the raw input videos need to be pre-processed into a suitable format.

Preprocessing steps include:

1. **Facial Landmark Detection:** Using a facial landmark detector, the system isolates the region of interest (ROI) around the lips. This ensures that only relevant visual information is used for feature extraction, eliminating unnecessary background details.

2. **Grayscale Conversion:** Each video frame is converted to grayscale to reduce computational complexity and focus on the structural details of the lips.
3. **Resizing and Normalization:** The cropped lip region is resized to a fixed resolution of 64×64 pixels, and pixel values are normalized to standardize the input for the CNN. Normalization ensures that the input data follows a consistent scale, improving the model's performance.
4. **Data Augmentation:** Augmentation techniques such as horizontal flipping and random frame removal are applied to increase the diversity of the training data. This helps the model generalize better to variations in lip shapes, orientations, and lighting conditions.

B. Visual Feature Extraction

The pre-processed video frames are passed through a spatial-temporal convolutional neural network (CNN). The CNN extracts spatial and temporal features from the lip region across multiple frames. The architecture includes:

- **3D Convolutional Layers:** These layers capture both spatial (lip movement) and temporal (across multiple frames) information simultaneously, making them ideal for video processing.
- **2D ResNet Layers:** A residual network (ResNet) is employed to reduce the spatial dimensions while preserving important features. ResNet’s skip connections prevent vanishing gradients and allow the network to learn deeper representations.

The output of the visual feature extraction stage is a sequence of feature vectors, one for each frame in the input video. These vectors contain high-level information about the movement and shape of the speaker’s lips.

C. Viseme Classification

Visemes are the visual equivalents of phonemes. While phonemes represent sounds, visemes represent the corresponding lip movements. In viseme classification, multiple phonemes can map to the same viseme, which creates ambiguity. The feature vectors generated by the CNN are fed into a transformer-based viseme classifier. The transformer model uses an attention mechanism to capture long-range dependencies between frames, allowing it to understand the sequence of lip movements over time.

- **Transformer Encoder:** The encoder processes the input sequence of feature vectors and generates context aware embeddings for each frame. The self-attention mechanism helps the model focus on important frames and their relationships to surrounding frames.
- **Transformer Decoder:** The decoder uses these embeddings to predict the corresponding visemes. The output of the decoder is a sequence of visemes, one for each segment of the input video.

The viseme classifier is trained to minimize cross-entropy loss over 17 output classes, which include 13 viseme classes and additional markers for padding, start-of-sentence, and end-of-sentence tokens.

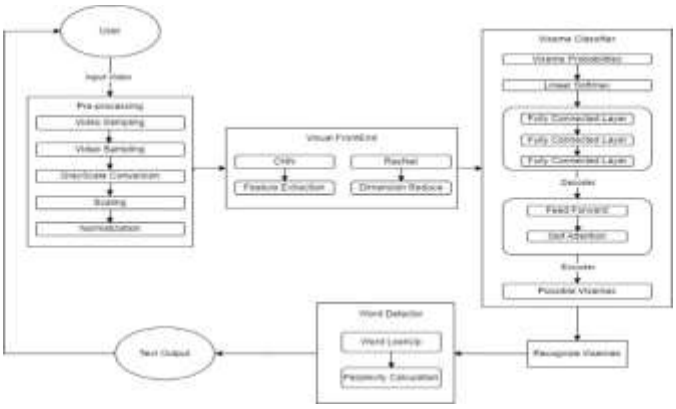
D. Word Detection via Perplexity Analysis :-

Once the visemes have been predicted, the next step is to convert them into meaningful word. This conversion is not straightforward due to the one-to-many mapping between visemes and words (i.e., different words can share the same lip movements). To address this, a perplexity analysis module
is introduced.

1. **Viseme-to-Word Mapping:** Words can correspond to the same viseme sequence, the system generates multiple candidate sentences.
2. **Beam Search:** A beam search algorithm is used to explore different combinations of words that match the viseme sequence. At each step, the algorithm keeps track of the top 50 most likely word combinations.
3. **Perplexity Scoring:** The likelihood of each candidate sentence is evaluated using a perplexity score. Perplexity measures the grammaticality and coherence of a sentence based on statistical language modelling. The sentence with the lowest perplexity score is selected as the final output.

By using perplexity analysis, the system can handle unseen words and ambiguous visemes more effectively, leading to more accurate sentence predictions.

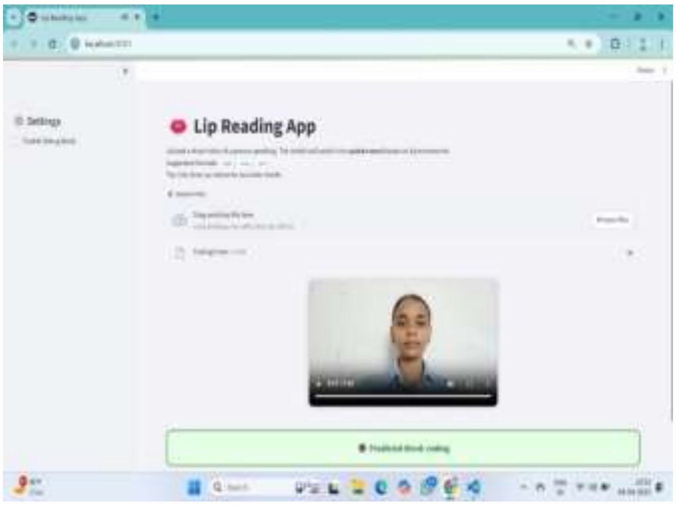
E. System Architecture



The system architecture consists of five main components: User Input, Pre-processing, Visual Frontend, Viseme Classifier, and Word Detector. The following is a detailed explanation of the architecture:

1. **User Input:** The user provides a video as input, which contains visual information of the speaker's
   lip movements.
2. **Pre-processing Module:** The pre-processing module is responsible for preparing the input video by performing:
   - **Video Sampling:** Extracting video frames at specific intervals.
   - **Grayscale Conversion:** Reducing the video to grayscale to decrease the computational load.
   - **Scaling and Normalization:** Standardizing frame dimensions and pixel values to ensure uniform input to the CNN.
3. **Visual Frontend:** The processed video frames are passed through a CNN-ResNet architecture for feature extraction:
   - CNN captures spatial and temporal features from the lip region.
   - ResNet reduces dimensionality while retaining important lip movement details.
4. **Viseme Classifier:** The visual features extracted from the frontend are classified into visemes using a transformer-based model:
   - The encoder-decoder architecture models the temporal relationships between frames and classifies lip movements into viseme sequences.
   - A Softmax layer outputs probabilities for each viseme class.
5. **Word Detection Module:** The classified visemes are mapped to words using a word lookup mechanism:
   - A beam search algorithm explores the best possible word sequence by using perplexity scores.
   - The sentence with the lowest perplexity is selected as the final text output.
6. **Text Output:** The final recognized sentence is displayed as the output, completing the lip-reading process.

3. RESULT



This Project processes video data by extracting frames and detecting the mouth region using MediaPipe. It resizes the mouth area to 64x64 pixels for each frame, normalizes the data, and organizes it into sequences. Then it prepares training and validation videos with labels, ensuring each class has enough samples.

During training, it uses a 3D CNN + LSTM model to learn spatial-temporal lip features, with K-Fold cross-validation. The output in your terminal shows how many classes were found, how many videos were processed, sequence lengths, class distribution, and model training progress with accuracy/loss per epoch. The best-performing model is saved as lip_reading_model.h5.

4. CONCLUSIONS

This paper presents a deep learning-based system for lip reading, capable of predicting sentences using only visual cues. The model achieves significant improvements in accuracy compared to existing methods and demonstrates robustness to varying environmental factors.

REFERENCES

[1] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, “Lip reading sentences in the wild,” IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[2] B. Shillingford, Y. Assael, M. Hoffman, et al., “Large-scale visual speech recognition,” INTERSPEECH, 2018.

[3] T. Afouras, J. S. Chung, and A. Zisserman, “Deep lip reading: A comparison of models and an online application,” INTERSPEECH,2018.

[4] S. Fenghour, D. Chen, and P. Xiao, “Decoder-encoder LSTM for lip reading,” 8th International Conference on Software and Information Engineering (ICSIE), 2019.

[5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, ‘‘Attention is all you need,’’ in Proc. NIPS, 2017, pp. 5998–6008.

[6] I. Matthews, T. Cootes, J. Bangham, S. Cox, and R. Harvey, "Extraction of visual features for lipreading," IEEE Trans. Pattern Anal. Mach. Intell., vol. 24, no. 2, pp. 198-213, 2002.