JAIST Repository
https://dspace.jaist.ac.jp/

Title	Study on Visual Speech Recognition Based on Multi-Region Information
Author(s)	曾, 鵬程
Citation	
Issue Date	2024-12
Type	Thesis or Dissertation
Text version	author
URL	http://hdl.handle.net/10119/19415
Rights	
Description	Supervisor: 吉高 淳夫, 先端科学研究科, 修士(情報科学)

JAIST
JAPAN
ADVANCED INSTITUTE OF
SCIENCE AND TECHNOLOGY

Japan Advanced Institute of Science and Technology

Master's Thesis

Study on Visual Speech Recognition Based on Multi-Region Information

ZENG Pengcheng

Supervisor: Atsuo Yoshitaka

Japan Advanced Institute of Science and Technology

Graduate School of Advanced Science and Technology

Information Science

December 2024

Abstract

Visual Speech Recognition (VSR) is a technology that recognizes and interprets spoken language by analyzing facial and lip movements in video. Its primary goal is to decode language content using visual cues, which is particularly valuable when audio information is limited or absent. VSR has made significant progress, with the current mainstream approach focusing on extracting lip features and using deep learning to enable the model to understand a speaker's content from video alone. However, one might question whether visual language recognition is synonymous with lipreading. Can we extract additional information beyond lip movements to improve model performance? In this study, by developing the Lip-Face-Surrounding model to comprehensively extract information from videos. This model supports three input channels, utilizes 3DCNN for feature extraction, and applies a CTC layer to align the extracted features with the text sequence. 3D Convolutional Neural Networks (3DCNN) excel at extracting spatial and temporal features, making them well-suited for visual speech recognition tasks. By employing 3DCNN, the model captures dynamic changes across facial, lip, and surrounding cues within video sequences. The Connectionist Temporal Classification (CTC) layer effectively addresses alignment, allowing extracted features to align with the target text sequence without requiring predefined alignment, thus enhancing the model's capability to handle variable-length input. The findings show that direct information beyond the lips—such as eye corner movements, jaw movements, nostril movements, throat movements, and shoulder movements—are captured by the model and serve as discriminative features for visual speech recognition. This direct information is also applicable to handcrafted datasets. Additionally, indirect information such as the speaker's body language and interactions with the surrounding can impact model

performance. Sometimes, this information provides extra context that enhances performance, while other times, it introduces noise that affects model convergence. This model achieved promising results on the CN-CELEB and GRID datasets, with a 5% absolute performance improvement over the lip-only approach.

CONTENTS

Figure Contents

Table Contents

I. INTRODUCTION

1.1 Background

This section primarily introduces the background of Visual Speech Recognition (VSR) and the existing challenges in the field. A detailed description of current research approaches will be provided in the related work section.

1.1.1 Visual Speech Recognition

The primary goal of Visual Speech Recognition (VSR) is to recognize language based on visual cues, particularly those extracted from the speaker's facial movements. These cues include lip movements, facial expressions, and occasionally even head or neck movements. VSR models typically use only video signals as input, but some multimodal language recognition studies incorporate both clean and noisy audio along with video signals to verify the role of visual information in enhancing audio recognition [1]. Unlike traditional lip reading, which focuses solely on interpreting spoken words through lip movements, VSR encompasses a broader range of visual cues, including facial expressions, eye movements, and other contextual elements that aid in understanding speech. This broader focus allows VSR to capture more comprehensive visual information, enhancing the accuracy and robustness of the recognition process.

For humans, language communication is inherently multimodal, as we often rely on visual information to comprehend spoken language. VSR models simulate this multimodal process by incorporating visual cues into language recognition, which significantly enhances the robustness of recognition, especially in high-noise environments. VSR also provides an essential tool for understanding spoken content without sound, benefiting people with hearing impairments and demonstrating its potential application in diverse scenarios. By integrating multimodal information,

VSR not only improves recognition accuracy but also contributes to the diversified development of language recognition technology.

1.1.2 Visual Speech Recognition and Lip-reading

Lip reading is a technique used to understand spoken language by observing the movements of the speaker's lips. Human lip reading primarily relies on noticing lip shapes, mouth openings, and facial expressions to assist in understanding speech, especially in situations where hearing is impaired or audio quality is poor. Lip reading compensates for auditory limitations and serves as an essential form of non-verbal communication.

In the field of computer vision research, distinctions between visual language recognition and lip reading are not always clearly defined. The primary difference between Visual Speech Recognition (VSR) and traditional lip reading lies in the range of visual information utilized. Lip reading typically focuses exclusively on lip movements to interpret spoken words, while VSR extends beyond lip movements to incorporate additional visual cues, such as facial expressions, eye movements, and other contextual signals. This broader scope allows VSR to capture more comprehensive visual information, enhancing the accuracy and robustness of recognition models. However, this expansion also adds complexity to model development, as it necessitates sophisticated techniques for effectively extracting and integrating diverse visual cues [2].

1.1.3 Auto Visual Speech Recognition

HMM-Based Automatic Lip Reading: The earliest computerized lip-reading technologies relied on Hidden Markov Models (HMM), using statistical methods to analyze the relationship between lip movements and phonemes. While HMM-based approaches performed well in small-scale, controlled scenarios, they suffered from

poor scalability, making it challenging to adapt to diverse speakers, complex language environments, and real-world applications [3].

Neural Network-Based Automatic Lip Reading: With advancements in deep learning, neural networks—especially Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN)—became widely used in the field of automatic lip reading. This approach allowed for reading simple sentences in controlled environments, achieving higher recognition accuracy than HMM-based methods. However, it still had significant limitations. The speaker must maintain a certain posture, and the spoken content is usually restricted to specific phrases or words, presenting challenges for real-world applications [4].

Automatic Lip Reading Using "Wild" Datasets: In recent years, progress in VSR has been supported by large-scale, labeled datasets collected from natural, unstructured environments ("wild" datasets). While these datasets have enabled significant improvements in model performance, creating such complex datasets is resource-intensive and costly, limiting the scalability of these approaches. Moreover, research indicates that error-prone words in audio-visual datasets often follow certain patterns, suggesting a need for alternative methods to capture nuanced information present at the edges of samples [5] [6] [7]. Current models generally focus on extracting the speaker's lip movements from videos, often with limited attention to other critical regions of interest. Most studies directly use pre-cropped images of the lips, omitting broader visual contexts that could improve recognition accuracy. This lip-centric approach can lead to ambiguities, particularly with homophones or similarly articulated words. Integrating broader contextual information, such as facial expressions and head movements, may help disambiguate these cases, though doing so introduces further challenges in feature extraction and fusion.

1.1.4 VSR from an Anthropological Perspective

The process of "speaking" in humans is not limited to lip movements alone; it involves a coordinated effort among multiple muscle groups, including those in the jaw, throat, and tongue. Zhang et al. [8] demonstrated that capturing the entire face, instead of just the lips, substantially improves model recognition accuracy. Building on this insight, further inquiries is: could visual information beyond the face—such as the surrounding environment, the speaker's attire, emotions, and body movements—enhance the effectiveness of recognition models? If so, capturing more comprehensive visual information might contribute to higher accuracy.

Relevant psychological research also aligns with this perspective. Liu et al. [9] found that language choice varies by context; formal reports often use more academic language, while casual conversations are typically more colloquial. Mohammad et al. [10] noted that a speaker's emotional state is closely linked to their speech patterns. By integrating as much directly or indirectly relevant visual information as possible, enhance model performance, leveraging these nuances to more effectively interpret spoken language.

1.2 Problem / Objectives / Contributions / Research Significance

Problem Statement: Despite advancements in VSR, the field faces several significant challenges. First, variability in speakers' facial features, expressions, and movements creates substantial obstacles to achieving consistent recognition accuracy. This variability is further complicated by external factors such as lighting conditions, camera angles, and ambient noise. Additionally, even when speaking the same word, lip movements differ widely between individuals, making it difficult for models to generalize across various speakers. Furthermore, occlusions, such as hands or objects covering the face, and rapid head movements can lead to the loss

of critical visual information, hindering the model's ability to accurately interpret speech.

Objectives: The most important task in this study is to develop a four-channel input network that separates inputs into four categories: lips, face, surrounding environment, and voice. By comparing the model's performance with individual inputs and various combinations, assess the impact of each input on the VSR system. The goal is to improve the accuracy of visual speech recognition models by integrating additional visual information beyond just the lips.

Contributions: By developing a lip-face-surrounding model to capture information from different video regions, including the speaker's lip movements, facial expressions, and the surrounding environment. This approach enables the model to utilize multiple visual cues, enhancing overall performance. Validated the model's effectiveness on two datasets: GRID [11] (a clean, controlled environment dataset) and CN-CELEB [5] (a complex, real-world dataset). Results demonstrate that incorporating face and environmental context positively impacts visual language recognition. Specifically, on the GRID dataset, the configuration with lips + face + surrounding achieved a word error rate (WER) that was 1% lower than with lips alone, with an absolute WER of 2.3%. On the CN-CELEB dataset, although the introduction of surrounding information reduced model stability, it improved overall performance. The average character error rate (CER) for single-speaker tasks was 37.04%, a 5% improvement over the lip-only baseline, while multi-speaker tasks saw an average CER of 49.31%, a 3% improvement.

By evaluating the impact of each input on the VSR system, provide insights into selecting key regions of interest (RoI) for visual language recognition tasks. Although incorporating environmental context can, in some cases, reduce model stability, it generally improves overall performance. Also conducted an initial

investigation into the causes of reduced stability, laying a foundation for future research in this area.

Research Significance: VSR systems can be applied in various scenarios where high-quality audio is challenging to capture, such as identifying athletes' speech during sporting events, enhancing broadcast quality. Additionally, this research seeks to identify the most critical aspects of human communication beyond language, providing guidance for training programs for individuals with hearing impairments and fostering better communication within this community.

1.3 Structure

This thesis is organized into five chapters. Chapter 1 provides an overview of the background of visual language recognition and a summary of this work. Chapter 2 covers related work, including datasets, models, and research on regions of interest (RoI) in VSR. Analyzed some of the unexplored issues in these studies and their relevance to this research. Chapter 3 details this model, including feature extraction, feature analysis, and methods for multimodal alignment. Chapter 4 presents these experiments and results, comparing outcomes under various input conditions and focusing on the influence of environmental information on results. Chapter 5 summarizes the experiment, offers recommendations on leveraging environmental information to improve model accuracy, highlights the limitations of this study, and provides an outline of potential directions for future research.

II. RELATED WORK

2.1 Dataset

This study will introduction begin with datasets, as all research in automatic visual speech recognition (AVSR) is directly dependent on the dataset used. For instance, a model that performs well on the GRID dataset, which consists of simple sentences, may not achieve similar results on the LRS dataset (a complex English sentence dataset, discussed later). Researchers often adjust models according to the focus of their tasks, demonstrating that the structure of a dataset and the quality of labeled data directly influence the quality of the entire research study.

Firstly, certain datasets used in previous studies are no longer valuable for modern VSR research. For example, the Tulips dataset [10] includes only one speaker who randomly utters numbers between 0 and 4 in English; the JR dataset [3] contains a single speaker saying one of ten random Japanese subway station names; and the CUAVE dataset [12] includes one speaker randomly reciting a series of phone numbers in English. These datasets are characterized by their simplicity and limited data volume, making them insufficient for complex AVSR tasks today.

Next, will introduce two relatively complex datasets created in controlled lab environments. These datasets are characterized by simple sentence structures and well-annotated content. They are typically seen as transitional datasets from simple lip-reading tasks to more complex ones. Sometimes, they are also used as pre-training datasets or as evaluation sets to assess model performance.

GRID: This dataset includes 32,746 usable videos from 34 speakers, featuring a fixed sentence structure that comprises commands (4 types), colors (4 types), prepositions (4 types), letters (25 types), digits (10 types), and adverbs (4 types). The sentence structure follows this format: command (e.g., bin, lay, place, set) + color

(e.g., blue, green, red, white) + preposition (e.g., at, by, in, with) + letter (e.g., a, b, c, ..., z excluding w) + digit (0, 1, 2, ..., 9) + adverb (e.g., please, soon, now, again). For example, a sample sentence in GRID could be, “bin blue at f2 please.” In this study, finally selected GRID as the evaluation set.

OuluVS2 [13]: This dataset features input from multiple angles, making it a valuable resource for research into optimal lip-reading angles. Researchers use OuluVS2 to explore the impact of different visual perspectives on lip-reading accuracy.

Finally, will introduce modern VSR datasets. Modern visual speech recognition datasets feature more complex sentences and scenes, typically collected from various online platforms. These videos include interactions between speakers and audiences, diverse facial expressions (e.g., silence, confusion, happiness), and non-frontal facial orientations, offering a richer range of visual cues.

CN-CELEB: Fan et al. developed CN-CELEB, a large audiovisual dataset specifically focused on Chinese. This dataset is currently the largest Chinese field dataset, containing over 130,000 utterances with a total duration exceeding 240 hours. CN-CELEB is categorized into “news” and “speech” sections and is fully annotated manually to ensure high quality. The utterances in CN-CELEB do not follow a fixed format and are sourced from platforms such as Bilibili. For example, an utterance from CN-CELEB could be, “ta jiu shi zhe me yi ge sui xing de ren.” In this experiment, primarily used CN-CELEB for training, tuning, and evaluating the model.

Studies based on this dataset have shown that previously well-performing models experience a decrease in performance on more complex datasets, highlighting the need to enhance the robustness of audiovisual language recognition models [5].

2.2 Automated Visual Speech Recognition

In the early days, many lipreading models were based on Hidden Markov Models (HMM). These models typically relied on small datasets, had limited task capabilities, and suffered from poor scalability. For instance, Movellan's [10] model could only recognize five digits (0-4), Sugahara et al.'s [3] model recognized ten subway station names, and Dupont et al.'s [12] model recognized phone numbers. While these HMM-based models often achieved near-perfect accuracy for their specific tasks, their applicability was limited to narrowly defined scenarios, rendering them inadequate for the demands of modern, complex visual language recognition.

Modern research in visual language recognition primarily focuses on the following aspects to enhance model performance:

1) **Building more complex datasets for model training**: Chung et al. developed three outdoor lip-reading datasets—LRW [14], LRS [15], and LRS2 [16]—specifically designed to train visual language recognition models. Ma et al. generated new datasets by augmenting old datasets, resulting in a substantial increase in data volume. These efforts have been validated, showing that the inclusion of additional training data indeed enhances model performance.

2) **Developing more complex networks to handle increasingly complex datasets**: Assael et al. [4] developed LipNet, a network using a 3D CNN [21] to extract lip movement features, followed by a fully connected layer processed by CTC to handle alignment issues. LipNet achieved a 5% word error rate (WER) on the GRID dataset, a significant improvement over earlier HMM-based VSR models for GRID, which had error rates close to 50%. Chung et al. developed a multimodal language recognition model that emphasizes the

impact of visual signals on multimodal language recognition in noisy environments. Their model extracts feature separately from the lips and audio using CNNs, then processes these features with a transformer. This model achieved a 13.9% word error rate (WER) on the LRS dataset. With video input alone (lips only), the WER rose to 50.2%, compared to a WER of 73.8% for human experts. However, their model underperformed compared to the CTC/Attention model by Ma et al. [17], which achieved a WER of 15.2% on LRS. The CTC/Attention model is more advanced, combining the strengths of CTC for handling audio-text alignment with attention for multimodal alignment. This model eliminates the LSTM network, which is less effective with long sequences, and leverages additional training data. It is worth noting, however, that all of these studies have focused solely on cropped images of the lips.

While collecting large and complex datasets has proven an effective method for enhancing VSR capabilities, this approach has notable drawbacks: it is time-consuming, resource-intensive, and costly. Furthermore, as FAN et al. [5] observed, even with extensive datasets, certain challenging words maintain high error rates, suggesting unresolved edge cases in current models. This indicates that simply expanding datasets may yield diminishing returns in accuracy, especially for easily confused words. Therefore, exploring alternative feature extraction methods to address these challenges more effectively is necessary. Although HMM-based models excelled in specific tasks, their limited applicability in broader, more complex scenarios reduces their relevance to modern visual language recognition needs [18].

2.3 Regions of Interest (RoI)

In the fields of psychology and medicine, related research has shown that when performing audiovisual language recognition, humans do not limit their focus to the lips. Bateson et al. [15] also demonstrated that the movement of the jaw and eye muscles positively impacts visual speech recognition. Furthermore, perceivers tend to focus on the speaker's lips, eyes, and body movements, and this tendency becomes more pronounced in noisy environments. Cvejic et al. [16] studied the movement patterns of all relevant head muscles during speech, emphasizing that eyebrow raises are often synchronized with changes in speech prosody. This rhythmic consistency is particularly important for visual language recognition tasks.

Research on regions of interest (ROIs) in the field of visual recognition remains relatively limited. Intuitively, neural networks could possess a capacity similar to that of humans for capturing information from facial expressions, body language, and environmental context. In other words, isolating only lip movement data may overlook the speaker's expressions and surrounding contextual cues. Currently, the size of the cropping frame is often determined by researchers' intuition or experience. Zhang et al. [8] demonstrated that movements of the jaw and eye muscles also positively impact visual speech recognition. By cropping the entire face, they achieved an absolute improvement of 2% over the baseline system on the GRID dataset. They also noted that while additional information can boost performance, it may also introduce noise. Therefore, it is crucial to carefully incorporate useful information while preserving clean data (lip movement), meaning that any new information should prompt a corresponding model update. Chung et al. [16] explored the feasibility of side-view lipreading, using the OuluVS2 dataset, which captures speakers from multiple angles, enabling training

across various viewpoints. They showed that lipreading from a slightly off-center frontal angle yields the best results; however, this difference is minimal and is unlikely to revolutionize current visual language models.

III. ARCHITECTURE

In this chapter, models will be provided with a detailed description of the lip-face-surrounding model. First, will introduce the overall structure of the model. Next, will discuss each component in sequence: the 3D CNN used for feature extraction, the fully connected layers, and the CTC for handling features. Additionally, will explain this approach for cropping the lip, face, and surrounding regions to standardize the selection of regions of interest.

3.1 Lip-Face-Surrounding Architecture

This work is based on LipNet, which employs a 3D Convolutional Neural Network (3DCNN) combined with CTC [19] to extract and process features. To enable the model to capture more information from the video (rather than limiting it to the lip area alone), extended the original LipNet to support three inputs: lip, face, and surrounding. This design serves two purposes: first, to retain the cleanest information (the lip region), as simply expanding the cropped area would significantly increase computational load, and do not want to compromise the original lip performance. Second, by changing inputs and observing the resulting performance, can validate which specific input proves most effective, thereby gaining insights into the areas the model should focus on. chose 3D CNN over LSTM or Transformer for the following two reasons:

1) Sentences in modern lipreading datasets can be particularly long, which would lead to a significant drop in LSTM's computational efficiency. Due to its recursive structure, LSTM requires step-by-step information transmission, and as the sequence length increases, training time can extend considerably. In contrast, CNN typically has higher computational efficiency because convolution operations are parallelized rather than performed recursively.

2) Based on real-time requirements, prefer that the model doesn't always wait to view the entire sentence before making a judgment. Instead of prioritizing the full sequence, emphasize local information, which made us opt not to use Transformer. Although this may somewhat affect model performance, 3D CNN is sufficient for this study's needs.

This model has three input channels of size 120×120×3 each, designated for LIP, FACE, and Surrounding information, respectively. The goal of standardizing the input size across these three image channels is to reduce the influence of the LIP region on the FACE and Surrounding layers. By resizing the FACE and Surrounding layers, the LIP becomes blurred in these two layers, which aligns with this intended focus. Specifically, the FACE layer should emphasize facial expressions, moods, and emotions, while the Surrounding layer should prioritize body language and the surrounding context.

For feature extraction, employ three layers of 3D CNNs with max-pooling. This network is not particularly deep; selected this structure to avoid overfitting and to balance efficiency and performance. After several trials, determined that a three-layer configuration was a relatively optimal choice.

In the feature processing step, the extracted features are flattened into a 1D vector and passed through a fully connected layer. This layer provides an initial output that has not yet addressed alignment, which is then inputted to the CTC layer to produce the final text output with the most likely alignment. Figure 1 illustrates the architecture of this model, Figure 2 shows the workflow, and Table 1 provides a detailed breakdown of this model's structure.

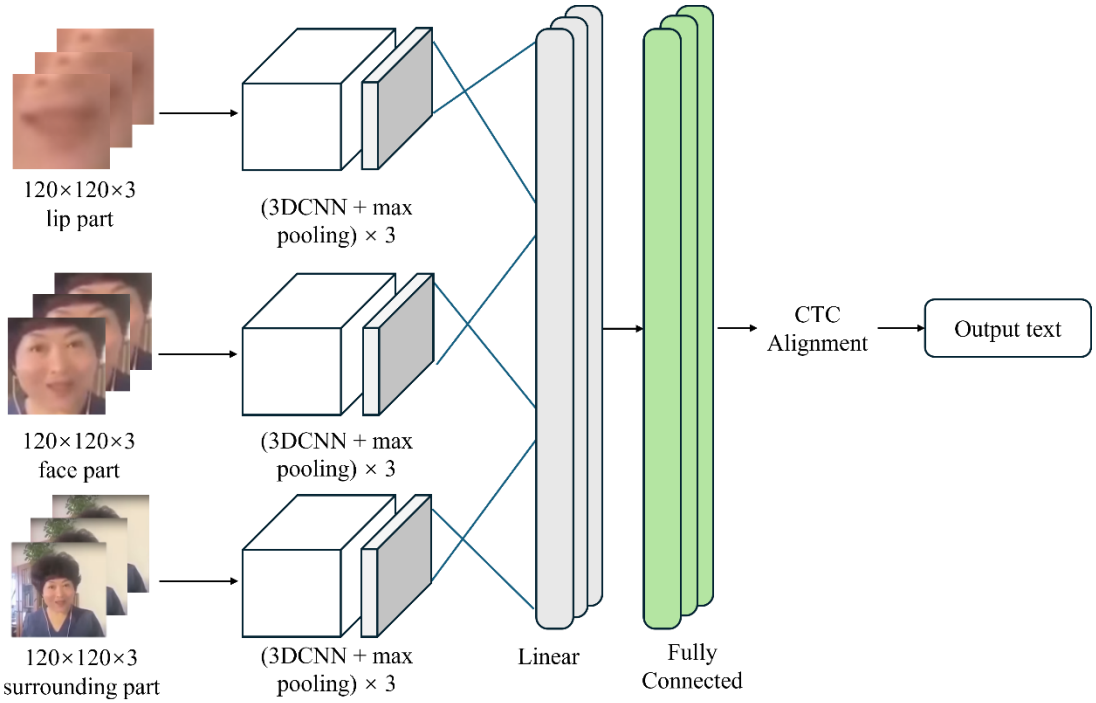

Fig 1 Architecture of Lip-Face-Surrounding Model

Table 1 Lip-Face-Surrounding Model details

Layer	Size/Stride/Pad	Input size	Dimension order
3DCNN1	3×6×6/1×2×2/1,2,2	75×3×120×120	T×C×H×W
Pool1	1×2×2/1×2×2	75×32×60×60	T×C×H×W
3DCNN2	3×6×6/1×2×2/1,2,2	75×32×30×30	T×C×H×W
Pool2	1×3×3/1×3×3	75×64×15×15	T×C×H×W
3DCNN3	3×3×3/1×2×2/1,1,1	75×64×5×5	T×C×H×W
Pool3	1×3×3/1×3×3	75×96×3×3	T×C×H×W
Linear		75×96×1×1	T×C×H×W

Layer	Input size	Output size	Dimension order
Linear	75×(96×3)1×1	75×288	T×C×M×H×W
Fully Connected	75×288	75×4469	T×Words
CTC	75×4469 Ground truth	CTC-loss	

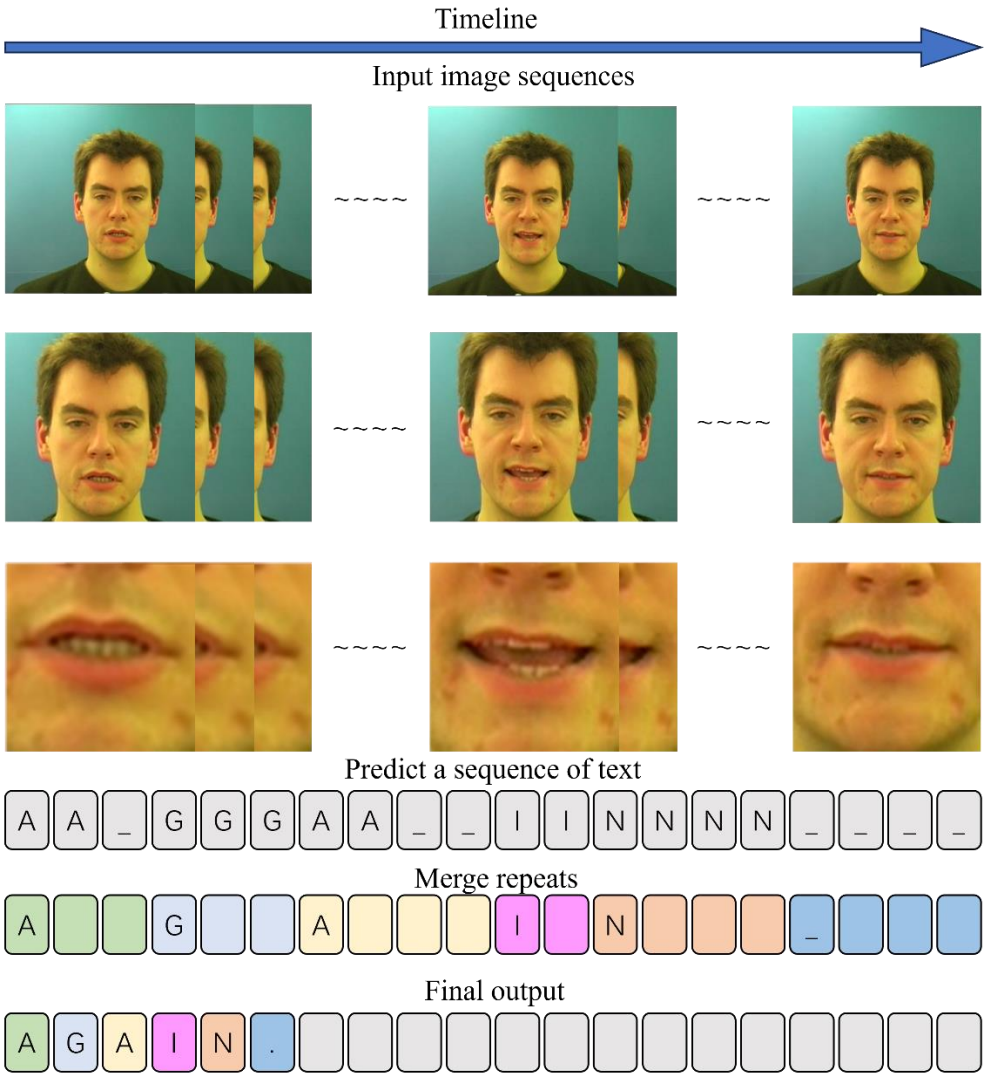

Fig 2 Work of Lip-Face-Surrounding Model

3.2 RoI Cropping

The regions of interest (ROIs) are a focal point of this study. In this section, will detail the elements included in each ROI.

First, for the initial video input that does not contain audio signals, crop it into 256 × 256, 3-channel images, organized into a sequence of images at a frame rate of 25 frames per second. (The reason for not directly cropping to 120×120 is that will later enlarge the lip layer while simultaneously reducing the face and surrounding layers

to minimize the influence of the lip region in both the face and surrounding layers.) For the Surrounding layer, directly resize it to a 120×120 image sequence. This layer includes a somewhat unclear image of the lips, a slightly clearer image of the face, and the surrounding environment. aim for the Surrounding layer to focus on the subject's body language and the surrounding context.

In another branch, the original image sequence is processed through the FAN detector, which is a pretrained model based on ResNet50 that extracts 68 facial features. Among these, the first 27 points represent the boundary information of the face. use the x_{max}, x_{min}, y_{max}, and y_{min} of these 27 points as reference boundaries, extending outward by 2 pixels to define the final boundaries of the Face layer. This layer is then normalized into a 120×120 image sequence, which includes a relatively unclear image of the lips along with the entire face. want the model to focus on the deformation occurring in the facial muscles.

Finally, the Lip layer is determined by the 49th to 60th points outputted by the FAN detector. Similarly, take the x_{max}, x_{min}, y_{max}, and y_{min} of these 11 points as reference boundaries and extend outward by 2 pixels to establish the final boundaries of the Lip layer. This layer is then normalized into a 120×120 image sequence (which typically involves enlargement, though in rare cases it may involve reduction). This layer includes only the image of the lips, serving as the benchmark for assessing the impact of the other two inputs on model performance. Figure 3 illustrates the process of cropping the RoIs.

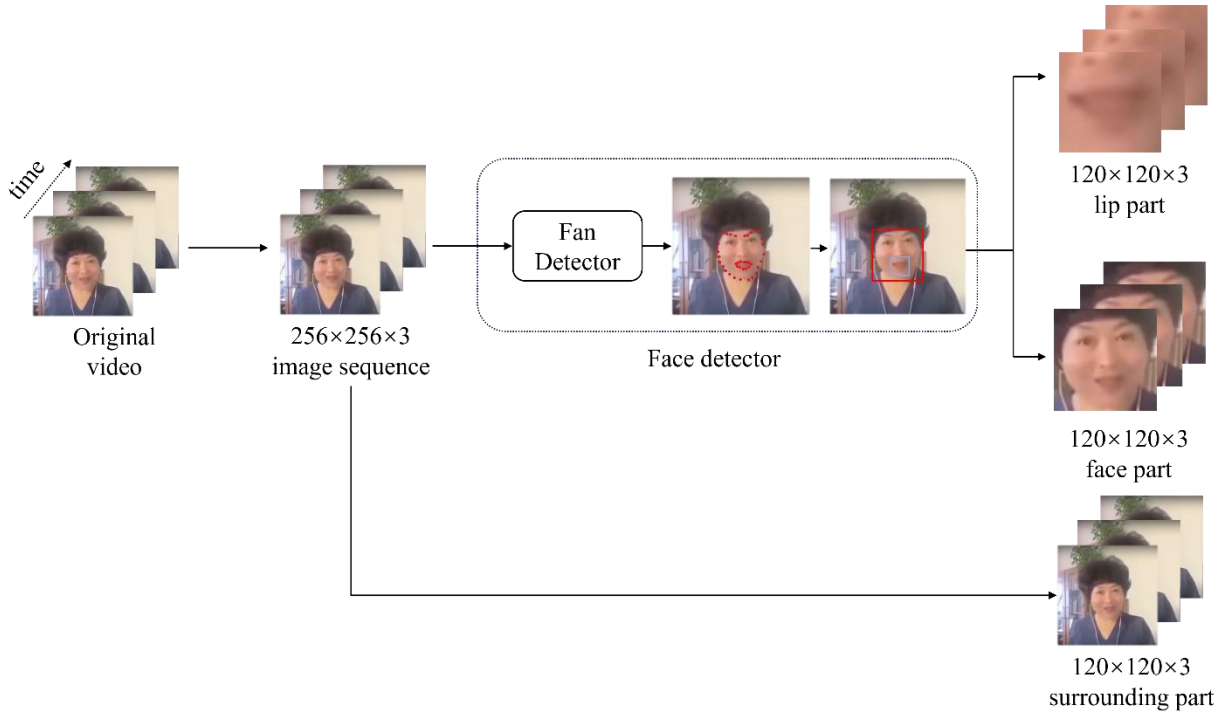

Fig 3 the process of cropping the RoIs.

IV. EXPERIMENTS AND RESULTS

This section provides a detailed overview of the training details and results of the Lip-Face-Surrounding model, including the training process setup, parameters used, and the model's performance across different datasets. have evaluated the model's performance under various inputs, including single-input comparisons of lip only, face only, and surrounding only. Additionally, compare multi-input performance using lip only as a baseline against combinations such as lip + face, lip + surrounding, and lip + face + surrounding. also present a performance comparison of this model on the GRID dataset with other datasets. Results indicate that this model outperforms the majority of existing lipreading models and even surpasses many models that utilize additional training data.

4.1 Training

This network was trained on two datasets, GRID and CN-CELEB, with the objective of comparing model performance in controlled (laboratory) and field environments, as well as exploring the potential for cross-linguistic visual language recognition. The network was implemented using PyTorch and trained on a single NVIDIA GeForce RTX 3080 GPU with 12GB of VRAM. In the fully connected layer, set the learning rate to 0.002, weights to 0.03, and biases to zero. No data augmentation or additional training data was used.

For the CN-CELEB dataset, evaluated the model’s Character Error Rate (CER) with single-input configurations: lip only, face only, and surrounding only. selected Ma et al.’s [17] results as a baseline, as it is a multimodal network that includes audio input and has a similar structure to these with lip-only input, making it suitable for assessing this model’s overall performance. also evaluated CER across four different input configurations: lip only, lip + face, lip + surrounding, and

lip + face + surrounding. Nine models were trained using cross-validation, with average CER, standard deviation, and the best CER achieved in these experiments reported.

For the GRID dataset, followed the same procedure, evaluating Word Error Rate (WER) under single-input configurations as well as composite-input configurations. As a baseline, chose Assael et al.'s [4] results, which is a network that supports only lip image sequence input and provides useful reference for single-input structure comparisons.

Training on the CN-CELEB dataset took 19 days, while training on the GRID dataset took one night. WER and CER are defined as follows:

$$WER = \frac{W_{\text{insertions}} + W_{\text{substitutions}} + W_{\text{deletions}}}{W_{\text{total words}}} \times 100\%$$

$$CER = \frac{C_{\text{insertions}} + C_{\text{substitutions}} + C_{\text{deletions}}}{C_{\text{total characters}}} \times 100\%$$

4.2 Result

Table 2 and *Table 3* show composite-input and single-input results for the LFS model on the CN-CELEB dataset, respectively. *Table 4* and *Table 5* show composite-input and single-input results for the LFS model on the GRID dataset. *Table 6* presents a comparison of this model with state-of-the-art VSR research on the GRID dataset.

Figures 4 and 5 illustrate the convergence of the LFS model on the CN-CELEB dataset for single-input and multi-input settings, respectively. Figures 6 and 7 illustrate the convergence of the LFS model on the GRID dataset for single-input and composite-input settings, respectively.

Table 2: multi-input Results on CN-CELEB. "Average CER in Test Set" represents the average CER and the corresponding standard deviation from 9 tests. "Best CER in Test Set" indicates the lowest (best) CER. "L" stands for "Lip," "F" stands for "Face," and "S" stands for "Surround."

Region	**Average CER in Test Set**	**Best CER in Test Set**	**CER in Evaluation Set**
Lip only	44.5±0.7	43.8	43.7
L + F	39.1±0.3	38.9	39.3
L + S	40.4±1.1	38.8	40.3
L + F + S	37.0±1.1	36.5	36.5
Baseline	N/A	41.2	39.6

Table 3: Single-input Results on CN-CELEB.

Region	**Average CER in Test Set**	**Best CER in Test Set**	**CER in Evaluation Set**
Lip only	44.5±0.7	43.8	43.7
Face only	40.2±0.3	38.9	39.6
Surrounding	40.4±2.6	38.8	40.3
Baseline	N/A	41.2	39.6

On the CN-CELEB dataset, observe an instability in prediction results introduced by the *surrounding* input. In the *surrounding-only* input mode, fluctuations reach ±2.6%, which is significantly higher than in other input modes. will discuss the potential causes of this phenomenon in the following analysis.

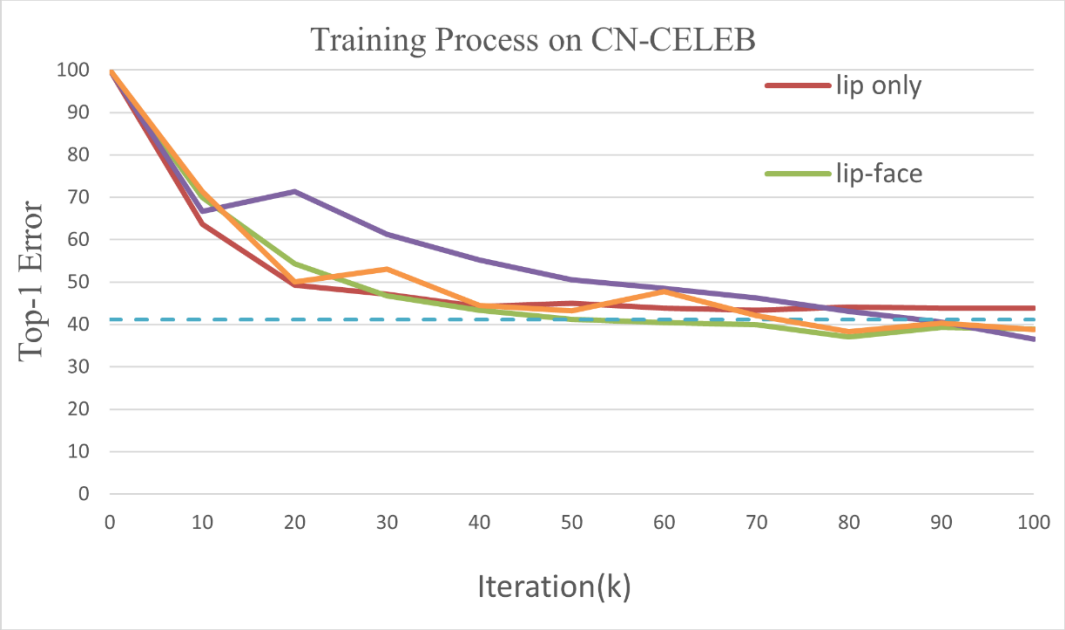

Fig 4 Training process in CN-CELEB(Multi-input)

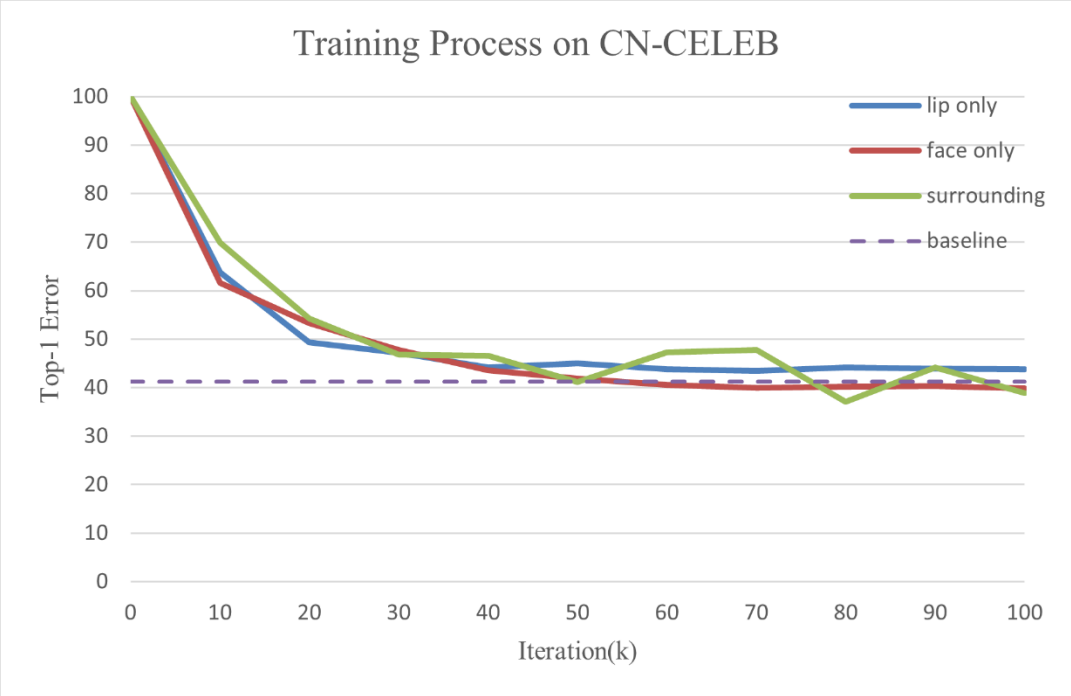

Fig 5 Training process in CN-CELEB(Single-input)

In the CN-CELEB dataset, while this performance in the *lip only* mode does not surpass that of the baseline system, several *multi-input* configurations do outperform the baseline. This supports this first conclusion: information beyond the lip region aids VSR systems in interpreting language. Furthermore, the best performance is achieved with the *lip-face-surrounding* input configuration, aligning with these expectations. Compared to the face layer alone, the surrounding layer captures additional cues, such as throat movements, which contribute to the model's decision-making.

When comparing face only and *surrounding only* modes, the final performance after convergence is similar. However, the *surrounding only* mode exhibits less stability during training than *face only*. believe this is because, while the surrounding information generally aids model training, it also introduces a variety of elements, which can sometimes act as noise—especially with speakers who use frequent gestures or are in complex environments. In these cases, the extra details from the surrounding layer can disrupt convergence. Overall, while the surrounding layer is effective, its stability is lower than that of the face layer alone. Finally, both *face only* and *surrounding* configurations outperform the baseline system, further reinforcing that information beyond the lip region is beneficial for VSR accuracy.

Table 4: multi-input Results on GRID

Region	**Average WER in Test Set**	**Best WER in Test Set**	**WER in Evaluation Set**
Lip only	5.2±0.4	4.9	5.4
L + F	3.8±0.3	3.5	3.9
L + S	3.9±0.2	3.7	3.7
L + F + S	3.5±0.2	3.4	3.2
Baseline	N/A	4.6	N/A

Table 5: single-input Results on GRID

Region	Average WER in Test Set	Best WER in Test Set	WER in Evaluation Set
Lip only	5.2±0.4	4.9	5.4
Face only	3.8±0.4	3.4	3.5
Surrounding only	3.9±0.5	3.5	4.0
Baseline	N/A	4.6	N/A

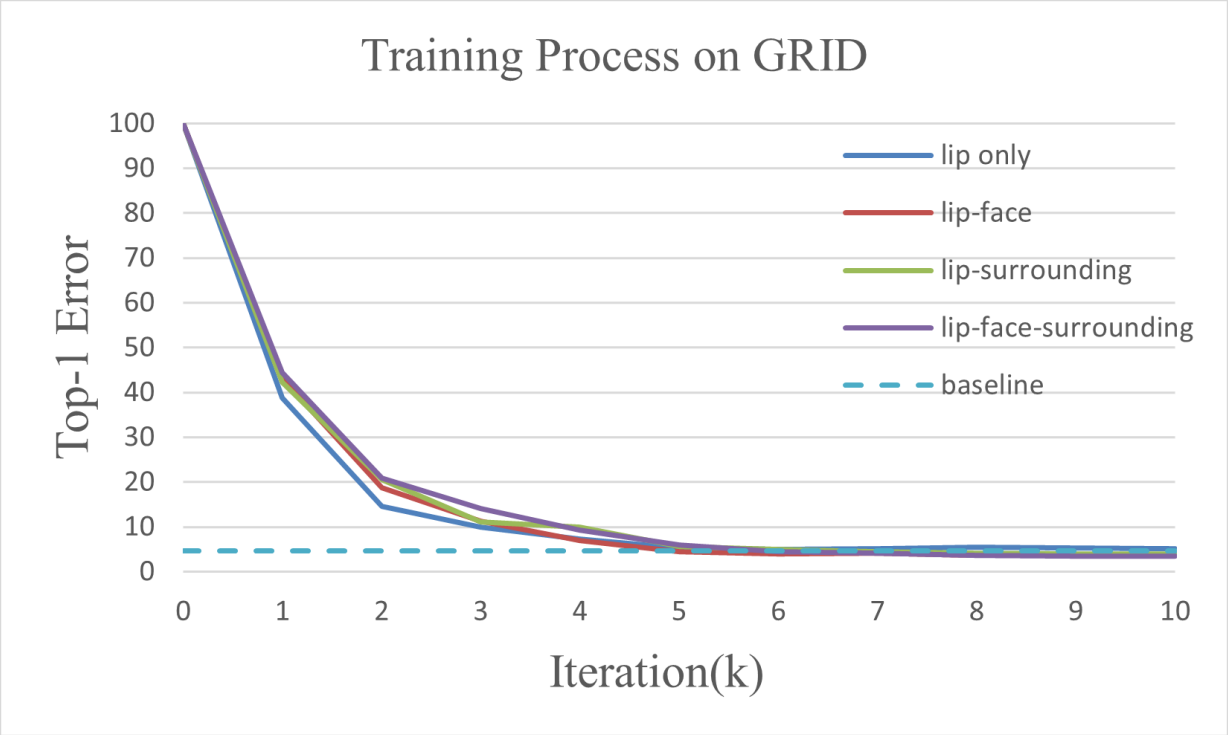

Fig 6 Training process in GRID(Multi-input)

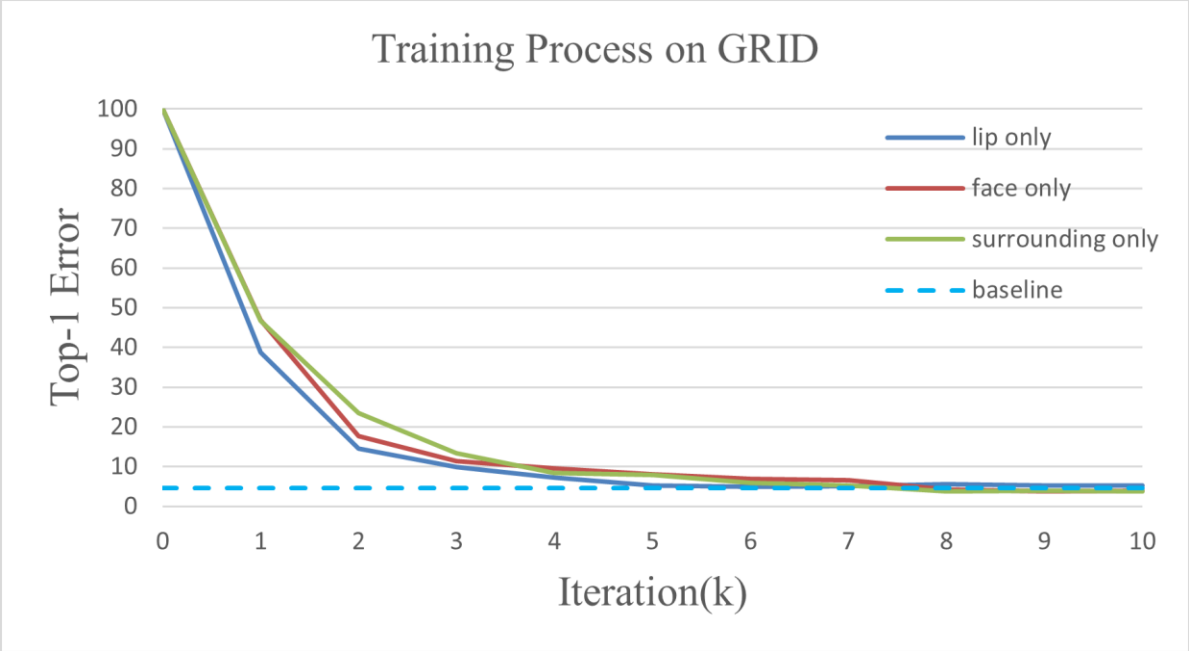

Fig 7 Training process in GRID(Single-input)

GRID is a dataset recorded in a controlled laboratory environment, without gestures or environmental variations. This purpose in using this dataset is to verify this hypothesis in reverse: in a dataset lacking surrounding information, including the surrounding layer should not impact model accuracy. According to the results, while the *lip only* mode underperforms relative to the baseline, this model achieves performance gains when either *face* or *surrounding* information is included. Throughout the training process, the surrounding input introduced minimal fluctuations and converged with high stability.

Table 6 summarizes a performance comparison between this model and other state-of-the-art VSR models, with data sourced from Paper with Code. Since some experiments utilized additional training data, have marked this variable for clarity and fairness. As shown, this model achieves performance on par with the top-tier models in the field.

Table 6: Comparison of Results on GRID with Other Models

Method	Year	Extra Training Data	Best WER
LipNet	2016	×	4.6
LFS(Ours)	2024	×	3.2
WAS	2016	√	3.0
LipNet-face	2020	×	2.9
LCANET	2018	×	2.9
CTC/Attention	2022	√	1.2

Here is a comparison of this model's performance on CN-CELEB and GRID datasets, revealing some interesting insights:

1. On CN-CELEB, the model's stability was lowest when only the surrounding input was used, with a slight improvement in the lip-surrounding configuration. This suggests that the information provided by the lip input is relatively clean, facilitating model convergence.

2. The surrounding input, which caused fluctuations on CN-CELEB, did not introduce similar instability on GRID. On one hand, this is due to GRID's controlled lab setting, which lacks environmental and body language cues. On the other hand, it indicates that this model attempts to extract meaningful information from surrounding cues. However, in the absence of robust labels, this did not yield the anticipated improvements.

3. A key point on GRID is that although the surrounding input lacks environmental and body language information, it captures details missed by the face input alone—specifically, throat movement. were pleased to observe that this contributed positively to model accuracy. This suggests that even a full-face crop may not encompass all relevant "lipreading" cues,

indicating that there is still much to explore regarding optimal Regions of Interest (RoIs) in visual speech recognition research.

4.3 Visualization of Model Attention Areas

conducted occlusion experiments to verify the regions the model focuses on, which in turn guides the selection of Regions of Interest (RoI). This approach is based on the method described in [20], with a key difference: instead of manually selecting occlusion areas, allow the occlusion to sequentially traverse the entire image. Since are working with image sequences, calculate the relative performance drop for each time step's image. The issue of positional shifts due to the speaker's movement or camera movement was disregarded in this study. For the cropped image sequences, a black square mask was used to cover portions of the image, and the relative drop in model performance was calculated. This process generates a saliency map, which is then superimposed on the corresponding image. The occlusion size was set to 8x8 with a stride of 4.

conducted occlusion experiments on selected data from both the GRID and CN-CELEB datasets. Due to technical limitations, dynamic tracking of the occluded region (e.g., keeping the occlusion fixed on the eyes as the subject moves) is currently not possible. Therefore, opted for a fixed occlusion area in this experiment. While this method has certain limitations, it still provides valuable preliminary data to help understand the importance of different regions in visual language recognition tasks. Additionally, given the computational intensity of this method, selected a small number of objects of interest for experimentation. Specifically, these include some outdoor speech videos from the CN-CELEB dataset and videos from the GRID dataset.

The GRID dataset's primary advantage is the relatively consistent positioning of individuals within the frame, which allows for the estimation of the specific regions impacting the model's performance even when occlusion does not move in sync with the speaker. The results indicated that lip movements had the most significant effect on model performance, particularly for words like "please" and "lay," which involve substantial lip movement. Conversely, for words like "soon" and "two," which involve minimal lip movement, the model increased the feature weights for nasal and throat movements. Additionally, observed that the model occasionally captured shoulder movements, despite their lower weight, suggesting that the model may be picking up on specific speaker habits. This could potentially lead to overfitting but also demonstrates the model's ability to capture and effectively utilize information beyond just facial features.

On the CN-CELEB dataset, the model responded to both the speaker's body language and environmental changes. hypothesize that the use of multi-head attention mechanisms mitigated the impact of redundant information, preventing significant fluctuations in model performance. For a relatively simple dataset like GRID, the model can converge quickly, but for a more complex dataset like CN-CELEB, less epoch of training might be insufficient. Furthermore, issues such as occlusion not moving with the speaker and camera movements exacerbated the difficulty of convergence in the surrounding layer. It can be concluded that the regions receiving the highest attention are the lips, followed by the corners of the eyes, the wings of the nose, the throat, and the shoulders. Notably, the corners of the eyes and the wings of the nose cannot be captured in the lip layer, while the throat and shoulders are beyond the reach of the face layer.

Figure 8(a), (b), (c) shows examples of the occlusion experiment's results on the GRID datasets. From top to bottom, the words being spoken by the speaker are "please," "soon," "lay," and "two".

Fig 8(a) Example of the effects of occlusion in the GRID dataset (SOON)

Fig 9(b) Example of the effects of occlusion in the GRID dataset (LAY)

Fig 10(c) Example of the effects of occlusion in the GRID dataset (TWO)

V. CONCLUSION

In this study, this primary contribution is demonstrating the effectiveness of incorporating information beyond just lip movements in visual language recognition tasks. By utilizing the Lip-Face-Surrounding model, achieved a 5% absolute improvement in performance over the baseline system. This result underscores the value of including facial expressions and surrounding context alongside lip movements.

Regarding the selection of RoI in visual language recognition, these experiments confirm that the lips are the most critical area. Additionally, movements around the corners of the eyes, the wings of the nose, the throat, and the shoulders also provide valuable information for recognition. Overall, the key information is primarily concentrated on the speaker's body. Therefore, recommend cropping the entire speaker and focusing particularly on the lips, as this approach can significantly enhance the performance of visual language recognition models.

Further analysis of the impact of surrounding factors on the performance of visual language recognition models reveals some intriguing findings. In the CN-CELEB dataset, the model indeed captured various surrounding cues, such as the speaker's body movements and the surrounding crowd's reactions. While anticipated that these elements would influence the recognition process, there were instances where they actually led to a decline in model performance. hypothesize that the model sometimes picked up additional contextual clues, which may enhance performance, but in other cases, the introduction of noise disrupted the model's convergence, resulting in reduced accuracy. Although cannot fully elucidate the underlying logic at this point, it is clear that visual language recognition models are sensitive to

surrounding information. The challenge moving forward will be determining how to effectively gather and apply this data in complex real-world scenarios.

Considerations on the application and ethical aspects of visual language recognition technology. One technology related to visual language recognition is generating realistic lip movements during speech. This research indicates that, in addition to lip movements, speaking also involves changes in the throat and facial muscles. The resulting micro-expressions are not perfectly generated, and this research can guide studies related to speech generation by improving the realism of generated expressions. On the other hand, technologies that can identify a speaker's content from video alone may lead to privacy infringements and generating more realistic speech animations could result in the misuse of technology in ways that violate laws.

VI. ACKNOWLEDGMENTS

My advisor provided numerous valuable suggestions for this paper, which I have sincerely accepted. However, I do not fully agree with his suggestion to remove all subjects from sentences, though I have followed his guidance.

Throughout my Master's journey, there have been so many people who helped me along the way that it's impossible to list them all. Here, I want to express my deepest gratitude to three individuals. Given the importance of this acknowledgement, I would like to convey it in English and Japanese.

First, I am especially grateful to my secondary supervisor, Professor Fumihiko Asano. He taught me to examine the world through mathematical and analytical lenses. Rather than simply completing a project, this perspective has expanded my outlook on life itself, offering a more scientific way of thinking that has profoundly influenced my approach to my primary research topic.

The second person I would like to thank is my Ph.D. senior, Youming Fan. Without him delivering food to me, I might have starved to death up on the mountain—JAIST's cafeteria is really hard to bear.

The third is my former girlfriend, Yutong Zhuang. Even a faint star can light the way forward. I am truly grateful for her support.

Finally, I must also mention Professor Yoshitaka, who has always been very kind. His warm approach means there's no one specific memory, just an overall sense of gratitude.

Reference

[1] G. Potamianos, C. Neti, J. Luettin and I. Matthews, "Audio-Visual Automatic Speech Recognition: An Overview," *Issues in visual and audio-visual speech processing*, vol. 22, no. 23, 2004.

[2] A. Torfi, S. M. Iranmanesh, N. Nasrabadi and J. Dawson, "3D Convolutional Neural Networks for Cross Audio-Visual Matching Recognition," *IEEE Access*, vol. 7, pp. 22081-22091, 2017.

[3] K. Sugahara, T. Shinchi, M. Kishino and R. Konishi, "Real-time lip reading system on personal computer," *Transactions of the Society of Instrument and Control Engineers*, vol. 36, no. 12, pp. 1145-1151, 2000.

[4] Y. M. Assael, B. Shillingford, S. Whiteson and et al., "Lipnet: End-to-end sentence-level lipreading," 2016. https://doi.org/10.48550/arXiv.1611.01599

[5] Y. Fan, J. Kang, L. Li, K. Li, H. Chen, S. Cheng, P. Zhang, Z. Zhou, Y. Cai and D. Wang, "Cn-celeb: a challenging chinese speaker recognition dataset," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 39-51, 2020.

[6] L. Li, R. Liu, J. Kang, Y. Fan, Y. Cai, R. Vipperla, T. F. Zheng and D. Wang, "Cn-celeb: multi-genre speaker recognition," *Speech Communication*, vol. 137, pp. 77-91, 2022.

[7] P. Mapetridis and M. PanticS, "Visual speech recognition for multiple languages in the wild," *Nature Machine Intelligence*, vol. 4, pp. 930-939, 2022.

[8] Y. Zhang, S. Yang, J. Xiao, S. Shan and X. Chen, "Can We Read Speech Beyond the Lips? Rethinking RoI Selection for Deep Visual Speech Recognition," in *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, Buenos Aires, Argentina, pp. 1193-1211, 2020.

[9] J. Liu and L. Han, "A corpus-based environmental academic word list building and its validity test," *English for Specific Purposes*, vol. 39, pp. 1-11, 2015.

[10] S. Mohammad and P. Turney, "Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon," in *NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pp. 145-151, 2010.

[11] M. Cooke, J. Barker, S. Cunningham and X. Shao, "n audio-visual corpus for speech perception and automatic speech recognition," *The Journal of the Acoustical Society of America,* vol. 120, no. 5, pp. 2421-2424, 2006.

[12] E. Patterson and et al., "CUAVE: A new audio-visual database for multimodal human-computer interface research," in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 2353-2360, 2002.

[13] I. Anina, Z. Zhou, G. Zhao and M. Pietikainen, "OuluVS2: A multi-view audiovisual database for non-rigid mouth motion analysis," in *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pp. 1-11, 2015.

[14] J. S. Chung and A. Zisserman, "Lip reading in the wild," in *Computer Vision – ACCV 2016: 13th Asian Conference on Computer Vision, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part II*, Springer International Publishing, pp. 87-103, 2017.

[15] J. S. S. Chung, A. Senior, O. Vinyals and et al., "Lip reading sentences in the wild," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1145-1151, 2017.

[16] J. Chung and A. Zisserman, "Lip Reading in Profile," in *British Machine Vision Conference*, 2017. https://arxiv.org/pdf/2403.16071

[17] P. Ma, A. Haliassos, A. Fernandez-Lopez, H. Chen, S. Petridis and M. Pantic, "Auto-AVSR: Audio-Visual Speech Recognition with Automatic Labels," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1139-1174, 2023.

[18] A. Fernandez-Lopez and F. M. Sukno, "Survey on automatic lip-reading in the era of deep learning," *Image and Vision Computing,* vol. 78, pp. 53-72, 2018.

[19] G. Alex, F. Santiago, G. Faustino and S. Jürgen, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning*, pp. 27-35, 2006.

[20] M. D. Zeiler and R. Fergus, "Visualizing and Understanding Convolutional Networks," in *Computer Vision–ECCV 2014: 13th European Conference*, Zurich, Switzerland, September 6-12, pp. 156-172, 2014.