Improving Preprocessing for LipReading

Dongwon Kim
Department of Statistics, SNU
dongwonida@snu.ac.kr

Jihoo Jung
Department of Economics, SNU
jjh123579@snu.ac.kr

Junhyeong Kong
Department of Plant Science, SNU
denovokjh@snu.ac.kr

Sejun Park
Department of Plant Science, SNU
aprimelonge@snu.ac.kr

Abstract

Lipreading decodes speech by watching how someone's mouth moves. Advancements in deep lipreading usually have two steps: preprocessing and using deep neural networks. While most focus on improving neural networks, we concentrated on refining preprocessing. In this project, we investigated the impact of introducing a preprocessing technique that highlights the segmentation around the lips on the accurancy of lipreading. Through the utilization of BlazeFace and BiseNet, we emphasized the lip segment within each video and the LipNet model was trained using this modified data. We evaluated five preprocessing methods using Miracl-VC1 for English and AI Hub Korean LipReading data. Contrary to our expectations, lip region utilization resulted in poor outcomes. But, when the contour addition and cropped lip methods were applied to Korean data, it performed a slightly better than baseline. However, this improvement was not significant. We discussed the reasons behind obtaining these experimental results, as well as the limitations and future work in our study.

1. Introduction

Lipreading is the ability that recognizes what is being said only with visual information around the lips. Lipreading plays an important role in communication and speech understanding for human, as highlighted in the McGurk effect [1], where dubbing the audio of one phoneme over the video of a person speaking another phoneme recognizes the third phoneme. Nowadays, lipreading has recently been used in a wide range of applications such as silent dictation, speech recognition in noisy environments.

However, in the field of computer vision, lipreading is still a difficult task. In fact, lipreading is also difficult for human either, because there are inherent ambiguity due to homophemes and heteronyms [2]. For example, 'tear' and 'tier' produce exactly the same lip sequence and if 'tear' is used as a verb, its pronunciation is different from when it is a noun.

Recent advancements in deep learning have significantly enhanced performance of lipreading. However, our research motivation stems form the fact that preprocessing stage is equally crucial in developing a robust lipreading system for additional improvements. [3] This is essential for mitigating visual ambiguity and achieving the goal of predicting words solely by analyzing the speaker's lip movements.

In our work, we aim to improve lipreading performance by utilizing techniques such as lip contour and lip region detection in preprocessing stage. In order to compare performance between the baseline preprocessing technique and the techniques we present, we will use the LipNet deep neural network model commonly used in lipreading. Additionally, given that most studies related to lipreading have been conducted with English datasets, we apply our technique with the best performance to Korean dataset.

This paper is organized as follows. In Section 2, we review related work on lipreading and lip segmentation. Section 3 describes our overall architecture and detail for lip contour and region detection we present. Section 4 introduces dataset and experimental process, and summarizes experimental results. Finally, conclusions of this project are reported in Section 5.

2. Related Works

In this section, we review a prior work on lipreading and lip segmentation.

LipReading In the realm of lipreading, most approaches historically employed machine learning techniques, not

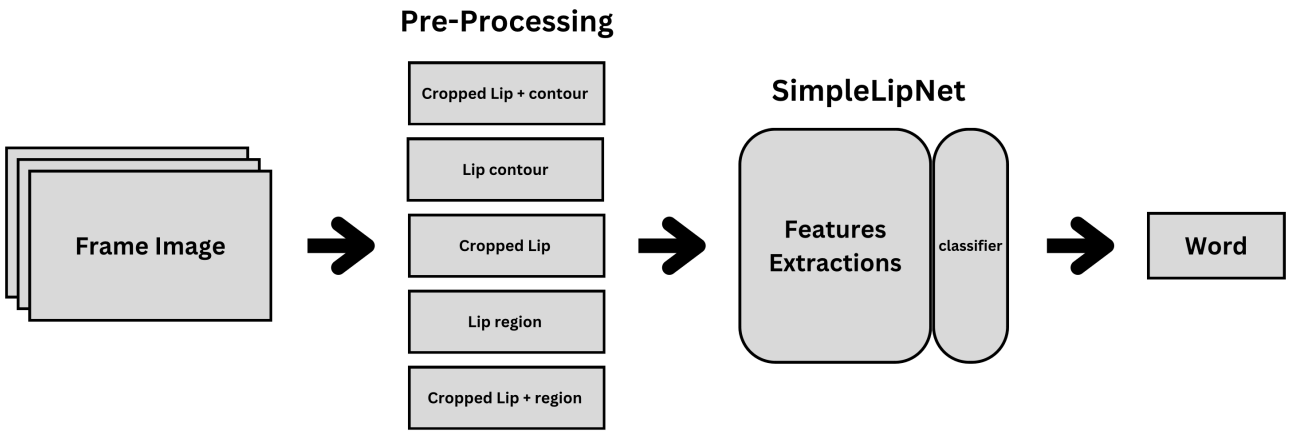

Figure 1. Overview of our lipreading system

deep learning. It is only very recently that deep learning methods have emerged and outperformed machine learning techniques. Pei et al. [4] employed Random Forest Manifold Alignment, focusing on extracting spatiotemporal trajectories through patched regions. These features were then mapped to motion patterns. Rekik et al. [5] used Hidden Markov Models (HMMs) with color and depth representations of images. They generated a 3D rendition of the speaker's mouth and derived diverse features from it, achieving a classification accuracy of 62.1% using HMMs on the MIRACL-VC1 dataset with speaker-independent testing.

After the advent of deep learning, early studies in lipreading have used Convolutional Neural Net works (CNNs) to predict phonemes [6] or visemes [7] from static images, rather than recognizing entire words or sentences. Phoneme represents the smallest distinguishable unit of sound that collectively forms a spoken word, while viseme is its visual counterpart. For the recognition of complete words, Petridis et al. [8] trained an LSTM classifier on discrete cosine transform (DCT) and deep bottleneck features (DBF). Similarly, Wand et al. [9] employed an LSTM with Histogram of Oriented Gradients (HOG) input features to recognize short phrases.

A recent work by Assael et al. [10] introduced LipNet, a spatiotemporal CNN and LSTM-based network utilizing Connectionist Temporal Classification (CTC) loss [11] for labeling computation. This study demonstrated strong performance on the constrained 51-word vocabulary of the GRID dataset [12] and surpassed the capabilities of experienced human lip readers. Our model draws inspiration mainly from this study.

Lip Segmentation Lipreading systems play a crucial role in pattern recognition, human-computer interaction, and artificial intelligence. These systems first identify the face region in images, extracting distinct mouth variation features from speakers. Subsequently, a recognition model is employed to determine the pronunciations of these features, enabling the system to recognize the speech contents. This approach has garnered increasing attention in recent years, particularly as it addresses challenges faced by audio speech recognition systems, such as a significant decline in recognition rates due to noises. In lipreading systems, Lip segmentation holds a fundamental role, as the accuracy of the segmentation directly influences the overall recognition rate. [13] Thus, we consider two types of lip segmentation as a pre-treatment for our lipreading systems : Lip Contour and Lip Region.

Early studies for lip segmentation employed computationally efficient techniques such as Fuzzy clustering [14] and k-means clustering [15] to distinguish pigmented lip-associated regions from non-lip regions. Other studies have focused on segmenting human lips based on contour and shape features. Notably, Le and Savvides introduced the Shape Constrained Feature-based Active Contour (SCFAC) model [16]. This involves integrating feature-based active contour (FAC) with prior shape constraints (CS) to achieve a high level of precision in segmentation.

With the advancement of deep learning, various models such as EHANet [17], Mask2Former [18] and PIDNet [19] have been introduced in the field of lip segmentation. Each model has merits under particular purposes and circumstances. In our project, we used Mediapipe Face Mesh [20] of Google to obtain lip contour and BiseNet V2 [21] to obtain lip region. The above models will be described in detail

in Section 3.

3. Methods

In this section, we describe our proposed lipreading system. Our system comprises primarily two main blocks. In the initial block, we focus on extracting and segmenting the lip region and solve them as a Computer Vision problem. BlazeFace [20] is employed for lip contour detection, and BiseNet V2 [21]is utilized for lip region detection. The resultant lip segment is then combined with the original image. Moving to the second block, outputs derived from the first block serves as input and undergoes classification through the LipNet model. Figure 1 illustrates our proposed lipreading system.

3.1. Lip Segmentation

Lip Contour Lip contour in each frame were detected using the faceMesh model of Mediapipe, which is same to BlazeFace [20]. BlazeFace is a deep learning model developed to swiftly and effectively address a broad spectrum of tasks associated with face detection. Designed for high-speed predictions on mobile GPUs, it's well-suited for lipreading that demand real-time resolution. The model's architecture prioritizes maintaining high accuracy while optimizing deployment on devices with limited computational resources.

BlazeFace achieves rapid and precise face detection through a combination of lightweight feature extraction, a GPU-friendly anchor scheme adapted from the Single Shot MultiBox Detector (SSD), and an enhanced strategy for tie resolution, offering an alternative to non-maximum suppression.

Figure 2 illustrates the results of applying our BlazeFace architecture to Miracl-VC1 dataset to get lip contour.

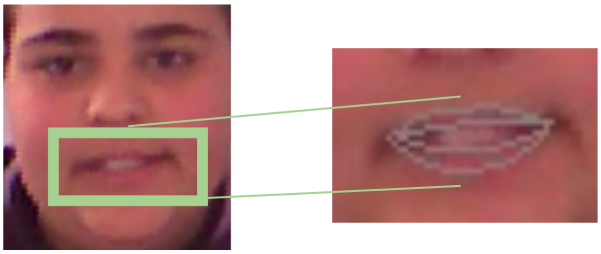

Figure 2. (left to right) image in the MIRACL-VC1 dataset; Original image + Lip contour

Lip Region We used network based on BiseNet V2 [21] and ResNet to obtain the lip region from the input image. BiseNet V2 represents a real-time semantic segmentation model that performs both high accuracy and preservation of low-level details. The architecture comprises the Detail Branch and the Semantic Branch. The Detail Branch is designed to capture low-level details through wide channels and shallow layers, thereby generating a high-resolution feature representation. In contrast, the Semantic Branch focuses on obtaining high-level semantic context using narrow channels and deep layers. To effectively fuse these two distinct layers and enhance their mutual connections, Guided Aggregation Layer is used. Additionally, a booster training strategy is employed to enhance segmentation performance without increasing the inference cost.

BiSeNet V2 achieves impressive results, demonstrating a Mean Intersection-over-Union (IoU) of 72.6% on the Cityscapes Dataset [22] for a 2,048×1,024 input.

Figure 3 illustrates the results of applying our BiseNet + ResNet architecture to Miracl-VC1 dataset to get lip region.

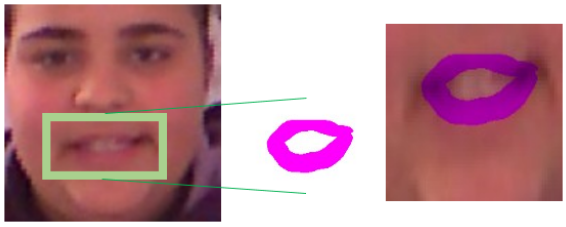

Figure 3. (left to right) image in the MIRACL-VC1 dataset; Lip region; Original image + Lip region

3.2. LIPNET

Traditional techniques for lipreading typically involved a two-stage process, wherein the problem was divided into designing or learning visual features, followed by prediction. However, recent deep learning approaches for lipreading have shifted towards end-to-end models. Notably, LipNet stands out as the pioneer in introducing an end-to-end model designed for sentence-level sequence prediction. [10] Human lipreading performance is known to increase with longer words, from which LipNet capturing temporal context in an ambiguous communication channel.

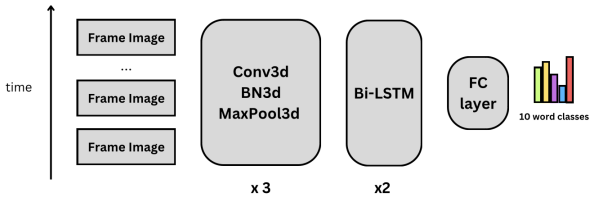

Figure 4. Overview of LipNet Architecture

Figure 4 depicts the LipNet architecture. This starts with a sequence of three blocks, each block consists of

spatiotemporal convolutions, channel-wise dropout, spatial max-pooling. The extracted features then undergo two Bidirectional Gated Recurrent Units (Bi-GRUs), playing a crucial role in effectively aggregating the output from the spatiotemporal convolutional neural network (STCNN). Finally, a linear transformation is applied at each time-step, followed by a softmax operation over the vocabulary augmented with the CTC (Connectionist Temporal Classification) blank, and subsequently, the CTC loss is computed. All layers in the architecture employ rectified linear unit (ReLU) activation functions.

LipNet achieves a strong performance of 95.2% accuracy on a Grid Corpus Dataset, particularly in overlapped speaker split task. This surpasses the performance of experienced human lip readers and the previous state-of-the-art accuracy of 86.4% at the word level.

Baseline technique The LipNet paper introduced a technique involving the processing of each video frame using the Dlib face detector and the iBug face landmark predictor with 68 landmarks. Using these facial landmarks, an affine transformation was applied to extract a mouth-centered crop, specifically sized at 100 × 50 pixels for each frame.

Figure 5. (left to right) image in the MIRACL-VC1 dataset; final cropped image using LipNet paper technique

4. Experiments

4.1. Datasets

MIRACL-VC1 The MIRACL-VC1 [23] dataset was used for performance comparison between preprocessing techniques. We used MIRACL-VC1 dataset because it is word-level lipreading data and appropriate size for initial evaluation process of deep learning models. MIRACL-VC1 contains both depth and color images of 15 speakers uttering 10 words and 10 phrases, 10 times each. That is, the dataset contains 3000 ($2 \times 15 \times 10 \times 10$) sequences, with varying lengths. The images within each sequence have dimensions of 640 x 480 pixels. They were captured at a rate of 15 frames per second and sequence lengths range from 4 to 27 image frames. The words and phrases are as follows:

Words: begin, choose, connection, navigation, next, previous, start, stop, hello, web.

Phrases: Stop navigation, Excuse me, I am sorry, Thank you, Good bye, I love this game, Nice to meet you, You are welcome, How are you, Have a good time

In our work, we used only words data and color images. Also, 3 of the 15 speakers who had difficulty in lip segmentation process were excluded from the experiment. The remaining 12 speakers were divided into sets, with 9 for training, 1 for validation, and 2 for testing. This partitioning resulted in datasets with 900, 100, and 200 examples for training, testing, and validation, respectively.

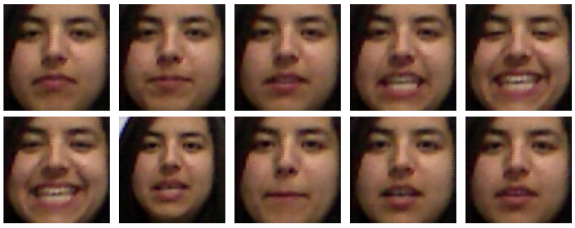

Figure 6. Example full input sequence where the subject is uttering the word "begin."

AIHUB Korean LipReading AIHUB Korean LipReading dataset [24] was used to apply the lipreading model specifically to the Korean language. We used this dataset because it contains word-level lipreading data similar to MIRACL-VC1. The AIHUB Korean LipReading dataset consists of videos, color images, lip keypoints and 3d model of 200 speakers uttering a total of 1000 words and phrases, repeated 5 times each. The images within each sequence have dimensions of 1600 × 1200 pixels. They were captured at a rate of 15 frames per second, and sequence lengths range from 3 to 16 image frames.

In our work, we exclusively utilized color images depicting 10 specific phrases uttered by 16 speakers. The 16 speakers were divided into sets: 13 for training, and 3 for testing. This partitioning resulted in datasets with 650 and 150 examples for training and testing, respectively.

Phrases: 좋아요, 주세요, 어디에 있어요, 이름이 뭐에요, 이게 뭐예요, 정리하세요, 가세요, 잠시만요, 대단해요, 모르겠어요.

4.2. Experimental Settings

Data Augmentation In our project, we employed a data augmentation to enhance the performance of the model

Figure 7. Example full input sequence where the subject is uttering the word "좋아요."

and overcome the limitations of the original dataset. Since the model exhibited subpar performance with the original dataset, it became necessary for us to implement data augmentation techniques. We manipulated the original image data in several ways to create a more diverse dataset, which in turn could help the model generalize better. Specifically, we implemented (1) Random Rotation of 15 degrees, (2) Random Horizontal Flip with a probability of 0.5, and (3) Random Resized Crop with a size of (30, 45) and a scale between 0.9 and 1.1. We ensured that the data augmentation was applied only to the training dataset and not to the validation and test datasets. This approach prevents the model from being influenced by any potential biases or anomalies that the augmented data might introduce during the validation and testing phases.

Hyperparameter Using the training and validation sets, we determined the optimal learning rate and L2 norm weight decay values for our model. The learning rate was set to 0.0001 and the L2 norm weight decay was set to 0.005.

Hyparameter	Value
Learning Rate (lr)	0.0001
L2 norm Weight Decay	0.005

Table 1. Optimum of Hyparameter

4.3. Experimental Results

In this work, two main experiments were conducted. The first experiment used the MIRACL-VC1 dataset to determine the best performance among our proposed techniques : (1) Baseline, (2) Lip Contour, (3) Orginal image + Lip Contour, (4) Lip Region, (5) Orginal image + Lip Region. In the second experiment, the preprocessing technique, which showed the best performance in the first experiment, was applied to the korean AIHUB LipReading dataset.

Preprocessing Techniques Selection The results of applying each preprocessing technique to the MIRACL-VC1 dataset are as follows.

Method	Managing policy	Accuracy
Dlib	Img	52.2
BlazeFace	Contour	45.8
BlazeFace	Img + Contour	51.9
BiseNet V2	Region	25.2
BiseNet V2	Img + Region	23.7

Table 2. Test Results on MIRACL-VC1 dataset

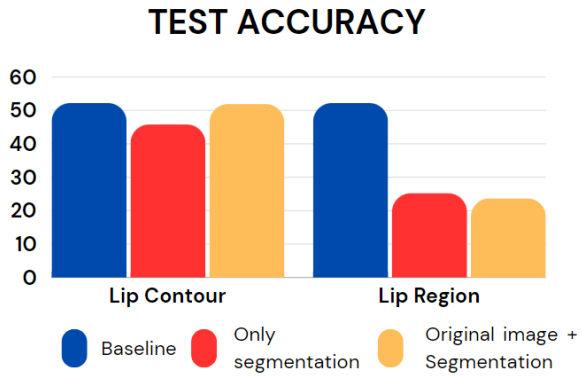

Figure 8. Comparison Graph on MIRACL-VC1 dataset

Table2 and Figure8 illustrate a result of experiment 1. The best performance was obtained from (1) the Dlib-based preprocessing method, which is a baseline model, and (3) the preprocessing method in which the Lip Contour was combined with the original image. Next, the performance followed (2) when only the Lip Contour was input. (4, 5) The preprocessing method using Lip Region performed very poorly regardless of the use of the original image.

The reason why the performance was poor in techniques using Lip Region compared to other techniques seems to be due to accuracy issue in finding lip region. The limitations were evident when the lips' shape was unclear or when the speaker's background color is similar to the face or lip color. Inaccurate lip region affects model's feature learning, resulting in poor performance. Figure9 below illustrates an example of a problem situation.

Korean Dataset In the second experiment, the (1) and (3) techniques that showed the best performance in the first experiment were applied to the korean AIHUB LipReading dataset.

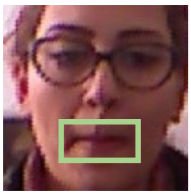

Figure 9. (left to right) Example of problem image; Shape of lip was unclear

Method	Managing policy	Accuracy
Dlib	Img	42.1
BlazeFace	Img + Contour	43.4

Table 3. Test Results on korean AIHUB LipReading dataset

The preprocessing technique using (3) original image + lip contour showed better performance than (1) the baseline preprocessing technique using only the original image. That is, We found that our processing technique improved performance even for Korean dataset.

However, it is difficult to say that this improvement is significant. There will be several reasons for obtaining the above results. First, as in the first experiment, the performance of lip contour was better than lip region, but not perfect in some case. Next possible reason is overfitting due to lip contour. According to previous studies, While LipNet could benefit from utilizing lip features in a "wild" environment [2], our dataset used in this project was generated in a controlled lab environment and was difficult to obtain its benefits. Last one is insufficient information of lip contour. In lipreading, other visual features may be more important than lip contour such as lip movement. [25] Therefore, depending on the situation, considering other features may be more helpful in improving performance.

5. Conclusion

Discussion This section aims to delve into two crucial discussions concerning our experimental outcomes. Firstly, we will examine the probable causes behind the observed limitations in our methodology related to lip region detection. During the preprocessing stages, we encountered challenges related to accuracy. Specifically, instances where the lip's shape was ambiguous or when the speaker's background closely resembled facial or lip tones significantly impacted the effectiveness of lip region detection using BiseNet V2. Secondly, we explore potential reasons why our methods fail to surpass the baseline. An accuracy issue emerged during the preprocessing steps, where the incorrect identification of lip boundaries resulted in the transmission of misleading information. Additionally, overfitting stemming from an emphasis on lip contour may contribute to the observed challenges. While LipNet could potentially benefit from utilizing lip features in a more dynamic environment, our dataset, generated in a controlled lab setting, posed difficulties in realizing these benefits. Furthermore, we discuss the issue of insufficient information regarding lip contour. In the context of lipreading, other visual features, such as lip movement, may play a more critical role than lip contour.

Further works Although two lip preprocessing techniques were employed, there was no significant improvement in lipreading accuracy. This outcome is likely attributable to the inaccuracies in BlazeFace and BiseNet, as well as the simplicity of the model architecture. Further research should explore alternative preprocessing models. Additionally, the current model, constrained by both its small size and outdated architecture, could benefit from replacement with a more intricate and contemporary model. Moreover, extending this project to incorporate a different dataset for addressing tasks beyond word classification, such as sentence reconstruction, represents the next frontier of research in advancing lipreading capabilities.

References

[1] Harry McGurk and John MacDonald. Hearing lips and seeing voices. *Nature*, 264:746–748, 12 1976. 1

[2] Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. Lip reading sentences in the wild. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3444–3453, 2017. 1, 6

[3] Malek Miled, Mohammed Anouar Ben Messaoud, and Aicha Bouzid. Lip reading of words with lip segmentation and deep learning. *Multimedia Tools Appl.*, 82(1):551–571, jun 2022. 1

[4] Yuru Pei, Tae-Kyun Kim, and Hongbin Zha. Unsupervised random forest manifold alignment for lipreading. In *2013 IEEE International Conference on Computer Vision*, pages 129–136, 2013. 2

[5] Ahmed Rekik, Achraf Ben-Hamadou, and Walid Mahdi. Human machine interaction via visual speech spotting. In *Advanced Concepts for Intelligent Vision Systems: 16th International Conference, ACIVS 2015, Catania, Italy, October 26-29, 2015. Proceedings 16*, pages 566–574. Springer, 2015. 2

[6] Youssef Mroueh, Etienne Marcheret, and Vaibhava Goel. Deep multimodal learning for audio-visual speech recognition. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2130–2134, 2015. 2

[7] Oscar Koller, Hermann Ney, and Richard Bowden. Deep learning of mouth shapes for sign language. In *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*, pages 477–483, 2015. 2

[8] Stavros Petridis and Maja Pantic. Deep complementary bottleneck features for visual speech recognition. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2304–2308, 2016. 2

[9] Michael Wand, Jan Koutník, and Jürgen Schmidhuber. Lipreading with long short-term memory. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6115–6119, 2016. 2

[10] Yannis M. Assael, Brendan Shillingford, Shimon Whiteson, and Nando de Freitas. Lipnet: End-to-end sentence-level lipreading, 2016. 2, 3

[11] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, page 369–376, New York, NY, USA, 2006. Association for Computing Machinery. 2

[12] Martin Cooke, Jon Barker, Stuart Cunningham, and Xu Shao. An audio-visual corpus for speech perception and automatic speech recognition (l). *The Journal of the Acoustical Society of America*, 120:2421–4, 12 2006. 2

[13] Yuanyao Lu and Tenghe Zhou. Lip segmentation using localized active contour model with automatic initial contour. *Neural Comput. Appl.*, 29(5):1417–1424, mar 2018. 2

[14] A.W.-C. Liew, Shu Hung Leung, and Wing Hong Lau. Segmentation of color lip images by spatial fuzzy clustering. *IEEE Transactions on Fuzzy Systems*, 11(4):542–549, 2003. 2

[15] Evangelos Skodras and Nikolaos Fakotakis. An unconstrained method for lip detection in color images. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1013–1016, 2011. 2

[16] T. Hoang Ngan Le and Marios Savvides. A novel shape constrained feature-based active contour model for lips/mouth segmentation in the wild. *Pattern Recognition*, 54:23–33, 2016. 2

[17] Ling Luo, Dingyu Xue, and Xinglong Feng. Ehanet: An effective hierarchical aggregation network for face parsing. *Applied Sciences*, 10(9), 2020. 2

[18] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1280–1289, 2022. 2

[19] J. Xu, Z. Xiong, and S. P. Bhattacharyya. Pidnet: A real-time semantic segmentation network inspired by pid controllers. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19529–19539, Los Alamitos, CA, USA, jun 2023. IEEE Computer Society. 2

[20] Valentin Bazarevsky, Yury Kartynnik, Andrey Vakunov, Karthik Raveendran, and Matthias Grundmann. Blazeface: Sub-millisecond neural face detection on mobile gpus, 2019. 2, 3

[21] Changqian Yu, Changxin Gao, Jingbo Wang, Gang Yu, Chunhua Shen, and Nong Sang. Bisenet v2: Bilateral network with guided aggregation for real-time semantic segmentation. *Int. J. Comput. Vision*, 129(11):3051–3068, nov 2021. 2, 3

[22] Alberto Garcia-Garcia, Sergio Orts-Escolano, Sergiu Oprea, Victor Villena-Martinez, and Jose Garcia-Rodriguez. A review on deep learning techniques applied to semantic segmentation, 2017. 3

[23] Ahmed Rekik, Achraf Ben-Hamadou, and Walid Mahdi. A new visual speech recognition approach for rgb-d cameras. In Aurélio Campilho and Mohamed Kamel, editors, *Image Analysis and Recognition*, pages 21–28, Cham, 2014. Springer International Publishing. 4

[24] National Information Society Agency. The open ai dataset project (ai-hub, s. korea) lip reading dataset. 4

[25] Meng Li and Yiu-ming Cheung. A novel motion based lip feature extraction for lip-reading. In *Proceedings of the 2008 International Conference on Computational Intelligence and Security - Volume 01*, CIS '08, page 361–365, USA, 2008. IEEE Computer Society. 6