



An AI Lip-reading System Based on Deep Learning

Chung-Min Wu¹, Yu-Sheng Huang², Yeou-Jiunn Chen², Shih-Chung Chen^{2*}

¹Department of Intelligent Automation Engineering, National Chin-Yi University of Technology, Taichung, Taiwan

²Department of Electrical Engineering, Southern Taiwan University of Science and Technology, Tainan, Taiwan

Paper No. K250067

INTRODUCTION

There are already well-developed speech recognition systems used in smartphones and smart assistants. While convenient for most normal people, they are not helpful for hearing- or speech-impaired individuals. In recent years, artificial intelligence (AI) based systems have explored alternative input methods for the people with hearing or speech disabilities, such as sign language and lip-reading. Lip-reading allows silent communication by analyzing visual lip movements. This study presents an artificial intelligence-based lip-reading system designed to recognize both static and dynamic lip motions. It was implemented using Python and PyTorch, with facial landmarks extracted via MediaPipe Face Mesh and RGB lip images as input[1]. The model includes a 50-layer residual network (ResNet-50) for spatial feature extraction[2-3], followed by bidirectional and unidirectional Long Short-Term Memory networks for temporal modeling. Connectionist temporal classification loss (CTC Loss) was adopted to train the AI model for lip-reading recognition without requiring (LSTM) sequence alignment. The AI Lip-reading system achieved a recognition accuracy of approximately 74.7% and identified 20 lip movement categories successfully, including both isolated words and short visual sentences during preliminary tests.

METHODS & MATERIALS

The AI lip-reading system was developed by using Python, with PyTorch as the deep learning framework and MediaPipe Face Mesh for facial landmark extraction. In the data preprocessing stage, a large number of lip movement videos were collected, including both RGB facial images and 3D landmark data. These sequences, performed by different speakers, included both static and dynamic lip motions. Data augmentation such as horizontal flipping and noise injection was applied to enhance model generalization. The architecture includes three main components: first, ResNet-50 is used for extracting spatial features from static lip images; second, a combination of Bi-LSTM and Uni-LSTM layers is used for temporal modeling to capture contextual dependencies from both directions, with Dropout added to prevent overfitting; finally, CTC Loss is used to train the model without requiring aligned input-output sequences, which is ideal for lip-reading tasks of which word boundaries are ambiguous. The main advantage of this system is its dual-input design—combining images and landmarks—along with its multilayered temporal modeling, allowing it to retain both spatial and sequential features across static and dynamic inputs. The use of CTC Loss also reduces the complexity of annotation and enables sentence-level decoding. The complete system architecture and flow chart are illustrated in Fig. 1.

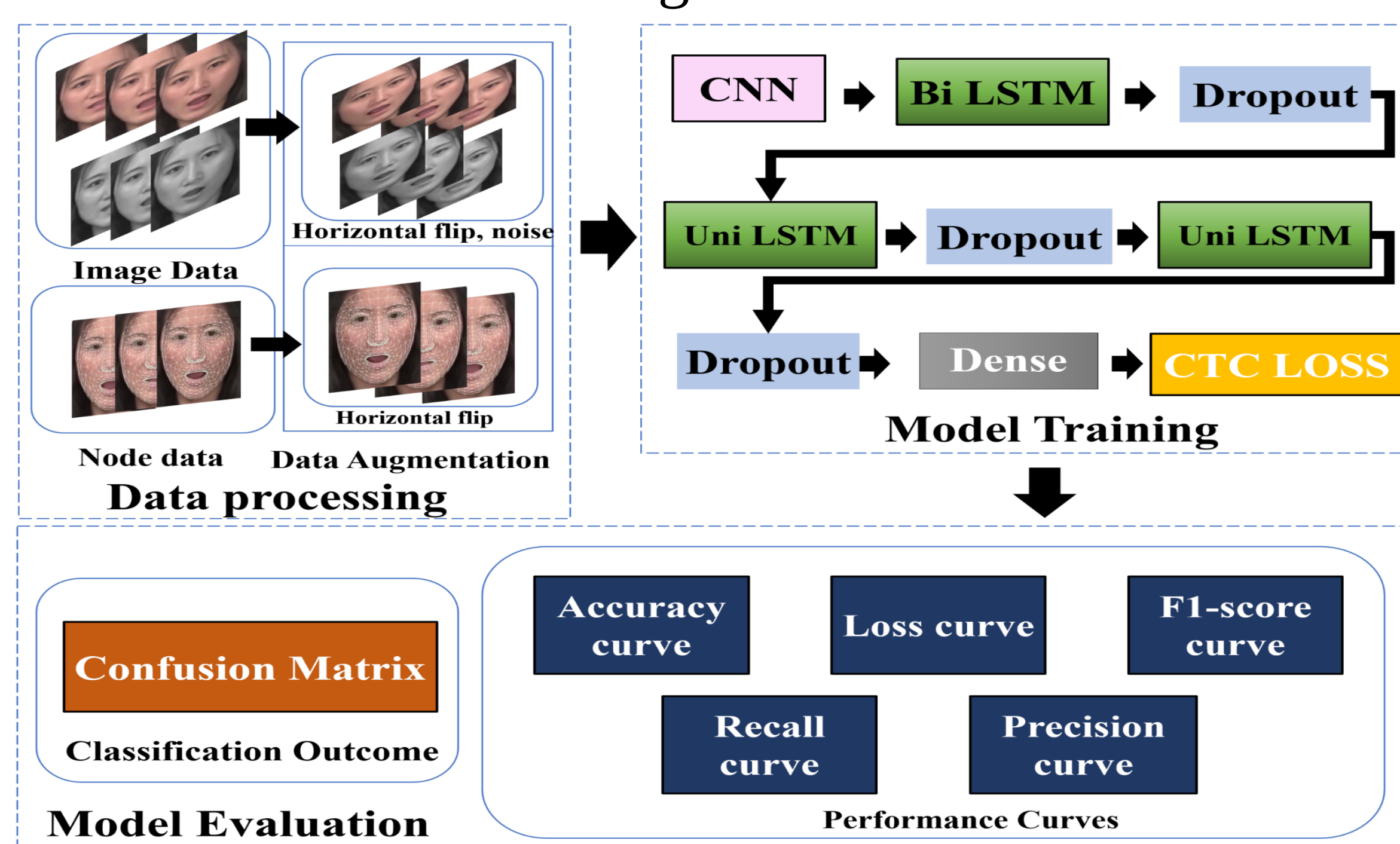


Figure 1. Architecture of the proposed AI-based lip-reading system with multi-source input and temporal modeling structure.

RESULTS AND DISCUSSION

As shown in Figures 2, both accuracy and F1-score increased rapidly during early training and stabilized after epoch 10, reaching approximately 74.7% and 0.72.

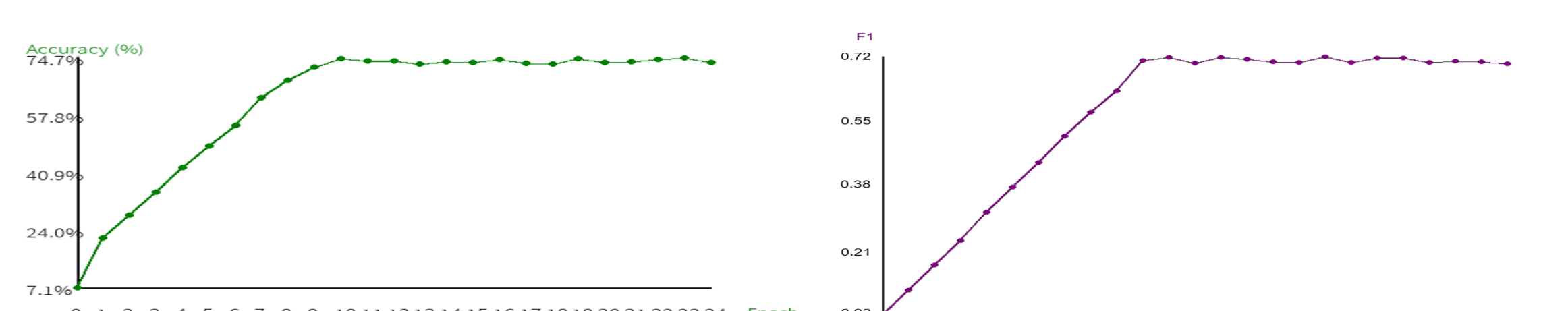


Figure 2. Training curves of the lip-reading model. The left graph shows the accuracy trend, while the right graph presents the F1-score progression across epochs.

This indicates that the model effectively learned lip features and maintained stable performance throughout training. The close trend between accuracy and F1-score suggests balanced predictions across all categories, with no significant skew toward precision or recall. The confusion matrix (Figure 3) shows evenly distributed errors, with no dominant misclassification patterns. This implies consistent classification logic and good generalization ability. Overall, the model demonstrates fast convergence, balanced prediction, and stable outcomes, validating its architectural effectiveness and potential for further improvement.

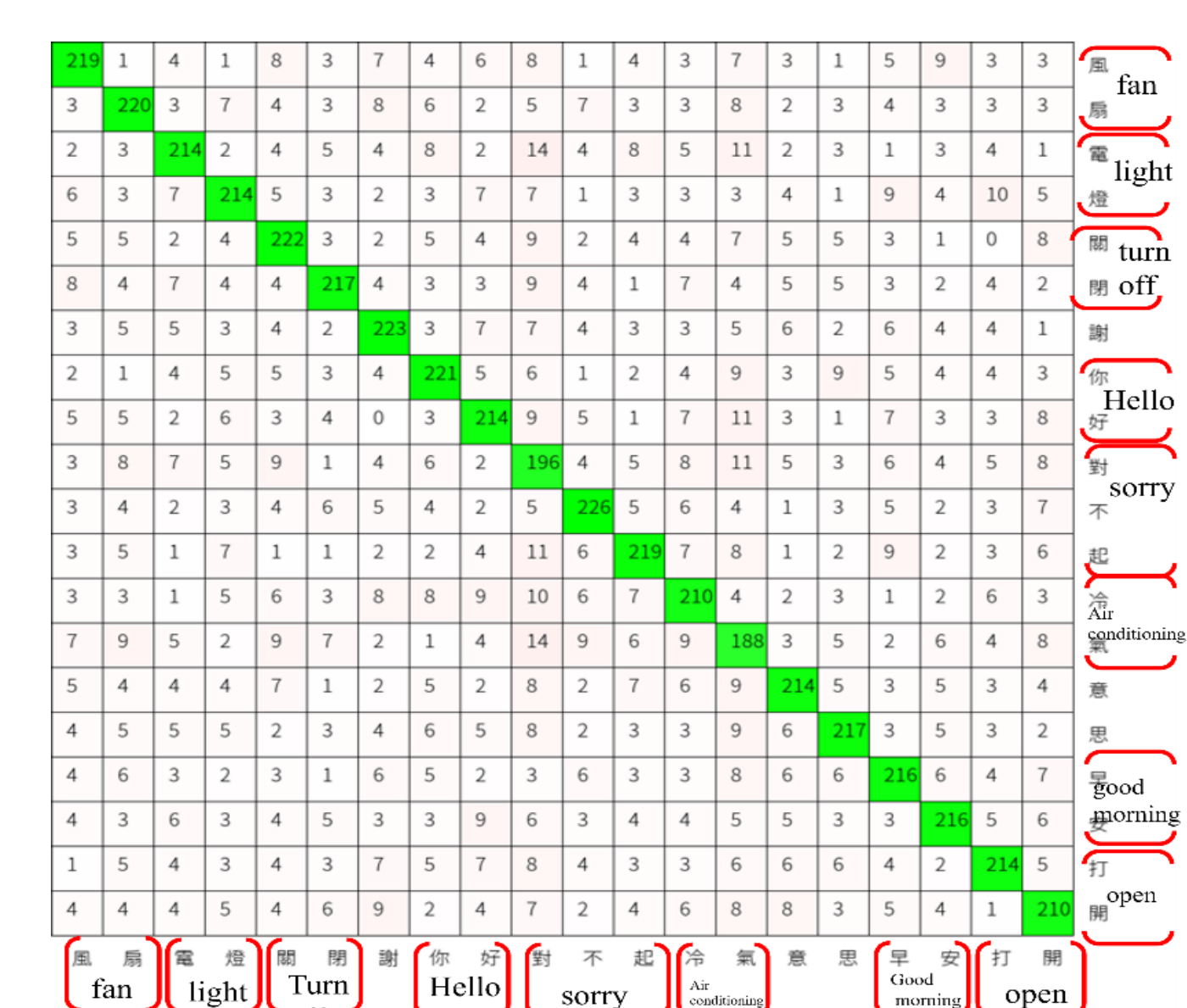


Figure 3. Confusion matrix of the lip-reading model.

CONCLUSIONS

This study developed an AI-based lip-reading system that combines image and landmark inputs. The model applies ResNet-50 for feature extraction, Bidirectional and Unidirectional Long Short-Term Memory networks for temporal modeling, and Connectionist Temporal Classification Loss for sequence output. Training curves and the confusion matrix indicate stable convergence, balanced classification, and good generalization, validating its effectiveness for static and dynamic lip-reading. Although real-time recognition is not yet achieved and the dataset is limited, the preliminary experimental results are promising. Future work may focus on dataset expansion and model compression to enhance performance and broaden applications.

ACKNOWLEDGEMENTS

We appreciate National Science and Technology Council (NSTC) Taiwan for supporting grants by NSTC 113-2221-E-218-005.

REFERENCES

- Lê, Manh & Doan, Minh & Nguyen, Duy Dieu & Nguyen, Minh Son. (2022). Smart Desk in Hybrid Classroom: Detecting student's lack of concentration when studying. 13-18. 10.1109/NICS56915. 2022.10013468.
- K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 770778, doi:10.1109/CVPR.2016.90.keywords:{Training;Degradation;Complexitytheory;Imagerecognition;Neuralnetworks;Visualization;Image segmentation},
- Fenghour, Souheil & Chen, Daqing & Guo, Kun & Li, Bo & Xiao, Perry. (2021). Deep Learning-Based Automated Lip-Reading: A Survey. IEEE Access. 9. 121184 - 121205. 10.1109/ACCESS. 2021.3107946.