SRS document for final project:

Automated Lip Reading

**Members:**

Tal Noam - 204397715

Evgeny Vendrov - 313962193

Amitai Zamir - 204324305

**Moderator:**

Lee-Ad Gottlieb

**Table of contents:**

1. Introduction
2. Dataset description
3. Architecture, tools and concept
4. Future steps

Date:20/01/2020

1. Introduction

In this document we will describe our project, the dataset we will use, the method of operation, tool and technologies, the idea behind the architecture, and future steps and conflicts.

2. Dataset description

We will use dataset - which is a combination of several datasets we obtained from several sources.

The dataset consists of high-quality videos - facial recordings of sentences spoken by various talkers.

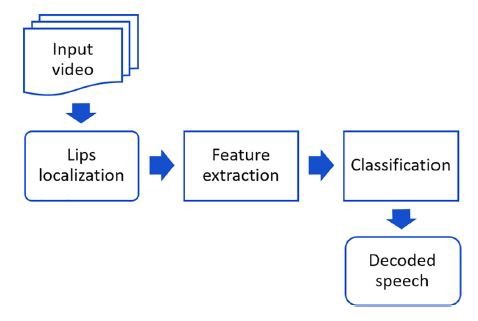
Sentences are of the form "put red at G9 now" i.e. a lot of different lip movement, different words and sounds "made".

The dataset also contains the transcriptions for the sentences the talker said in specific video.

License wise - the datasets, together with transcriptions, is freely available for research use.

\*Note: we sent an email to [rob.cooper@bbc.co.uk](mailto:rob.cooper@bbc.co.uk), and ask to use a database they use for a similar research at the University of Oxford. Hopefully we will use it as well.

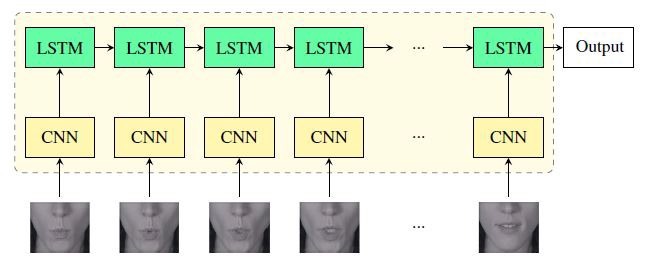
3. Architecture, tools and concept

This image below demonstrates the idea of the system we're about to develop.

The first two steps as seen in the chart – lips localization and feature extraction are a computer vision problem while classification and outputting text are an NLP problem, so we'll combine two sorts of models, one to deal with videos and pictures and the other to deal with the raw textual data received from the video – outputting it as a correct English sentence.

The whole model will receive a video as input, the first step will be the lips localization and detection, this problem is already solved and researched – so we will use a mixture of open source tools-based on CNN as the open CV library, to deal with it.

Second step – will be extracting relevant features from the frames of lips we obtained from video, this is tricky – we don’t know yet which parameters of the speech will be used as features for the NLP component.

Third step - will be taking the output raw-data from the first component as some kind of formatted text data (as csv), classify it to probably syllables, and then use it as feature for the second NLP, LSTM based component.

This sort of network matches our dataset as described earlier, it will require a video as input and text as output.

Loss function will be some sort of MSE between what been actually said in the video and what the model predicted.

4. Future steps

\*we are standing here:

A. We checked several databases and create a dataset which contains 33k videos and transcripts for now, this is still growing, i.e. for this point, it seems we have enough of data.

B. We managed to detect lips in several human face still images using open CV library, we also managed to output only the lips as separated image.

C. We have studied many researches in the field of ALR using deep learning methodologies.

**\*what next?**

A. Figure out which features will be needed for second part of the model, i.e. which parts of the speaking – lips movement wise are needed to effectually "learn" what has been said in the video, we assume these will be pictures focusing only on the lips, the question is which amount of pictures is needed per frames\second.

B. Next step will be extracting these features mentioned in A from the video, we already managed to use OpenCV for this purpose on still images, so extracting relevant frames of the lips in certain frame pace shouldn't be a problem.

C. Next, is figuring out in which format these features will need to be to work with the NLP component effectually, we assume we'll have to represent these lips images with text data as some sort of CSV for example, this is also a question of which NLP method will be chosen for this mission, what brings us to:

D. Figure out which NLP method will be used for this sort of job, we assume that we'll classify the lips movement data to syllables – this is already has been done so we know it's possible, we seen researches in which they managed to classify mouth movement – as images to syllables with CNN(VGG blocks) based DNN methods.

E. Next step will be combining these syllables to words and phrases – this is purely NLP; we assume that knowing which NLP tool we'll use will answer this question.