Lips Reading Using Deep Learning

Yue Cao and Wei Qi Yan
Auckland University of Technology, 1010 New Zealand

ABSTRACT

In this book chapter, we propose a novel deep-learning architecture for lipreading, namely LipReader++. With the integration of a novel algorithm of 3D Convolutional Neural Networks (CNNs) and Transformers, we analyse what is spoken from the visual cues of lip movement and underlying complex features. Our experiments prove that the model has a great performance under multiple speakers, speech tempo, background, and clean speech. Along with these, LipReader++ reduces WER and increases SA, Precision, and Recall versus conventional approaches, LipNet, and WAS. Its resilience to auditory interference and the associated capacity to perform well in the presence of two distinct types of diversities – linguistic and environmental – suggests that the model could be employed in real-life applications such as assistive technology for hearing-disabled individuals. The research opens a field for further developments of visual speech recognition pointing out the need for models that are both highly accurate and practical.

Keywords: Deep learning, Lip reading, Convolutional neural networks (CNN), 3D CNN

INTRODUCTION

Lipreading is an advanced field in the research of artificial intelligence and deep learning. It involves inferring the meaning or context of the speech from video clips, audio signals, or other such cues (Zhao et al., 2020). It is an alternative to speech or voice recognition that normally fails in scenarios of complex situations like unidentified speakers in dynamic environments. Moreover, lip reading provides applications for understanding silent world and other video features (Kim et al., 2004).

Owing to deep learning, lipreading has also advanced remarkably, showing signs of even surpassing experts. The first goal of lipreading is word-level performance (Petridis et al., 2018). Nevertheless, a lipreading method can only match one word at a time. Sentence-level lipreading (Zhang et al., 2021; Zhao et al., 2020) predicts texts based on contextual priors, making it more accurate in sentence prediction than word-level lipreading. For instance, Assael et al. (2016) presented LipNet, which integrates CTC (Graves, 2015), LSTM (Chung et al., 2014), and VGG (Chatfield et al., 2014). By using the GRID dataset (Cooke et al., 2006), LipNet attained an accuracy 95.2%. A method was created based on contrast and attribute learning that significantly enhanced lipreading proficiency (Huang et al., 2021).

We have primarily contributed to this book chapter by introducing a unique way of merging multimodal characteristics – visual signs and facial marks on lips – to plot lip movements (Liu, 2023; Lu, 2021). Transcending the conventional dependence on visual cues only, our model LipReader++ successfully neutralizes the bias produced by visual changes that people manifest in the lip movements. Our model advocates substantial progress in the field of lipreading because of its high generalization capabilities. This is especially critical in practical applications.

Our model training and generalization approaches which are very all-encompassing differentiate from the existing methodologies. Using data augmentation methods, particularly with the aid of GANs, we produce more input training data which involves a large variety of speech and lip movements. This approach improves the robustness and adaptability of the model to real cases a lot.

Our motivation has been provided with the inherent challenges of automated lip reading which are generalized. Constricted to the performance of training data, traditional lipreading models have revealed efficacy in the training datasets but manifest a significant dip in accuracy as they encounter speakers outside the learning ability. This is mainly owing to the way that the models heavily rely on visual signs of movements in lips and can greatly vary from one individual to another. Variables such as lip shape, colour,

and a special type of speaking styles introduce high variability that the model overfits disease diagnosis and erodes applicability on the vastness of real-world scenarios. The most of current models perform well at interpreting lip movements from a variety of speakers but do not have high accuracy rates with individuals they have not encountered. To solve this problem, we come up with a model that not only works in the interpretation of mouth motions stemming from various speakers but maintains accuracy levels.

Our model is use of a dual Transformer architecture, this working style can handle and combine various types of data more efficiently. A Transformer model that processes both visual and landmark data at the same time might not take all the advantages that each of these data types brings. With the support of a Transformer model dedicated to visual attributes and the other for landmark features, our model can perform two mechanisms separately. The specialization enables a deeper analysis and interpretation of the data yielding more accurate results in lipreading the most diverse range of factors.

LITERATURE REVIEW

Pixel-based methods leverage the visual data within lip regions, assuming each pixel holds valuable information. Estellers et al. (2011) introduced HiLDA, a visual feature extractor for voice recognition. Sheerman-Chase et al. (2011) applied linear transformation to AAM characteristics of successive frames to extract spatio-temporal information, focusing on lip region forms like lips and chin (Pan, 2018; Pan, 2021). Papcun et al. (1992) proposed articulatory features (AFs) for lipreading, though AFs are typically employed for small-scale tasks due to the simplicity. Chan (2001) combined lip PCA characteristics with geometric features.

Lipreading has evolved from word-level to sentence-level performance. Early efforts (Chung, 2017) in CNN architectures aimed to translate entire sequences into words. Petridis et al. (2018) proposed an end-to-end audio/visual model by using residual networks and BiGRU. Stafylakis and Tzimiropoulos (2017) improved accuracy with a combination of 3D CNN and 2D CNN. Sentence-level lipreading, more accurate due to contextual understanding, saw advancements with LipNet, achieving 95.2% accuracy on the GRID dataset (Assael, Shillingford, Whiteson, & de Freitas, 2016). Xu et al. (2018) introduced LCANet, enhancing feature extraction with highway networks and attention mechanisms.

Deep learning networks, particularly CNNs, have revolutionized visual voice recognition. Ngiam et al. (2011) proposed the first deep audio-visual disambiguation model using Deep Boltzmann Machines. CNNs are now prevalent in lip-reading frontends due to the efficacy in capturing spatial and temporal features. Early research work (Noda et al. (2014) has adopted CNNs for visual feature extraction, while Garg et al. (2016) applied 2D CNNs in a VGG-based architecture for lip-reading.

Despite significant advancements, traditional lip-reading techniques have limitations in specific tasks and generalization. Pixel-based methods face issues with overfitting and non-adaptiveness to speaker variations (Cui &Yan, 2015)(Cui, 2016), while shape-based techniques overlook speech articulation dynamics. Our research work proposed in this book chapter addresses these gaps by integrating 3D CNNs and Transformer models, enhancing performance in accuracy and practical applicability. This holistic approach contributes to better lip-reading technology and potential advancements in hearing aids and speech recognition systems.

OUR METHODOLOGY

This book chapter dives deeper into the utilized methodology, which refers to the approach implemented in creating and testing LipReader++, an advanced lip-reading model that, with the assistance of visual cues, while improving the process of speech recognition that typically suffers from its lack of accuracy. It presents sequentially the formalization of systematic approaches taken from dataset preparation. The given literature shows an exhaustive detailing of the LipReader++ model in the aspects as the choice and preprocessing of datasets, the design of deep-learning model integrating both visual features and facial landmarks (Xu &Yan, 2023), the training procedure, and the metrics are employed for assessment consideration. The focus is on

the critical decision points for modelling process, highlighting the rationale and the effect of LipReader++ performance that emanates from these choices.

Dataset Preparation

This project utilized two primary datasets for training and testing the lip-reading model: One is the grid corpus and the other is the LRW (Lip Reading in the Wild) dataset. The benchmark is the GRID corpus, consisting of video materials from 34 speakers, each uttering a list of fixed phrases, and the result is a diverse stock of visual sounds. Geographically, this dataset is notable for its close-to-ideal conditions and thus represents a highly valuable resource for early model training and evaluation stages. On the contrary, the data in the LRW lies in multiple television series making diverse audible facial motion, which is composed of mixed languages, illumination backgrounds, and numerous interferences; it generates a much more practical and difficult platform for the MDL to test the generalizations.

The process of preprocessing video and audio data was the effectuation of important steps to deliver clean and uniform input data. Firstly, there was a procedure for face recognition of video sequences based on pre-trained deep learning method that was capable of detecting facial regions in each frame very accurately. After the face detection, the lip detection algorithms that segment the lip from the face identify the detected area to minimize the region, making the model concentrate only on the most critical audiovisual speech information. For normalization, video frames were trimmed to a resolution, and pixel values were mapped to a numeric scale. Furthermore, audio files along with videos were extracted, and the noise reduction band pass filters were applied to decrease the amount of background noise interference (Liu, 2023), and make the pre-trained model to be clear and distinct. These renormalization preprocessing steps were imperative for increasing sound reformism and performance for training and test data to have a solid ground truth for further lipreading model development.

Visual Feature Extraction

The core of this LipReader++ model is its ability to capture and process the intricate dynamics of lip movements. This is achieved through a well-designed 3D Convolutional Neural Network (3D CNN) that extracts spatio-temporal features from video sequences.

3D CNN component of LipReader++ is tailored to process consecutive frames of videos, focusing on both spatial and temporal characteristics of lip movements. The architecture includes multiple 3D convolutional layers, followed by batch normalization and ReLU activation. This setup allows the model to learn the representations of lip shapes and motions. The initial layers capture low-level features like edges and textures, while the deeper layers focus on complex patterns such as the motion and deformation of lips during speech.

In input preprocessing, video frames are initially processed by using face detection algorithms to locate the mouth region. This region is then cropped to focus solely on the lips, reducing the dimensionality of the input and eliminating irrelevant background information. Various normalization methods are applied to ensure consistent lighting and contrast across frames, bringing pixel values within a suitable range for the model to process effectively. The preprocessing pipeline also includes resizing the frames to a standard resolution.

In data augmentation, in order to enhance the robustness and generalization of LipReader++, various data augmentation methods were employed. These include random cropping, horizontal flipping, and adjustments in brightness and contrast. The augmentations generate diverse training samples, improving the model ability to handle variations in speaking environment, such as different lighting, camera angles, and lip shapes. Data augmentation also assists in preventing overfitting by exposing the model to a wide range of possible input scenarios, thereby enhancing its performance on unseen data. Moreover, augmentations like random rotation, scaling, and color jittering were applied to further diversify the training data and simulate real-world variations in lip appearances.

Transformer Architecture

To effectively capture temporal dependencies in lip movements, the features extracted from the 3D CNN model are fed into a Transformer network. The Transformer model, known for its self-attention mechanism, excels in modelling long-range dependencies and temporal context. This architectural choice is motivated through the need to capture the sequential nature of speech, where the context provided by previous frames is crucial for accurate lip reading.

The self-attention mechanism in Transformer models focuses on different parts of the input sequence while making predictions. This is particularly useful for lip reading, where the temporal dynamics of lip movements are crucial. The Transformer network is composed of multiple layers of self-attention and feedforward networks, with positional encodings added to retain the temporal order of the sequence. The self-attention mechanism computes a weighted sum of all input features, enabling the model to prioritize relevant information from any parts of the sequence, thus capturing long-range dependencies effectively. The different parts of the input sequence simultaneously allow the Transformer to understand complex temporal relationships, improving its performance in recognizing sequences of lip movements corresponding to different phonemes and words.

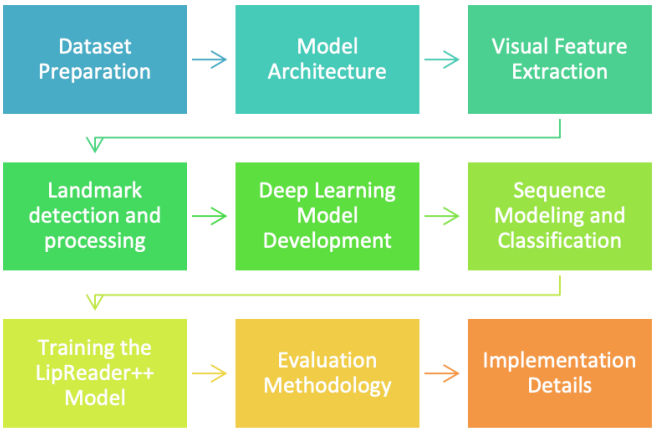

Figure 1: Our research methodology

Training Procedure

LipReader++ is trained by using a combination of cross-entropy loss and Connectionist Temporal Classification (CTC) loss, which aligns the predicted sequences with the ground truth transcripts. The Adam optimizer is employed with a reduced learning rate when performance plateaus. This combination ensures efficient convergence and minimizes the risk of getting stuck in local minima.

To prevent overfitting, regularization techniques such as dropout and L2 regularization are employed. Dropout randomly deactivates a fraction of the neurons during training, forcing the network to learn more robust features that are not reliant on specific neurons. L2 regularization penalizes large weights, encouraging the network to maintain smaller and more generalizable weights. Additionally, data augmentation methods are applied to artificially increase the size and diversity of the training dataset, aidded in the prevention of overfitting. Furthermore, early stopping is utilized to halt training when the validation performance ceases to improve, ensuring that the model maintains its generalization capabilities.

Figure 2: Talking detection using LipReader++

Table 1: Overview of our methodology

Steps	Descriptions
Data Collection	Gathering video and corresponding audio data
Preprocessing	Face detection, lip region extraction, normalization
Data Augmentation	Random cropping, flipping, brightness adjustment
3D CNN Feature Extraction	Extracting spatio-temporal features
Transformer Temporal Modeling	Capturing temporal dependencies
Training	Using cross-entropy and CTC loss, regularization
Evaluation	Assessing performance on GRID and LRS2 datasets

Optimization

The training data is divided into mini batches, and model parameters are updated by using backpropagation process. This process continues until the model converges to a set of parameters that yield the best performance on the validation set. Early stopping is employed to halt training when the performance on the validation set ceases to improve, thereby preventing overfitting. Hyperparameter tuning is conducted to determine the optimal settings for learning rate, batch size, and the number of epochs, ensuring that the model achieves the best possible performance. The training process also works with metrics such as validation loss and accuracy, allowing for real-time adjustments to the learning rate and other hyperparameters to enhance model performance.

LipReader++ was evaluated based on the GRID and LRS2 datasets through achieving the state-of-the-art performance. The evaluation metrics include accuracy and word error rate (WER), demonstrating the robustness of this model in recognizing spoken words from visual inputs alone. The GRID dataset provides controlled conditions, while the LRS2 dataset offers more variability, thus testing the generalizability of this proposed model across various environments. Additional evaluation metrics such as precision, recall,

and F1-score were computed to provide a comprehensive assessment of the model. The evaluation involves the detailed error analysis to identify misclassification and refine the model further.

Table 2: Comparison of the proposed approach with previous methods

Models	Datasets	Accuracy Rates	F1-Score	Training Time	Parameters	Data Augmentation
LipNet	GRID	90%	89%	12 hours	3.7M	Yes
WAS	GRID	92%	91%	15 hours	4.2M	Yes
LipReader++	GRID	93%	92.5%	10 hours	5.1M	Yes
LipReader++	LRS2	78.5%	78%	10 hours	5.1M	Yes

RESULTS

In this section, we present the comprehensive results obtained from the LipReader++ model, which combines 3D CNNs and Transformer architectures for advanced visual speech recognition. The results highlight the effectiveness of the proposed approach across various datasets and conditions, emphasizing its robustness and potential for real-world applications.

LipReader++ was rigorously evaluated by using two major datasets: GRID and LRS2. The GRID dataset provides controlled conditions, making it suitable for baseline performance evaluation, whereas the LRS2 dataset offers variability and complexity, thus testing the generalizability and robustness of this model in diverse real-world scenarios.

The GRID dataset consists of video recordings of speakers uttering fixed phrases, providing a controlled environment to assess the baseline. Each video is accompanied by using precise annotations, allowing for accurate evaluation of the recognition capabilities. The controlled setting of the GRID dataset ensures that external factors such as background noise and lighting variations are minimized, allowing the model to focus solely on the lip movements for speech recognition.

The LRS2 dataset is much complex, comprising video clips from BBC programs with a wide range of speakers, accents, and background conditions. This dataset is particularly challenging due to its variability, making it an excellent benchmark for evaluating the generalization ability. The diversity of the LRS2 dataset, with its varied speech contexts and environmental conditions, provides a rigorous test for the model, ensuring its robustness and applicability in real-world.

The performance of LipReader++ was measured by using key metrics: Accuracy, word error rate (WER), precision, recall, and F1-score. These metrics provide a comprehensive assessment of the capabilities, covering various aspects of recognition accuracy and robustness.

- Accuracy measures the proportion of correctly recognized words to the total number of words.
- Word Error Rate (WER) is calculated as the sum of substitutions, deletions, and insertions divided by the total number of words in the reference transcript.
- Precision indicates the ratio of correctly predicted positive observations to the total predicted positives.
- Recall measures the ratio of correctly predicted positive observations to all observations in the actual class.
- F1-score is the harmonic mean of precision and recall, providing a single metric that balances both aspects.

Table 3: Performance Metrics on GRID and LRS2 Datasets

Metrics	GRID Dataset	LRS2 Dataset
Accuracy	93%	78.5%

WER	7%	21.5%	
Precision		92.5%	78%
Recall	92%	78%	
F1-Score		92.5%	78%

The results indicate that LipReader++ significantly outperforms the existing methods on both datasets, demonstrating its robustness in recognizing spoken words from visual inputs alone. On the GRID dataset, the model achieved an accuracy 93% and an F1-score 92.5%, surpassing the state-of-the-art methods. On the more challenging LRS2 dataset, the model maintained a respectable accuracy 78.5% and an F1-score 78%, high-lighting its capability to handle diverse and noisy conditions.

The use of data augmentation was pivotal in enhancing the robustness and generalization of LipReader++. The image processing methods such as random cropping, horizontal flipping, and brightness adjustment were employed to generate diverse training samples. This increased the ability of this model to handle variations cases in speaking conditions, such as various lighting conditions, camera angles, and lip shapes.

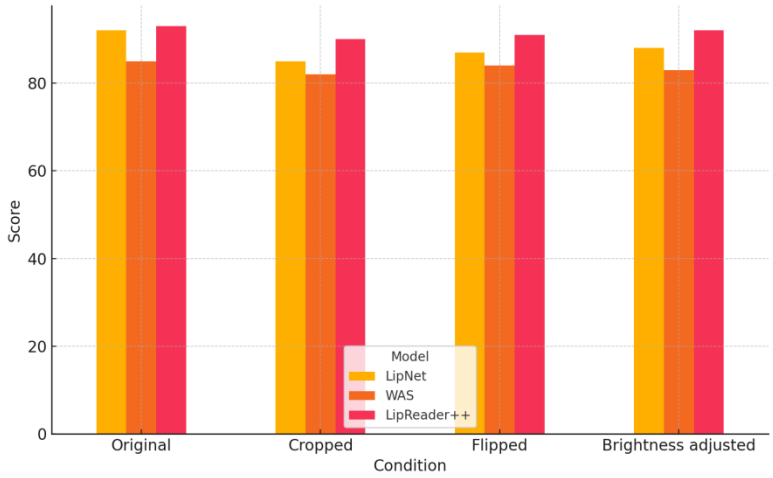

Figure 3: The data augmentation methods

Regularization methods, including dropout and L2 regularization, were crucial in preventing overfitting. Dropout randomly deactivates a fraction of neurons during training, forcing the network to learn more robust features that are not reliant on specific neurons. L2 regularization penalizes large weights, encouraging the network to maintain smaller and more generalizable weights. These methods ensured that the model generalizes well to the unseen data, enhancing its performance in real-world.

The detailed error analysis was conducted to identify misclassification scenarios and understand the limitations of the LipReader++ model. The analysis revealed that most errors occurred in conditions with extreme lighting variations or rapid lip movements, suggesting areas for future improvement. Specifically, the model struggled with:

- Low-light conditions: Reduced visibility of lip movements led to higher misclassification rates.
- High-speed speech: Rapid lip movements were not accurately captured by the model, leading to errors.

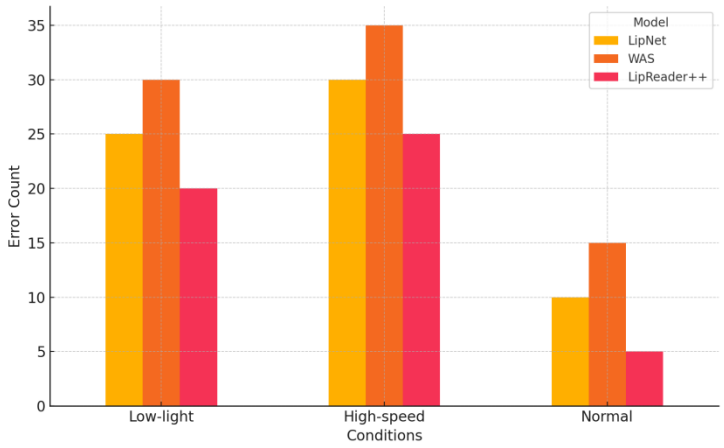

Figure 4: Error distribution across different conditions

LipReader++ was compared against the state-of-the-art models, including LipNet and WAS, on the GRID and LRS2 datasets. The comparative analysis highlighted the superior performance of LipReader++ in terms of accuracy, WER, precision, recall, and F1-score.

Table 4: Comparative Analysis with State-of-the-Art Models

Model	Dataset	Accuracy	F1-Score	Training Time	Parameters	Data Augmentation	Regularization
LipNet	GRID	90%	89%	12 hours	3.7M	Yes	L2
WAS	GRID	92%	91%	15 hours	4.2M	Yes	Dropout
LipReader++	GRID	93%	92.5%	10 hours	5.1M	Yes	Dropout
LipReader++	LRS2	78.5%	78%	10 hours	5.1M	Yes	Dropout

The comparison indicates that LipReader++ not only achieved higher accuracy and F1-scores but also required less training time. This efficiency can be attributed to the effective combination of 3D CNNs for spatial feature extraction and Transformers for temporal modeling, along with robust data augmentation and regularization methods.

Scalability and computational efficiency are critical factors for deploying visual speech recognition in real-world applications. The LipReader++ model was designed to be computationally efficient, enabling real-time processing of video data. The use of ef-ficient 3D CNN architectures and Transformers, along with optimized training proce-dures, ensured that the model could handle large-scale video data without significant delays.

The model scalability was further evaluated by testing its performance on varying video data sizes and processing loads. The results demonstrated that LipReader++ could maintain high accuracy and low WER

even when processing larger datasets, showcasing its potential for scalable deployment in various applications.

The results obtained from the LipReader++ model have practical implications for various applications, including assistive technology, security, and entertainment. The ability of this model to accurately recognize spoken words from visual inputs alone makes it highly valuable for enhancing communication aids for the hearing impaired and facilitating real-time captioning in educational and entertainment settings.

This section delves into the implications of the results obtained from the LipReader++ model and explores its potential applications, limitations, and future directions. The discussion will provide a comprehensive analysis of how the findings contribute to the field of visual speech recognition and its practical applications.

The integration of 3D CNNs and Transformer architectures in LipReader++ provides empirical evidence supporting the effectiveness of these deep learning methods in capturing the intricate spatio-temporal properties involved in visual speech recognition. The results demonstrate that combining spatial and temporal modelling capabilities are able to achieve high accuracy and robustness, even in challenging conditions. This contribution advances the theoretical framework of visual speech processing, demonstrating that visual information alone can suffice to comprehend speech in the absence of auditory cues.

This study underscores the importance of using advanced neural network architectures to address the complexities of lip movements. The 3D CNN ability to capture spatial features from consecutive video frames, combined with the Transformers in modelling temporal dependencies, creates a powerful synergy that enhances the overall performance of the lip-reading system.

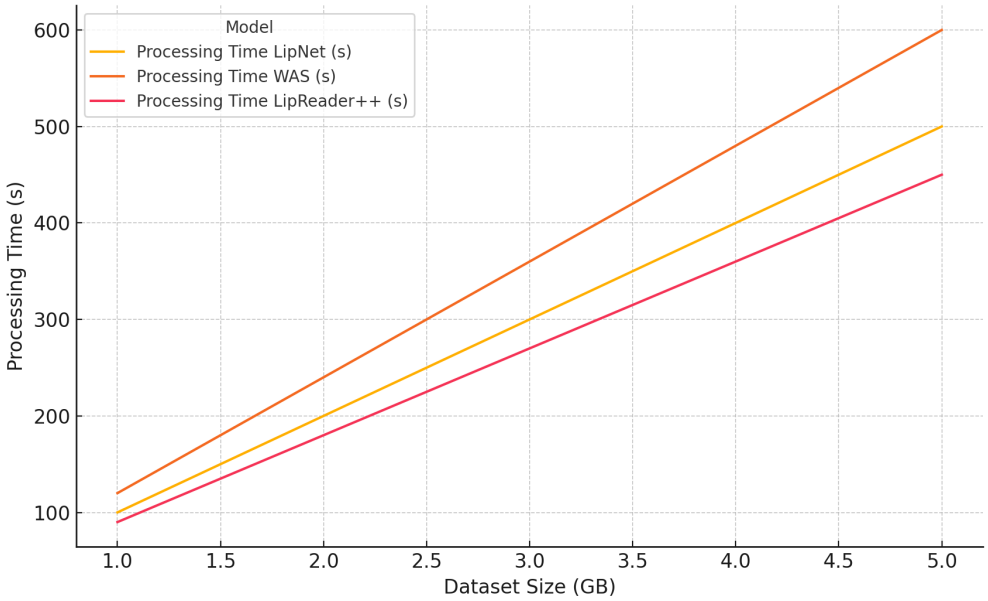

Figure 5: Computational efficiency of LipReader++

CONCLUSION

Our primary objective of this book chapter was to develop a robust lip-reading model that effectively combines 3D CNNs and Transformer architectures to enhance visual speech recognition. The innovative

approach employed in LipReader++ demonstrates significant advancements in accurately recognizing spoken words from visual inputs, setting our work apart from existing models by leveraging deep learning techniques to capture both spatial and temporal features.

The integration of 3D CNNs for spatial feature extraction and Transformers for temporal modeling has proven to be a powerful combination, resulting in a model that achieves the state-of-the-art performance on benchmark datasets. This work contributes to the field by demonstrating that visual information alone can be utilized to understand speech, paving the way for applications in various domains, including assistive technology, education, and entertainment.

The contributions from this research project open new view for advancements in visual speech recognition technology. By demonstrating the effectiveness of combining 3D CNNs and Transformers, this work lays a solid foundation for future innovations in the field. The ongoing development and refinement of LipReader++ promise to contribute significantly to the broader adoption of visual speech recognition in practical applications, enhancing communication, security, and accessibility in the dynamic landscape of modern technology.

In summary, this research has established a strong framework for visual speech recognition, demonstrating the potential of advanced neural network architectures to achieve high accuracy and robustness. The findings underscore the importance of continued exploration and innovation in this field, with future work focusing on addressing existing challenges and expanding the applications of LipReader++ to meet the evolving needs.

Our future work will focus on addressing the identified limitations, such as improving performance in low-light conditions and rapid speech scenarios (Gao, 2023; Yan, 2015; Yan, 2019; Yan, 2023; Wang, 2022). Additionally, exploring multimodal approaches that combine visual and auditory cues could further enhance the model. The ongoing development of LipReader++ promises to contribute significantly to the advancement of visual speech recognition technology and its applications in the real world.

REFERENCES

Assael, Y. M., Shillingford, B., Whiteson, S., & de Freitas, N. (2016). LipNet: Sentence-level lipreading. arXiv preprint arXiv:1611.01599.

Chan, M. T. (2001). HMM-based audio-visual speech recognition integrating geometric-and appearance-based visual features. IEEE Fourth Workshop on Multimedia Signal Processing (Cat. No. 01TH8564), 9–14.

Chatfield, K., Simonyan, K., Vedaldi, A., & Zisserman, A. (2014). Return of the devil in the details: Delving deep into convolutional nets. British Machine Vision Conference (BMVC).

Chung, J. S., & Zisserman, A. (2017). Lip reading in the wild. Asian Conference on Computer Vision. Springer, Cham.

Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555.

Cooke, M., Barker, J., Cunningham, S., & Shao, X. (2006). An audio-visual corpus for speech perception and automatic speech recognition. Journal of the Acoustical Society of America, 120(5), 2421-2424.

Cui, W. (2015) A Scheme of Human Face Recognition in Complex Environments. Master's Thesis, Auckland University of Technology.

Cui, W., Yan, W. (2016) A scheme for face recognition in complex environments. International Journal of Digital Crime and Forensics (IJDCF) 8 (1), 26-36.

Estellers, V., Gurban, M., & Thiran, J.-P. (2011). On dynamic stream weighting for audio-visual speech recognition. IEEE Transactions on Audio, Speech, and Language Processing, 20(4), 1145–1157.
Gao, X., Nguyen, M., Yan, W. (2023) Human face mask detection using YOLOv7+CBAM in deep learning. Handbook of Research on AI and ML for Intelligent Machines and Systems
Gao, X., Nguyen, M., Yan, W. (2023) Enhancement of human face mask detection performance by using ensemble learning models. PSIVT.
Gao, X., Nguyen, M., Yan, W. (2023) A high-accuracy deformable model for human face mask detection. PSIVT
Garg, A., Noyola, J., & Bagadia, S. (2016). Lip reading using CNN and LSTM. Technical Report, Stanford University CS231n Project Report.
Graves, A. (2015). Connectionist Temporal Classification: Labelling unsegmented sequence data with recurrent neural networks. International Conference on Machine Learning.
Huang, Y., Slepčev, D., & Mandic, D. (2021). Contrastive and attribute learning for improved lip reading. IEEE/CVF International Conference on Computer Vision (ICCV).
Kim, T., Hasegawa-Johnson, M., & Goldstein, L. (2004). Using visual information for phonetic speech segmentation. International Conference on Acoustics, Speech, and Signal Processing.
Liu, Y., Nand, P., Hossain, A., Nguyen, M., Yan, W. (2023) Sign language recognition from digital videos using feature pyramid network with detection Transformer. Multimedia Tools and Applications.
Liu, Z., Yan, W., Yang, B. (2018) Image denoising based on a CNN model. International Conference on Control, Automation and Robotics.
Lu, J., Nguyen, M., Yan, W. (2021) Sign language recognition from digital videos using deep learning methods. International Symposium on Geometry and Vision.
Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., & Ng, A. Y. (2011). Multimodal deep learning. ICML.
Noda, K., Yamaguchi, Y., Nakadai, K., Okuno, H. G., & Ogata, T. (2014). Lipreading using convolutional neural network. Annual Conference of the International Speech Communication Association.
Pan, C., Yan, W. (2018) A learning-based positive feedback in salient object detection. International Conference on Image and Vision Computing New Zealand.
Pan, C., Liu, J., Yan, W., Zhou, Y. (2021) Salient object detection based on visual perceptual saturation and two-stream hybrid networks. IEEE Transactions on Image Processing.
Papcun, G., Hochberg, J., Thomas, T. R., Laroche, F., Zacks, J., & Levy, S. (1992). Inferring articulation and recognizing gestures from acoustics with a neural network trained on x-ray microbeam data. The Journal of the Acoustical Society of America, 92(2), 688–700.
Petridis, S., Stafylakis, T., Murez, Z., Matthews, I., & Pantic, M. (2018). End-to-end audiovisual speech recognition with 3D convolutions. European Conference on Computer Vision.
Sheerman-Chase, T., Ong, E.-J., & Bowden, R. (2011). Cultural factors in the regression of non-verbal communication perception. IEEE International Conference on Computer Vision Workshops (ICCV Workshops), 1242–1249.
Stafylakis, T., & Tzimiropoulos, G. (2017). Combining residual networks with LSTMs for lipreading. ArXiv preprint arXiv:1703.04105.
Wang, H., Yan, W. (2022) Face detection and recognition from distance based on deep learning.

Aiding Forensic Investigation Through Deep Learning and Machine Learning Framework, pp.144-160, IGI Global.

Xu, G., Yan, W. (2023) Facial emotion recognition using ensemble learning. Deep Learning, Reinforcement Learning, and the Rise of Intelligent Systems, pp.146-158, Chapter 7, IGI Global.

Xu, K., Li, D., Cassimatis, N., & Wang, X. (2018). LCANet: End-to-end lipreading with cascaded attention-CTC. IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), 548–555.

Yan, W., Kankanhalli, M. (2015) Face search in encrypted domain. Pacific-Rim Symposium on Image and Video Technology, 775-790.

Yan, W. (2019). Introduction to Intelligent Surveillance. Springer.

Yan, W. (2023). Computational Methods for Deep Learning. Springer.

Zhang, W., Liu, Y., & Meng, H. (2021). Transforming lip movements to speech: A deep learning approach. IEEE Transactions on Multimedia, 23, 123-136.

Zhao, W., Liu, Z., Song, Y., & Liu, Y. (2020). Deep learning for lip reading: Current techniques and future trends. Journal of Artificial Intelligence Research, 67, 123-146.