VIDEO TRANSCRIBING USING LIP READING

Rahul Patel
rp3752@nyu.edu

Sai Preetham Bojja
sb9509@nyu.edu

Saiteja Siddana
ss14789@nyu.edu

Abstract

In the realm of Automatic Speech Recognition (ASR), where audio has seen significant success, the decoding of visual speech poses a compelling yet open challenge. Beyond its relevance in audio-less videos, visual speech decoding is poised to complement existing systems, aiming to elevate accuracy by harnessing information from lip movements and facial expressions. This becomes particularly valuable in environments marked by audio challenges or disturbances. The integration of visual cues offers promising avenues for generating spoken content when audio proves unreliable due to disruptions or background noise. Furthermore, its applications extend to areas such as face forgery detection, leveraging insights derived from intricate mouth movements. Additionally, envision silent speech interactions on mobile devices as a potential outcome of advancements in visual speech decoding.

1 Introduction

Communication is a multifaceted process, extending beyond mere auditory cues to include visual components crucial for comprehensive understanding. Lipreading, becomes particularly challenging when attempted in isolation, divorced from its acoustic context. Many experiments involves exposing individuals to conflicting audio and visual stimuli, leading to the perception of an entirely new sound, highlighting the intricate relationship between auditory and visual elements in speech comprehension. The difficulty in lipreading lies in the subtle movements of lips, tongue, and teeth, which lack clarity without contextual information.

Deciphering lipreading cues is a challenging task as most lip movements, apart from those associated with the lips and occasionally the tongue and teeth, remain latent and challenging to distinguish without additional context (Fisher, 1968; Woodward & Barber, 1960) [2]. Fisher (1968) categorizes 5 visual phonemes, referred to as visemes, from a total of 23 initial consonant phonemes. These visemes are frequently confused when individuals attempt to interpret a speaker's mouth movements. Fisher's observations extend to asymmetric confusion patterns, and similar trends are noted for final consonant phonemes. Consequently, human lipreading performance tends to be subpar. For individuals with hearing impairments, achieving accuracy remains a challenge, with reported rates of only 17±12 for a limited subset of 30 monosyllabic words and 21±11 for 30 compound words (Easton & Basala, 1982) [1].

2 Related work

In 1997, Goldschen et al. pioneered visual-only sentence-level lipreading, employing hidden Markov models (HMMs) on a limited dataset with manual segmentation. However, a number of studies[7] show that the use of statistical models alone may not be sufficient to capture the video dynamics. A solution is to encode temporal information to improve informativeness and stability of the extracted visual features. Subsequently, in 2000, Neti et al. achieved the first sentence-level audiovisual speech recognition by combining HMMs with hand-engineered features, demonstrating their approach on the IBM ViaVoice dataset. Notably, they enhanced speech recognition performance in noisy environments by integrating visual features with audio ones. In a different approach, Gergen et al. (2016) employed speaker-dependent training on an HMM/GMM system. They utilized an LDA-transformed version of the Discrete Cosine Transforms of mouth regions to improve the performance of their system. Most recent advancements have been made in lip reading by using densely convoluted temporal convolution networks[6].

3 Approach

We present a end-to-end model for visual speech recognition that takes image sequences as input and outputs token sequences using the Connectionist Temporal Classification (CTC)[4] loss eliminating the need for manual alignments during training. In data pre-processing we used two approaches for cropping the mouth region of the frame one is cropping the frame to our required shape and the other approach is using the pretrained weights of dlib architecture and use mouth detector points to crop the frame to the required amount of size for input of our model.

In designing our model, we've chosen a different path compared to the common use of complex structures like RNNs and vision transformers for video tasks. Recent research has shown that simpler frameworks can be quite effective. An example of this is SimVP[3], a model that predicts video frames and has become a benchmark for many video prediction datasets.

Inspired by this shift towards simplicity and efficiency, our model adopts a stack of 3D convolution layers applied to video frames. This choice is motivated by the capacity of 3D convolutions to simultaneously process spatial and temporal information. This concurrent processing allows our model to adeptly capture intricate motion dynamics and learn dynamic patterns within the video data.

Following the stacked convolutional layers, our architecture incorporates Long Short-Term Memory (LSTM) networks. LSTMs prove particularly effective in lip reading architectures due to their proficiency in modeling sequential information, accommodating variable-length sequences, retaining long-term dependencies, facilitating bidirectional processing, and possessing feature learning capabilities. This strategic integration of 3D convolutions and LSTMs enhances our model's ability to understand the nuanced temporal and spatial aspects of lip movements, contributing to its overall effectiveness in the realm of video understanding.

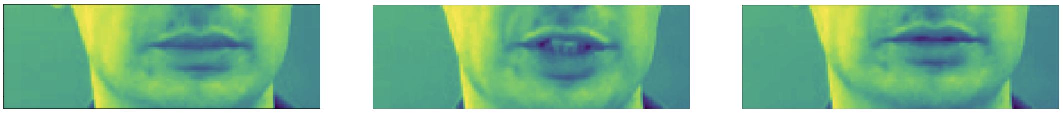
Table 1: Detailed architecture of the Model neural network.

Layer	Output Shape	Param #
Conv3d	$[1, 128, 75, 46, 140]$	3,584
ReLU	$[1, 128, 75, 46, 140]$	–
MaxPool3d	$[1, 128, 75, 23, 70]$	–
Conv3d	$[1, 256, 75, 23, 70]$	884,992
ReLU	$[1, 256, 75, 23, 70]$	–
MaxPool3d	$[1, 256, 75, 11, 35]$	–
Conv3d	$[1, 75, 75, 11, 35]$	518,475
ReLU	$[1, 75, 75, 11, 35]$	–
MaxPool3d	$[1, 75, 75, 5, 17]$	–
Linear	$[1, 75, 128]$	816,128
ReLU	$[1, 75, 128]$	–
LSTM	$[75, 1, 256]$	264,192
Dropout	$[75, 1, 256]$	–
LSTM	$[75, 1, 256]$	395,264
Dropout	$[75, 1, 256]$	–
Linear	$[75, 1, 40]$	10,280

4 Data processing

The GRID corpus encompasses 34 individuals, each tasked with narrating 1000 sentences. These sentences adhere to a straightforward grammar structure: command(4) + color(4) + preposition(4) + letter(25) + digit(10) + adverb(4). Here, the numbers denote the available word choices within each of the six categories, which include {bin, lay, place, set}, {blue, green, red, white}, {at, by, in, with}, {A, . . . , Z}, {zero, . . . , nine}, and {again, now, please, soon}. This structured approach results in a vast set of 64,000 possible sentences. For instance, examples from the dataset include sentences like "set blue by A four please" and "place red at C zero again."

For our model development, we specifically selected videos from one person, and during testing, we maintained consistency by evaluating the model on the same individual. We opted against testing on different individuals' facial features due to the inherent complexity of this task. Testing on untrained speakers poses a significant challenge, and even benchmark models struggle to achieve high accuracy in such scenarios. This difficulty arises from the variations in facial expressions, articulation, and speech patterns among different speakers, making it a challenging problem with limited accuracy in existing models. In selecting a dataset from a single individual, we began by cropping the frames to a size of (46,140) around the lip region, achieved through the cropping operation of pixels [190:236, 80:220] in the frame of shape (288,360).



(a) 1st frame (b) 25th frame (c) 50th frame

Figure 1: Frames Captured at Different Time Intervals

In this dataset, the person's face consistently occupies a confined region, ensuring stability and consistency in the captured visual features. To validate this, we employed dlib's pretrained weights to focus specifically on the mouth region, aligning the frame size to 46 by 140 pixels. The use of dlib yielded comparable results, affirming the consistent location of the speaker's face within the frames. Consequently, for this dataset with a single speaker, we opted for the cropping operation over utilizing the dlib architecture. This decision is based on the understanding that when dealing with multiple speakers, the lip region may vary significantly in terms of pixel positions for each speaker, making dlib a more suitable choice in such diverse scenarios.

Our model processes each input video by dividing it into 75 frames, focusing specifically on the cropped region containing the mouth. Following this, the frames undergo stacked spatiotemporal convolutional layers and bidirectional LSTMs, resulting in an output of shape (75, 1, 40) yielding a distribution over 40 characters, including a blank token used for the Connectionist Temporal Classification (CTC) loss. During training, a logarithmic softmax is applied along the character dimension for each frame, and the CTC loss is computed to optimize the model. CTC greedy decoding is employed to determine the final alignment of the predicted string during inference.

5 Experiments and Results

As dicussed in the above sections, we tailored the experiments to evaluate our architecture under various conditions, specially focusing on loss function, different decoding strategies, model initialization, optimization, learning rate scheduling, and using different pre-processed data to validate our results.

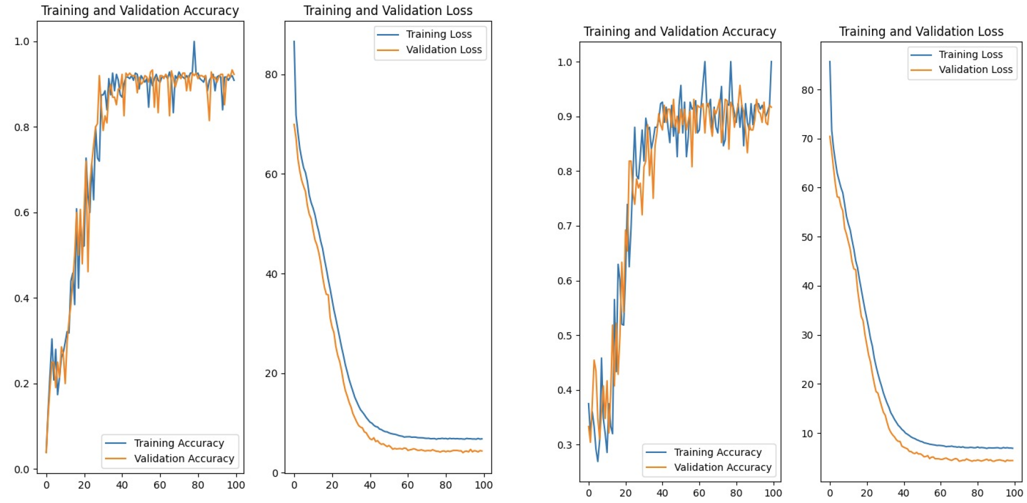
5.1 Experimental Settings

The dataset was partitioned into training, validation, and test subsets, following a 90%, 9%, and 1% split ratio, respectively. Batch processing was implemented using data loaders, and inputs were reshaped and permuted as required by the model. We utilized the Adam optimizer with an initial learning rate of 0.0001 and included a weight decay of 1e-4 to implement L2 regularization. A learning rate scheduler was employed, maintaining a constant rate for the first 30 epochs, followed by an exponential decay. We conducted the training of our model over a span of 100 epochs, after which we conducted character-by-character predictions on a test dataset to evaluate the performance of our model. Some of the sample predictions are shown in figure-3 Model evaluation was conducted using the Character Error Rate (CER), which measures the performance of the model in terms of character recognition accuracy. Accuracy was calculated using the edit distance between the original and predicted sentences. This method provided a quantitative measure of how closely the model's predictions matched the actual spoken sentences

5.2 Results

Throughout the training, the model demonstrated a consistent improvement in learning patterns, as indicated by the reduction in loss.

At the end of our training, our model with greedy decoding strategy in CTC loss, gave a stable accuracy of around 92% at the end of 100 epochs, achieving a peak of 98% accuracy in between. Validation loss exhibited a significant decrease, starting at 71 and dropping to 5.4 by the end of the 100th epoch. Similarly, with dp decoding strategy, we achieved an accuracy of around 93%. Validation loss mirrored the positive trend of the greedy approach, starting at 70.5 and decreasing to 5.1, suggesting slightly more efficient learning or better sequence alignment handling. Same configuration, with the integration of dlib mouth detector maintained the same accuracy and training, validation loss at the end of 100 epochs, affirming the robustness of our results.



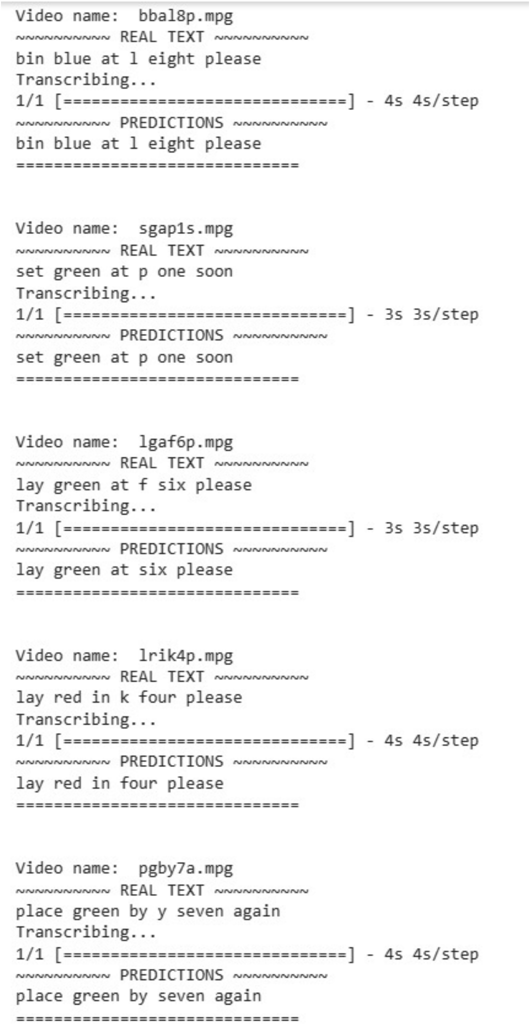
(a) CTC Greedy Decoding

(b) CTC DP Decoding

Figure 2: Training, Validation Loss and Accuracy plots vs Epoch Number

Method	Accuracy
Greedy	92.1
DP	93.4
+ Dlib	93.2

Table 2: Accuracy for different configurations.



```
Video name:  bbal8p.mpg
~~~~~~~~~~ REAL TEXT ~~~~~~~~~~
bin blue at l eight please
Transcribing...
1/1 [==============================] - 4s 4s/step
~~~~~~~~~~ PREDICTIONS ~~~~~~~~~~
bin blue at l eight please
==============================


Video name:  sgap1s.mpg
~~~~~~~~~~ REAL TEXT ~~~~~~~~~~
set green at p one soon
Transcribing...
1/1 [==============================] - 3s 3s/step
~~~~~~~~~~ PREDICTIONS ~~~~~~~~~~
set green at p one soon
==============================


Video name:  lgaf6p.mpg
~~~~~~~~~~ REAL TEXT ~~~~~~~~~~
lay green at f six please
Transcribing...
1/1 [==============================] - 3s 3s/step
~~~~~~~~~~ PREDICTIONS ~~~~~~~~~~
lay green at six please
==============================


Video name:  lrik4p.mpg
~~~~~~~~~~ REAL TEXT ~~~~~~~~~~
lay red in k four please
Transcribing...
1/1 [==============================] - 4s 4s/step
~~~~~~~~~~ PREDICTIONS ~~~~~~~~~~
lay red in four please
==============================


Video name:  pgby7a.mpg
~~~~~~~~~~ REAL TEXT ~~~~~~~~~~
place green by y seven again
Transcribing...
1/1 [==============================] - 4s 4s/step
~~~~~~~~~~ PREDICTIONS ~~~~~~~~~~
place green by seven again
==============================
```

Figure 3: Sample predictions of 5 videos.

6 Conclusion

In this project, we have developed a novel lip-reading model using a combination of advanced neural network techniques that can be trained end-to-end. Despite variations in performance throughout the training process and for different setups, the model consistently showed a significant reduction in validation loss, proving its ability to learn and adapt effectively. These outcomes not only validate our model's efficacy but also its potential application in real-world scenarios where robust lip-reading capabilities are required.

7 Future work

In future, we plan integration between Automatic Speech Recognition (ASR) and Visual Speech Recognition (VSR) in finding optimal combination strategies that contribute to the robustness of the system. Additionally, we plan to assess how changes in the frame size around the mouth region impact accuracy and predictions. By expanding the frame size to encompass features such as upper cheek movements and other facial expressions, potentially correlated with speech, we aim to capture a richer set of features that may enhance the system's accuracy. The consideration of these additional features is anticipated to provide a more nuanced understanding of speech. Furthermore, in terms of architecture, we intend to delve deeper into the utilization of vision transformers[5].

References

[1] Randolph D Easton and Marylu Basala. Perceptual dominance during lipreading. *Perception & Psychophysics*, 32:562–570, 1982.

[2] Cletus G Fisher. Confusions among visually perceived consonants. *Journal of speech and hearing research*, 11(4):796–804, 1968.

[3] Zhangyang Gao, Cheng Tan, Lirong Wu, and Stan Z Li. Simvp: Simpler yet better video prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3170–3180, 2022.

[4] 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*, pages 369–376, 2006.

[5] Pingchuan Ma, Stavros Petridis, and Maja Pantic. End-to-end audio-visual speech recognition with conformers. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7613–7617, 2021.

[6] Pingchuan Ma, Yujiang Wang, Jie Shen, Stavros Petridis, and Maja Pantic. Lip-reading with densely connected temporal convolutional networks. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2857–2866, 2021.

[7] Gerasimos Potamianos, Chalapathy Neti, Guillaume Gravier, Ashutosh Garg, and Andrew W Senior. Recent advances in the automatic recognition of audiovisual speech. *Proceedings of the IEEE*, 91(9):1306–1326, 2003.