100%

ASHESI UNIVERSITY

Machine Learning Final project proposal.

Lip Reading or Speechreading.

Submitted by:

Yussif Mustapha Tidoo, Jean-Sebastien Dovonon, Samuel Atule, Nutifafa Cudjoe Amedior.

On the

8th February 2019.

TABLE OF CONTENT

PROJECT SUMMARY

"Lip reading, also known as lipreading or speechreading, is a technique of understanding speech by visually interpreting the movements of the lips, face, and tongue when normal sound is not available. It relies also on information provided by the context, knowledge of the language, and any residual hearing." (Woodhouse, L; Hickson, L; Dodd, B., 2009). For decades, the idea of making machines detect speech from a silent video clip was conceived to be extremely difficult. However, projects such as Oxford University's LipNet has proven that this extremely difficult problem can be solved by machines, to a performance that beats professional lip readers. We are attempting to solve the problem of lip reading.

The difficulty in this problem comes from the fact that we need to consider factors such as the language been spoken as well as the knowledge of the context. Furthermore, we also need to decern visual clues from the videos. These factors involved makes the problem cumbersome.

Primarily, lipreading is often used by the deaf and hard-of-hearing people. However, there are other used cases where lipreading could be applied. This application could come in handy when a person loses their voice, and may not be able to speak audibly. Integrating this application in software can help the other person understand you, without the stress of forcing to speak.

METHODS

Algorithms

The nature of the problem means we are going to use video inputs for our program. This type of input needs to be redefined depending on the algorithm that is being used to fit the kind of input this algorithm takes. Also, we will use supervised learning in order to teach the algorithm on how to do a task that can be performed by human beings. Therefore, we need labels generated to follow the corresponding video data.

Given the nature of the data, the range of algorithms we can use includes Hidden Markov Models (HMM), and different flavors of neural networks combinations (Recurrent Convolutional Neural Networks, RCNN or Long-Term Recurrent Convolutional Neural Network, LRCNN).

Hidden Markov Models can be used on video data to model the probability of an output sequence given an input sequence and then choose the output with the highest probability. It will, however, require a preprocessing step which will segment the video in frames corresponding to each visemes (a shape that the lips can take). This causes an issue because the sequence of visemes is usually shorter than the corresponding text sequence. However, HMMs and hybrid SVM-HMMs (Hassanat, 2011) seem to reach good accuracy levels.

.

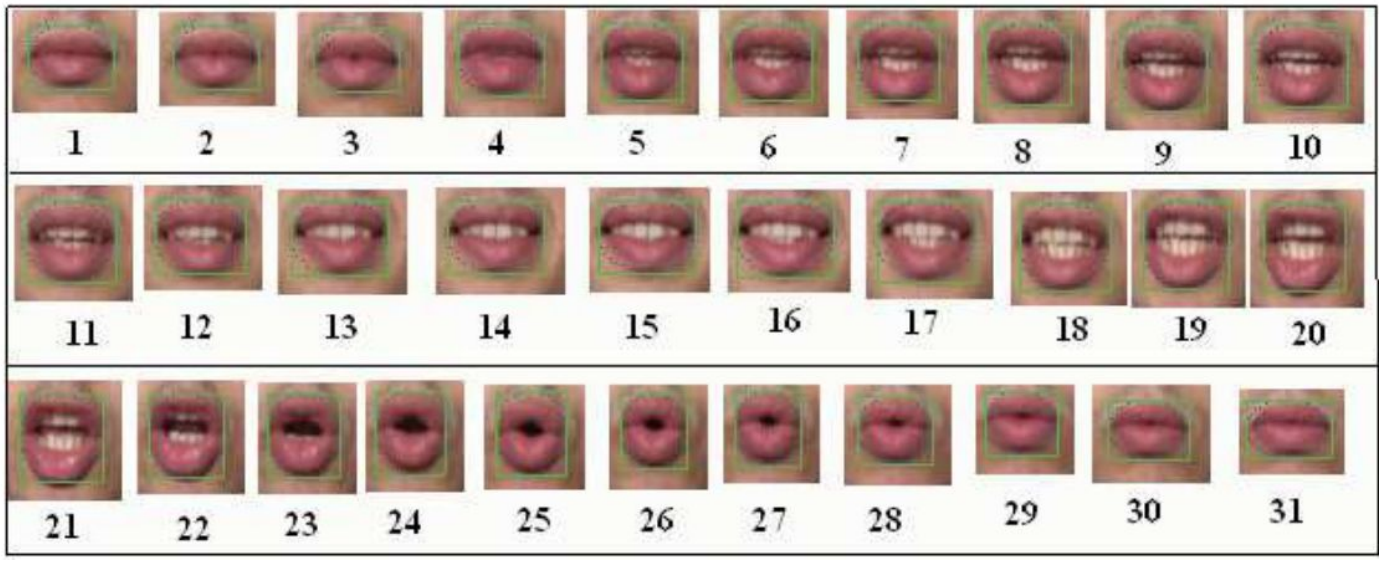

Figure 1: Example sequence of visemes for a video

Possible neural network architectures that can be used here, in the case of a sequence of inputs include Recurrent Neural Networks (RNNs) and Long Short Term Memory (LSTMs). Those can be used directly on a sequence of visemes. However, to make full use of the feature extraction capabilities of a neural network, we can use the raw videos as inputs. It will require us to use videos as a series of frames. It means we will need to include convolutions steps before the recurrent units.

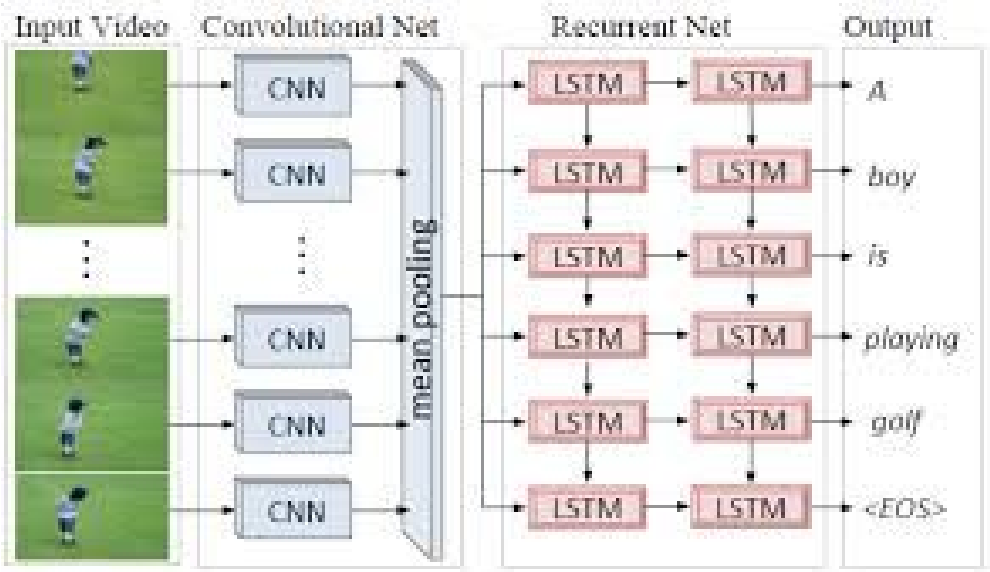

Figure 2: RCNN architecture for video input

Data processing

The data processing steps, in this case, will first include transforming each video in a sequence of frames. This can be stored in a tensor. For instance, if using 256x256 frames for a 10 seconds video with 60 frames per seconds, that would be 600 frames and the tensor representation of that video would be 600x3x256x256. 600 represents the number of frames, 3, the number of channels (here RGB so red, green and blue) and 256x256 the dimension of each frame.

The video might have to be segmented in visemes if we want to use HMMs, which will require an extra preprocessing step.

DATASETS

The data we want to train, validate and test our model on come from the LRW(Lip Reading in the wild), LRS2 (Lip Reading Sentence 2), and LRS3 (Lip Reading Sentence 3) audio-visual speech recognition datasets collated by University of Oxford.

Figure 3: LRW(BBC), LRS2(BBC), and LRS3(TED)datasets.

The first dataset, LRW, contains up to 100 utterances of 500 different words by different speakers. The metadata of each video in this dataset is provided which can be used to determine the start and the end time of the video. The LRS2 dataset contains 1000s of natural sentences from the British Television with each sentence having more at least 100 characters. The LRS2 training, validation and test sets are divided according to date they were broadcasted. Finally, the LRS3 comprises 1000s of natural sentences from TED and TEDx

videos. Each of these datasets has a train, validation and test sets. The statistics of the LRW, LRS2 and LRS3 datasets are summarized in the tables in the appendix.

The training and validation datasets.

The training data for our model will be the combination of the training and validation data in all the three datasets described above. This will ensure that our model gets enough data to learn from in order to generalize well (predict correctly). To get the sense of how well our modeling is training, we will use a 10-fold cross-validation technique to compute the average errors over all the 10 sub-dataset or validation sets. The errors compute in using this technique will be reported as the model training accuracy.

Test dataset.

In all the datasets, some portion of them is set aside as test data. The testing data for the project is a combined test data from all the test data.

Target Goals by Milestone Due Date

The following are the expectations the group seek to accomplish by the milestone due date.

1. Figured out the model to use (from the bunch of many capable models) for lipreading.
2. Processed and cleaned our dataset.
3. Developed a well tested model using the algorithms discussed above.

APPENDIX

Set	Dates	# class	# per class
Train	01/01/2010 - 31/08/2015	500	800-1000
Validation	01/09/2015 - 24/12/2015	500	50
Test	01/01/2016 - 30/09/2016	500	50

Figure 4: The statistics of the LWR dataset

Set	Dates	# utterances	# word instances	Vocab
Pre-train	11/2010-06/2016	96,318	2,064,118	41,427
Train	11/2010-06/2016	45,839	329,180	17,660
Validation	06/2016-09/2016	1,082	7,866	1,984
Test	09/2016-03/2017	1,243	6,663	1,698

Figure 5: The statistics of the LRS2 dataset.

Set	# videos	# utterances	# word instances	Vocab
Pre-train	5,090	118,516	3.9M	51k
Trainval	4,004	31,982	358k	17k
Test	451	1,452	11k	2,136

Figure 6: The statistics of the LRS3 dataset.

Reference

Chung, J. S., Senior, A., Vinyals, O., & Zisserman, A. (2016). Lip Reading Sentences in the Wild. *arXiv:1611.05358 [Cs]*. Retrieved from http://arxiv.org/abs/1611.05358

Google's AI can now lip read better than humans after watching thousands of hours of TV - The Verge. (n.d.). Retrieved February 8, 2019, from https://www.theverge.com/2016/11/24/13740798/google-deepmind-ai-lip-reading tv

Hassanat, A. B. A. (2011). Visual Speech Recognition. *arXiv:1409.1411 [Cs]*. https://doi.org/10.5772/19361

The Challenges and Threats of Automated Lip Reading - MIT Technology Review. (n.d.). Retrieved February 8, 2019, from https://www.technologyreview.com/s/530641/the-challenges-and-threats-of-automated-lip-reading/

VGG Lip Reading datasets. (n.d.). Retrieved February 8, 2019, from http://www.robots.ox.ac.uk/~vgg/data/lip_reading/

Yargıç, A., & Doğan, M. (2013). A lip reading application on MS Kinect camera. In *2013 IEEE INISTA* (pp. 1–5). https://doi.org/10.1109/INISTA.2013.6577656