#12

Helen - An Accessibility Device for AI-based Lip Reading

KINAGI, Amrutavarsh KRISHNAMURTHY, Padmanabhan

Project Supervisor: Prof. Brian Mak

Abstract

In this project, we introduce Helen, an accessibility device that uses deep learning to perform lip reading. Helen consists of a camera connected to a RaspberryPi computer that streams video to a device running LipNet : a deep learning model that achieves high accuracy on automatic lip reading. Helen is compact and is designed to either be worn around one's neck or clipped on to one's glasses. Its software workflow has been optimized to ensure that the entire process from recording the video to obtaining annotations takes no more than 10 seconds. The LipNet model used for this proof-of-concept iteration of Helen achieves an accuracy of 60% on GRID-dataset-like sentences, and also works successfully - albeit sporadically - on a small custom-made dataset of everyday sentences. We believe that this project proves the concept that deep learning based lip reading transcription can be successfully deployed in real-world scenarios in order to supplement hearing aids, enable audio-less communication and enhance several other fields like forensic lipreading and automated subtitle generation.

Introduction

Lip reading is the process of transcribing the content of speech using only the movements of a speaker's lips. It is particularly beneficial when applied to hearing aids, audio-less communication scenarios, forensic lip reading, and automated subtitle generation, among other areas. Most hearing aids are currently hindered by their inability to function well in the presence of background noise. They either fail to zone in to the target's voice, thus producing very meek audio signals, or amplify the noise of the entire crowd, resulting in jarring results. As far as audio-less communication is concerned, we commonly find ourselves in scenarios where communication through audio or textual media is impossible. These scenarios include trying to communicate on a noisy train, or speaking with a disabled person who is unable to use his voice. In the domain of forensic lipreading, courts of law often have to infer what was said in a video clip with no audio (such as CCTV camera recordings). However, the transcriptions provided by court-appointed lip readers cannot always be relied upon, due to unfamiliarity of the lip reader with the accent of the speaker, or susceptibility to biases and stress.

All of these limitations can be circumvented by using automated lip reading systems, which open up several new avenues for human communication and interaction. By mapping changes in a speaker's lip movements to spoken words, it is possible to transcribe the content of a speech without using any audio information. Building upon LipNet [1], a refined deep learning model for automated lip reading developed by Oxford and Google's Deep Mind, our team has built Helen: a proof-of-concept wearable device that captures a speaker's lip movements and provides users with a transcription of the spoken content. Our system - which does not receive any audio inputs - outperforms human lip-readers, and promises to provide hearing impaired patients with a much needed alternative dimension to interact with people particularly in noisy surroundings.

Technical Description

Helen consists of 3 key components:

1. Sony's IMX219 8-megapixel camera sensor to capture spoken input
2. RaspberryPi's Zero W for high quality 802.11 b/g/n wireless LAN video streaming

3. A working implementation of LipNet to annotate the video input received from the Raspberry Pi

- **Hardware Design and Video Transmission**

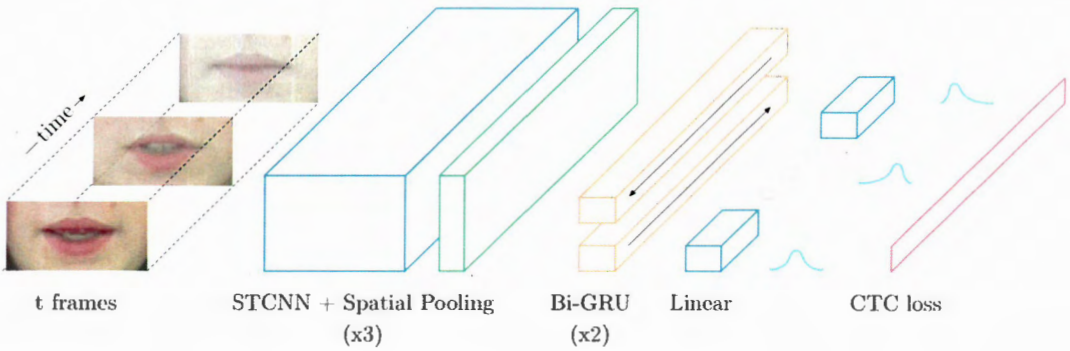
The Raspberry Pi board and the camera module are both housed in a compact enclosure (7cm * 3.5cm * 1.5cm) which is attached to a custom designed 3D-printed structure that ensures wearability. This enables Helen's hardware component to be connected to a user's body in varying configurations, most efficaciously around his or her neck.

Helen

To eliminate the latency that arises from post-video-recording file transfers, Helen streams video while recording and thus has a relatively rapid execution time. Helen utilises the inbuilt wireless LAN module to enable a UV4l streaming server on the Raspberry Pi. A script invokes the stream, which can be then be accessed by devices on the connected local network by directly accessing the mjpeg stream over the http protocol. Helen streams a 400*400 30 fps video stream to the client, which in turn utilises Opencv and ffmpeg to save the video and convert it to the mpeg format required for LipNet to run. The client script also partitions the stream into chunks of a pre determined frame length, which are then analyzed by LipNet. By changing this frame length, we can modify LipNet to annotate videos of different lengths. This multiprocessing between tasks ensures a low prediction time over longer phrases and enables near-real-time transcription while streaming.

- **LipNet: Underlying model used by Helen to perform lip reading:**

LipNet is a refined deep learning model created by researchers at Oxford and DeepMind to perform automated lip reading. The architecture of LipNet consists of Spatiotemporal Convolutional Neural Networks (ST-CNNs), Spatiotemporal Pooling Layers, Bi-Directional Gated Recurrent Units (Bi-GRUs) and Connectionist Temporal Classification (CTC) Loss. When LipNet receives a video as input, it uses ST-CNNs to extract features from the video frames, processes these features using the Bi-GRUs and finally examines the correctness of its output predictions using CTC loss. The CTC loss is then backpropagated through the model in order to optimize the model's parameters and improve LipNet's performance. For the sake of clarity and relevance to the theme of this project report, the detailed working of LipNet has been omitted from the report. We are, however, more than happy to explain the inner workings of LipNet and elucidate its intricacies to interested parties.



LipNet's Architecture as illustrated in [1]

For the sake of proof of concept, we have trained LipNet on 2 kinds of data:

1. The GRID Corpus: A large, standard dataset used for building automated lip reading systems. It consists of 64000 videos of 3-second sentences of the form *command + color + preposition + letter + digit + adverb.*
   Eg: *"Place red in A zero now"*
2. A custom dataset created by the authors of this project containing 15 examples everyday phrases likely to be processed by automated lip reading systems.
   Eg: *"Call an ambulance"*

It is important to note that since the GRID dataset is exponentially larger than the custom dataset created by us, the LipNet model trained on it is far more reliable and accurate than the model trained on our smaller, custom dataset. Consequently, Helen performs at its peak when tested on GRID-like sentences. However, should we be given more time and resources to build a dataset of comparible size to GRID, we are confident that Helen will yield GRID-like performance on a wide variety of sentences. These results will be discussed in depth in the evaluation section of this report.

- **Hardware-Model Linkage - Tying it all together:**
  This section will discuss the steps taken to integrate our hardware with LipNet and ensure that video captured by Helen is annotated and transcribed successfully. The workflow for our proof-of-concept is as follows:

  1. As mentioned in the video transmission section, the RaspberryPi and the device running LipNet are connected to a local wifi network. Video captured by the camera is streamed to LipNet across this network, and converted from '.avi' to '.mpg'.
  2. Again, our proof-of-concept device achieves best performance when given GRID-like sentences to annotate. These sentences are ideally 3 seconds long, as a result of which the camera is set up to record video in 3-second intervals. This interval can be modified to work with non GRID-like sentences.
  3. 100x50 crops of the speakers lip movements are extracted from the input video using dlib's face detector. These cropped lip movement features are passed forward through the LipNet model to ultimately obtain predicted transcriptions of the spoken content.

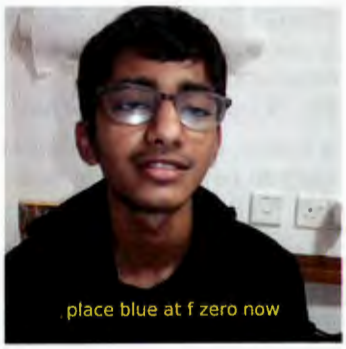
  Over the course of the development of Helen, this workflow has been extensively optimized so that even at this proof-of-concept stage, the entire process of recording a 3 second video, streaming it to a target device, feeding the video into LipNet and obtaining the transcribed content takes no more than 10 seconds.

Evaluation:

In this section, we will describe the results of deploying Helen in real-world scenarios, on individuals who's faces are vastly dissimilar to those faces used for training the LipNet. We have trained 2 different versions LipNet on both the GRID-dataset (to demonstrate how well Helen performs when trained on a large dataset) and a much smaller, custom-made dataset (to demonstrate LipNet's potential to transcribe sentences of varying nature). The results of evaluating both models on completely unseen speakers (speakers from HKUST who did not take part in the creation of either dataset) are shown below. It is important to note that best results were obtained when Helen was placed within a distance of 40cm from a speaker, whose lips were adequately illuminated.

- **Helen using LipNet model trained on GRID Dataset:**
  The model that has been trained **and evaluated** on the GRID dataset achieves a word error rate (WER) of 15.7%. WER is defined as the number of substitutions, deletions and insertions required to transform the ground truth of a sentence into the prediction made by the model. When evaluated on unseen speakers, Helen's LipNet performed surprisingly well, and yielded the following results.

(a) Truth:'bin red in o five please'



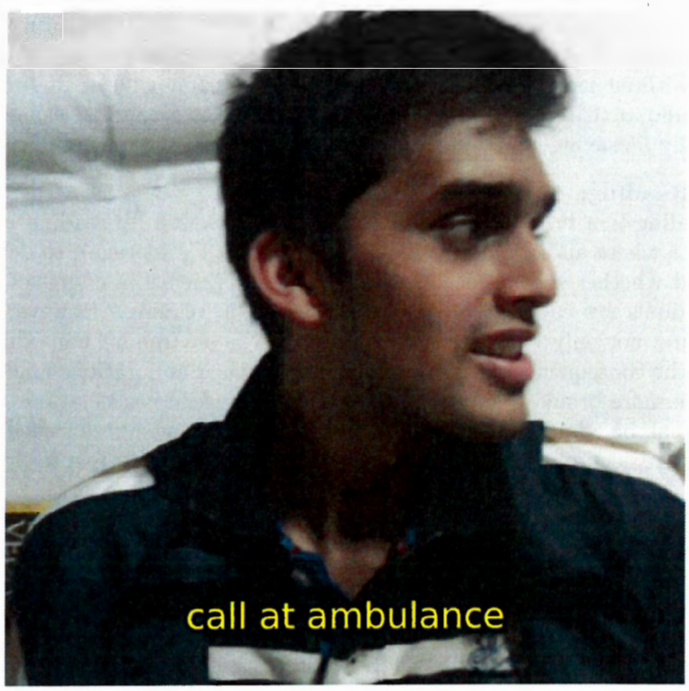
(b) Truth:'place blue at f two now'

(c) Truth:'set white with p two soon'

(d) Truth:'place red in a zero now'

Figure 1: Evaluating Helen on LipNet model trained on GRID dataset

- **Helen using LipNet model trained on custom dataset:**
  A custom dataset was created by the authors of this project to demonstrate that the model is capable of learning even non GRID-like sentences. This model was trained on 15 phrases everyday phrases that might necessitate lipreading. Unfortunately, due to a paucity of time, this dataset could not be made comprehensive enough to work on all non GRID-sentences.
  However, it performs well when given certain phrases like 'call an ambulance' or 'call the police', thus **proving that if we are ideally supplied with enough time and data, LipNet can be trained to work extremely well on sentences across multiple domains.** Between the time of submission of this report and the live demonstration of Helen, we hope to have increased the size of this dataset and improved the model's performance on it.



Truth:'call an ambulance'

As is apparent from these results, Helen's transcriptions of speech using LipNet are far from perfect. However, keeping in mind that professional lip-readers achieve an accuracy of only 53% on the GRID dataset, reaching comparable - and in many cases better - results with our proof-of-concept device is an unprecedentedly firm validation of the concept. Again, we reiterate that given larger datasets and enough time, LipNet - and hence Helen - can work successfully for multiple sentences across a variety of domains.

Applications:

Helen is applicable to any domain in which lip reading is already being used. These domains include hearing aids and accessibility devices, forensic lipreading, automated subtitle generation, etc. For the sake of brevity,

we have narrowed this set of domains to the following 3:

1. **Supplements to hearing aids:**
   Several experts in the hearing devices community believe that automated lip reading can be considered as a viable supplement to hearing aids that are incapable of isolating a speaker's noise from amongst a crowd of people or other noisy surroundings. In such scenarios, having a device that enables one to merely look at someone and comprehend what he or she is saying, would be extremely beneficial to hearing impaired patients. Models like LipNet enable such transcriptions to be performed with clean data under research settings, and devices like Helen enable such models to be deployed in the public domain, so that they can be put to use by patients who truly need them.

2. **Communication in extenuating circumstances where audio is unavailable:**
   Consider a scenario where a person taken hostage needs to call the police, but cannot use his voice to do so. Or where a disabled patient unable to use his hands (say, due to paralysis) needs to call an ambulance but cannot articulate his request. Or, on a less morbid note, where a person on a train needs to dictate a message into his phone, but cannot do so either because the train is too noisy for speech recognition to work well, or because the train is too silent for him to talk without drawing attention to himself. These are just a few of the many situations in which communication needs to be carried out without using audio or textual input, but cannot occur due to a lack of a device that enables such audio/text-free communication. In all of these scenarios, using a device like Helen can enable potentially lifesaving communication to be carried out without requiring any audio.

3. **Forensic Lip Reading:**
   Forensic lip reading is a branch of forensic linguistics in which lip reading is used to transcribe the content of videos where audio is not available (such as CCTV footage), to determine what transpired in the video and whether such videos are admissible as evidence in courts of law. Currently, hearing impaired individuals are called upon to perform such lip reading. However, they are placed under immense pressure, not only because of the gravity of the environment in which they are placed, and also because of the consequences and repercussions that their annotations might have. Moreover, being human, they are more prone to emotional biases and potential intimidation than automated systems. These concerns can be alleviated by using automated lip reading systems such as Helen, which can not only be made more accurate with additonal data, but is less susceptible to pressure and intimidation.

Future Work and Conclusion:

In its current proof-of-concept stage, we believe that Helen successfully demonstrates how automated lip reading systems like LipNet can be deployed in the real world, to aid hearing impaired patients, facilitate communication in noisy environments and ascertain what the content of an audio-less speech. However, it has tremendous scope for improvement. First and foremost, we aim to improve the performance of the LipNet model that is at Helen's core, so as to enable Helen to transcribe non GRID-like sentences from multiple domains. This can be done by preparing a much larger dataset containing sentences of varying length, complexity and speakers (as we would like Helen and LipNet to be as independent of the accent of the speaker as possible). Secondly, we aspire to make Helen even more compact than it already is, so that it is more intuitive to wear and carry. We propose to achieve this by using a custom-designed micro-controller that eliminates many of the unnecessary components like SD card slots and HDMI ports that come bundled with the RaspberryPi. Finally, in its current state, the device on which LipNet runs is a laptop. Therefore, a user is currently limited to viewing the results of a transcription on a laptop. However, since the uv4l server-based mechanism that we use to stream videos from Helen's camera to a target device is device-independent and requires no additional configuration on the part of the user, we hope to implement Helen on handheld devices like smartphones and tablets using emerging technology like TensorFlow JS that enables machine learning models to be deployed on computationally-weak devices.

In conclusion, we hope that we have been able to elucidate the need for automated lip reading systems, demonstrate a fully functioning device that performs this function, and display results that are indicative of the potential possessed by our device. While we acknowledge the inconsistency in the accuracy of annotations

that currently plagues our system, we urge the reader to consider the fact that our progress in automatic lip reading is akin to the early days of automated speech recognition, when even the most sophisticated systems could accurately recognize only a keywords/phrases at best. Just as automated speech recognition has evolved to increasingly refined levels over the past few decades, we believe that automated lip reading will not only see faster growth due to the availability of deep learning systems, but will also exhibit wider adoption by virtue of providing an entirely new dimension to interact with one's surroundings. We sincerely hope that our work with Helen is a precursor to a series of automated lip reading devices that adopt similar deep learning approaches to efficiently and accurately make communication easier for both disabled patients and society at large.

Acknowledgements

We would like to place on record our deepest gratitude to our supervisor, Prof. Brian Mak for his continuous feedback, support and resources that facilitated the evolution of this project. We would also like to acknowledge the work of Omar Salinas who's open source implementation of LipNet was invaluable. Finally, we would like to thank Barbara Hitchins, the secretary of ATLA (the United Kingdom Association of Teachers of Lipreading to Adults) for her insighful perspective on the merits and challenges of lip reading.

References

[1] Y.M.Assael et al., "LipNet: End-to-End Sentence-Level Lipreading", 2016. [Online].
Available: https://arxiv.org/pdf/1611.01599.pdf