

Real-Time Lip Reading Using A 3d Cnn–Bilstm Architecture for Visual Speech Recognition

Harshini R¹, Poojabharani S², Sanjana S³, Saranya R⁴

Dept. Of Artificial Intelligence And Data Science Sengunthar Engineering College Tiruchengode, India

Abstract—Lip reading is the task of recognising spoken words from the visible movements of the lips without using any audio signal. This paper presents a real-time visual speech recognition system that captures live video from an ordinary webcam, locates the mouth region, and predicts the spoken word or phrase as text. The system combines 3D convolutional layers, which learn spatial and temporal lip-movement features, with bidirectional LSTM layers that model the whole sequence in both directions, and it is trained with Connectionist Temporal Classification (CTC) loss so that no frame-level alignment is needed. A lightweight pipeline of MediaPipe landmark detection, mouth-region preprocessing and a sliding frame buffer allows the model to run on a standard computer. On a small custom dataset of four phrases, the system achieved a word error rate (WER) of 16.67% and a character error rate (CER) of about 4–5%. The results indicate that a compact, audio-free lip-reading system is feasible for assistive communication and for noisy environments.

Index Terms—lip reading, visual speech recognition, silent speech interface, 3D CNN, BiLSTM, CTC, real-time systems.

I. INTRODUCTION

Speech is the main means of human communication, but audio-based communication becomes difficult in the presence of background noise, hearing difficulty, or when a person cannot speak aloud. Human speech perception is multimodal: besides sound, listeners use the movements of the lips, jaw and face, and for people with hearing difficulties lip reading is often the main way of following a conversation. However, many different sounds look almost identical on the lips (visemes), so even human lip reading is not fully reliable. Automatic lip reading, also called visual speech recognition (VSR), uses deep neural networks to learn the required features directly from video frames.

Silent speech interfaces (SSIs) recognise speech-related information without audible speech. They use different sensors: cameras observe the lips and face, electromyography (EMG) records muscle activity, ultrasound observes the tongue, and EEG, ECoG or iEEG record brain signals. Among these, the camera-based approach is the least invasive and the easiest to deploy, since a webcam is available in almost every laptop and phone. Applications include assistive communication for people with hearing or speech impairments, recognition in noisy places, private silent commands, and support for audio-visual recognition when audio quality is poor.

The main difficulties are the similarity of lip shapes for different sounds, differences between speakers in face shape

and speaking speed, changes in lighting and head pose, and the limited availability of labelled video data. Most high-performing systems also depend on very large datasets and powerful GPUs, and many are evaluated only on recorded benchmark videos rather than on a live camera stream.

The contributions of this work are: (1) a real-time, audio-free lip-reading pipeline that works on a standard computer with a webcam; (2) a 3D CNN–BiLSTM model trained with CTC loss that learns spatiotemporal lip features without frame-level alignment; and (3) an evaluation of the system using CER and WER. The scope is limited to word- and phrase-level recognition of a predefined set of classes with a single webcam and no audio.

II. RELATED WORK

Early deep-learning lip-reading models combined convolutional networks with recurrent layers. Martinez et al. replaced recurrent layers by temporal convolutional networks (TCN) and introduced variable-length augmentation, evaluated on isolated-word datasets [1]. Zhao et al. transferred knowledge from a speech recognizer to a lip-reading model by cross-modal distillation [2], and Ma et al. reduced computation with a depthwise-separable TCN and self-distillation, at the cost of an accuracy–size trade-off [3]. A survey by Fenghour et al. reviews datasets, feature extraction and classification methods for automated lip reading [4].

Audio-visual approaches combine sound with lip movements. Ma et al. proposed an end-to-end system using a ResNet-18 front end and Conformer encoder with CTC/attention [5]; Shi et al. learned audio-visual representations by masked multimodal cluster prediction [8]; Afouras et al. studied deep audio-visual recognition [9]; and Auto-AVSR used automatically generated transcriptions of unlabeled videos to enlarge the training set, reporting 0.9% WER on LRS3 [10]. These systems achieve very low error rates, but they cannot be used when audio is

unavailable. Visual-only work has addressed multilingual recognition in the wild [6] and sub-word-level modelling with visual attention [7]. Most recently, large language models have been attached to visual encoders to resolve homophenes, i.e. words with similar lip movements, in VSP-LLM [11], and Llama-AVSR reports about 0.77% WER for audio-visual recognition on LRS3 [12].

The reviewed methods generally need large datasets, long training time and powerful GPUs, are often evaluated on recorded benchmark videos, or rely on audio or very large language models that are hard to run in real time on ordinary computers. This motivates a simple, lightweight, visual-only system that runs on a normal computer with a webcam and displays recognised text immediately. The proposed system addresses this need with a 3D CNN-BiLSTM architecture and CTC decoding.

III. PROPOSED SYSTEM

1. System Overview

The system captures live video, detects the face, extracts the lip region, preprocesses it, and arranges the frames into a temporal sequence. The sequence is passed to a trained 3D CNN-BiLSTM model whose output is decoded by CTC into text and displayed on screen together with the live feed and the processed mouth region. Fig. 1 shows the functional blocks of the system.

time, so each filter responds to a lip-movement pattern such as opening or closing over a few frames; each convolution is followed by a ReLU activation and 3D max pooling, which reduces computation and sensitivity to small shifts. Second, the convolution output is arranged as a sequence of feature vectors and processed by bidirectional LSTM layers, which read the sequence forwards and backwards so that every prediction uses both previous and following context. Third, dropout randomly switches off neurons during training to reduce overfitting. Finally, a dense layer with softmax activation gives a probability for every character or symbol class at each time step.

2. Training And Decoding

The network is trained with CTC loss, which considers all possible alignments between the output and the label through a blank symbol; therefore no manual frame-to-character alignment is needed and variable-length sequences are handled. Training uses the Adam optimiser with a learning rate of $1e-4$, 20–50 epochs and a batch size of 16 or 32. During inference, CTC decoding removes repeated characters and blanks; greedy decoding selects the most probable symbol at each step, while beam search keeps several candidate paths and selects the best one.

3. Real-Time Prediction

In live operation the system does not decide from a single frame. Each valid ROI frame is added to a sliding buffer; when the buffer is full the model is run, the decoded text is displayed over the video, and the buffer is refreshed for the next word. Frames without a detected face are skipped. The interface shows the live feed, the processed mouth region (so that the user can check correct lip detection) and the predicted text.

IV. EXPERIMENTAL SETUP



Fig. 1. Functional block diagram of the real-time lip-reading system.

1. Data Acquisition And Preprocessing

Frames are read continuously from the webcam using OpenCV. The face and lip landmarks are located with MediaPipe face mesh, and only the mouth region of interest (ROI) is retained, which removes irrelevant facial regions and background. The ROI is then (i) converted to grayscale to reduce data size and lighting effects, (ii) resized to a fixed dimension, and (iii) normalised by subtracting the mean and dividing by the standard deviation. A fixed number of consecutive frames is stacked in temporal order to form the input tensor; short sequences are padded and long ones trimmed, and the frame order is always preserved. Identical steps are applied to training videos and live frames. Because the dataset is small, simple augmentation such as slight brightness changes and small shifts of the cropped region is applied to training frames.

2. Model Architecture

The model is inspired by LipNet and has four parts. First, multiple Conv3D layers slide a 3D filter over height, width and

3. Dataset

The dataset consists of video samples of a speaker saying a predefined set of phrases; each video is converted into a frame sequence containing the lip movements. Table I lists the classes and the number of samples. The data is divided into training, validation and test subsets in the ratio 70%, 15% and 15%, so that the validation set monitors training and the test set evaluates the final model.

Table I Dataset Overview

Class (phrase)	Samples
Hello	20
Welcome	20
Thank You	20
How Are You	20
Total	80

4. Hardware And Software

Experiments were run on a computer with an AMD Ryzen 7 processor (4.40 GHz), 16 GB RAM and 1 TB disk, with a standard webcam, under Windows 10. The system is implemented in Python with OpenCV and MediaPipe for videoprocessing; Jupyter Notebook was used for development and Streamlit for the user interface.

5. Evaluation Metrics

Performance is measured with the character error rate (CER) and the word error rate (WER), both computed from the edit distance between the predicted and reference text:

$$WER \text{ (or CER)} = (S + D + I) / N \times 100 \quad (1)$$

where S, D and I are the numbers of substitutions, deletions and insertions, and N is the total number of words (or characters) in the reference text. Lower values indicate better accuracy.

6. Results And Discussion

The system was tested with live webcam input. It continuously captured frames, located the mouth region, passed the frame sequences to the trained model and displayed the predicted text; sample outputs such as the word “HELLO” and the phrase “How are you” were obtained in real time. Table II summarises the measured error rates.

Table II Recognition Error Rates

Metric	Value
Word error rate (WER)	16.67%
Character error rate (CER)	≈ 4–5%

A WER of 16.67% means that about one word in six was predicted incorrectly, whereas the CER is much lower

because a wrong word is often only partly wrong. The errors arise mainly from visually similar lip shapes (for example /p/, /b/ and /m/), poor lighting and the limited training data. Performance can also be affected by camera quality, head movement, speaking speed and differences between speakers. These results are obtained on a small four-phrase dataset, so they demonstrate feasibility of the pipeline rather than general performance; evaluation on larger, multi-speaker benchmarks is required to compare with the systems in Section II.

V. CONCLUSION AND FUTURE WORK

This paper presented a real-time, audio-free lip-reading system based on a 3D CNN-BiLSTM model trained with CTC loss. The complete pipeline, from webcam capture and mouth-region extraction to text display, was implemented and tested, and the model recognised the trained phrases with a WER of 16.67% and a CER of about 4–5%. The system needs no special sensors, so it is low-cost and non-contact, and can support silent communication and accessibility applications.

Future work includes training on larger and more diverse datasets with many speakers, environments and lighting conditions; larger vocabulary and sentence-level recognition; Transformer and attention-based models; language-model integration to resolve words with similar lip movements; multilingual support including regional languages; text-to-speech output for people who cannot speak aloud; audio-visual fusion; and model optimisation by pruning or quantisation for mobile and cloud deployment.

Acknowledgment

The authors thank their supervisor Ms. N. Indhuja, M.E., Assistant Professor, and the Head of the Department, Mrs. Santhipriya, M.E., Department of Artificial Intelligence and Data Science, Sengunthar Engineering College, Tiruchengode, for their guidance and support.

REFERENCES

1. B. Martinez, P. Ma, S. Petridis, and M. Pantic, “Lipreading using temporal convolutional networks,” in Proc. IEEE ICASSP, 2020, pp. 6319–6323.
2. Y. Zhao, R. Xu, X. Wang, P. Hou, H. Tang, and M. Song, “Hearing lips: Improving lip reading by distilling speech recognizers,” in Proc. AAAI Conf. Artif. Intell., vol. 34, no. 4, 2020, pp. 6917–6924.
3. P. Ma, B. Martinez, S. Petridis, and M. Pantic, “Towards practical lipreading with distilled and efficient models,” in Proc. IEEE ICASSP, 2021, pp. 7608–7612.

4. S. Fenghour, D. Chen, K. Guo, B. Li, and P. Xiao, “Deep learning-based automated lip-reading: A survey,” *IEEE Access*, vol. 9, pp. 121184–121205, 2021.
5. P. Ma, S. Petridis, and M. Pantic, “End-to-end audio-visual speech recognition with conformers,” in *Proc. IEEE ICASSP*, 2021, pp. 7613–7617.
6. P. Ma, S. Petridis, and M. Pantic, “Visual speech recognition for multiple languages in the wild,” *Nature Mach. Intell.*, vol. 4, pp. 930–939, 2022.
7. K. R. Prajwal, T. Afouras, and A. Zisserman, “Sub-word level lip reading with visual attention,” in *Proc. IEEE/CVF CVPR*, 2022, pp. 5162–5172.
8. B. Shi, W.-N. Hsu, K. Lakhotia, and A. Mohamed, “Learning audio-visual speech representation by masked multimodal cluster prediction,” in *Proc. ICLR*, 2022.
9. T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, “Deep audio-visual speech recognition,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 12, pp. 8717–8727, 2022.
10. P. Ma, A. Haliassos, A. Fernandez-Lopez, H. Chen, S. Petridis, and M. Pantic, “Auto-AVSR: Audio-visual speech recognition with automatic labels,” in *Proc. IEEE ICASSP*, 2023.
11. J. H. Yeo, S. Han, M. Kim, and Y. M. Ro, “Where visual speech meets language: VSP-LLM framework for efficient and context-aware visual speech processing,” in *Findings Assoc. Comput. Linguistics (ACL)*, 2024.
12. U. Cappellazzo, M. Kim, H. Chen, P. Ma, S. Petridis, D. Falavigna, A. Brutti, and M. Pantic, “Large language models are strong audio-visual speech recognition learners,” in *Proc. IEEE ICASSP*, 2025.