Improving Preprocessing for Lip Reading

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

Lip Reading is the task of extracting spoken content from video footage. Traditionally, the solution to this problem involves two main stages: preprocessing of video data and predicting the spoken content. In our project, we aim to enhance Lip Reading performance by focusing on classical methodologies such as active contour, Hough transform in preprocessing stage. Our objective is to improve evaluation metrics or reduce operational costs by efficiently implementing these classical techniques. In addition, we propose to apply our methodology to Korean dataset. We will evaluate our approach using the following two datasets: GRID corpus and AIHub LipReading.

1. Introduction

Lip Reading is a computer vision problem that involves extracting speech content from spoken video footage. Solving this problem is typically carried out in two main stages. Firstly, there is a preprocessing stage, which involves accurately identifying the region of the face containing the lips, aligning and cropping the detected lip region horizontally. Following this, the process involves encoding the speech video and performing the decoding to generate the expected spoken content.

In our project, we aim to improve Lip Reading performance by utilizing classical methodologies in the preprocessing stage instead of commonly used deep learning-based approaches. Additionally, given that most studies related to Lip Reading have been conducted with English datasets, we would like to apply our methodology to Korean dataset. We expect improvements, such as enhancing evaluation metrics for the final output or reducing the overall operational costs, through the efficient utilization of classical methodologies.

2. Related Works

Our process consists of two parts: Video Preprocessing of facial videos and Predicting sentence from lip features.

2.1. Video Preprocessing

To begin, meticulous preprocessing of facial videos is essential as it serves as the cornerstone for robust feature extraction in the later stages using deep learning. The preprocessing pipeline typically includes the following steps [1]: 1) Identifying and accurately locating faces within the video to extract crucial facial landmarks. 2) Adjusting the frames to ensure that the detected lips are horizontally aligned. 3) Employing a uniform cropping technique to ensure that the lip region remains consistently centered in the image crop. In previous studies, these preprocessing steps were executed using standard libraries like DLib [2].

However, this approach presents limitations. Firstly, dlib does not allow for the isolation and extraction of only the lip-related landmarks. Extracting complete facial landmarks, including unnecessary ones, is required, demanding substantial computational power. Secondly, for lip landmarks, we are confined to extracting a fixed set of 17 points. There is no flexibility to choose different lip landmark locations or explore potentially superior approaches, such as utilizing entire edges instead of points. Thus, our objective is to implement these preprocessing steps independently, avoiding dependence on standard libraries. This way, we can freely conduct experiments to identify the most vital aspects related to Lip Reading while maintaining computational efficiency.

To accomplish this, we plan to employ methods like active contours, the Hough transform, etc. The algorithm will involve iterative refinement of the lip contour frame by frame, enabling effective detection of time-variant lip images, horizontal alignment of the detected lip contour, cropping, and other essential tasks.

2.2. Predicting sentence from lip features

After video preprocessing, the extraction of lip features from time-variant lip images and the prediction of full sentences are performed consequently. Machine learning has been used to effectively communicate information across multiple domains for Lip Reading, which combines both computer vision and natural language processing. LipNet is the first end-to-end sentence-level machine learning model suggested for Lip Reading and consists of Spatiotemporal convolutional neural networks and bidirectional Gated Recurrent Unit [1]. Additionally, Bi-LSTM [3] and temporal CNN [4] were used for other Lip Reading-related tasks.

Our prediction model is the same as one used in LipNet [1] and comprises two modules: 1) the lip feature extracting module and 2) the lip feature-based full-sentence predicting module. For the extracting module, CNN-based neural network which is a conventional approach that implicitly extracts image features by neural networks will be applied. It is able to harvest higher-dimensional features compared to traditional methods. Full sentence predictions are related to natural language processing (NLP). Despite various NLP models being proposed such as transformer, we plan to use GRU model applied in original LipNet. Following all these steps, full sentences are reconstructed from the video.

3. Dataset

We will evaluate our approach on publicly available datasets that are used in Lip Reading studies. Considering our objectives, we will use the following two datasets : GRID corpus and AI hub LipReading.

GRID corpus In recent years, there have been more audio-visual transcribed datasets in English, but most only contain single words. One exception is the GRID corpus [5], where audio and video recordings of 34 speakers who each produced 1000 sentences are performed for a total of 28 hours across 34000 sentences. We will use the GRID corpus to evaluate our initial Lip Reading model, because it is sentence-level data and appropriate size for use in initial evaluation process of deep learning models. The sentences are extracted from simple grammar of <command:(4)> <color:(4)> <preposition:(4)> <letter:(25)> <digit:(10)> <adverb:(4)>. Here, (number) represents the number of words selected for each of the six word categories. The categories consist of {bin, lay, place, set}, {blue, green, red, white}, {at, by, in, with}, {A, B, ...,Y, Z}/{W}, {zero, one, ..., eight, nine} and {again, now, please, soon}, respectively. This allows us to get 64,000 possible sentences. For example, two sentences in the data are ”set red by C four please” and ”place blue at F nine now”.

AIHUB LipReading Unlike English, there are only a few number of audio-visual transcribed datasets in Korean. In research related to Lip Reading of Korean speakers, most researchers created and used small datasets themselves for their research. Fortunately, since last year, AI Hub began offering publicly available sentence-level Lip Reading dataset. The AIHub Lip Reading dataset contains videos from more than 100 speakers on various topics such as current events and culture. And, each video consists of 2,500 short sentence videos of 5-10 seconds and audio and text files extracted from them. We will use this dataset to check whether our model, which performed well on the GRID corpus, also performs well in Korean.

References

[1] Yannis M. Assael, Brendan Shillingford, Shimon Whiteson, and Nando de Freitas. Lipnet: End-to-end sentence-level lipreading, 2016. 1, 2

[2] D. E. King. Dlib-ml: A machine learning toolkit. *J. Mach. Learn. Res.*, 10:1755–1758, Jul. 2009. 1

[3] Ümit Atila and Furkan Sabaz. Turkish lip-reading using bi-lstm and deep learning models. *Engineering Science and Technology, an International Journal*, 35:101206, 2022. 2

[4] Brais Martinez, Pingchuan Ma, Stavros Petridis, and Maja Pantic. Lipreading using temporal convolutional networks, 2020. 2

[5] 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