GuNiLeo: Lip Reading From Videos with STCNN

Beray Nil Atabey (*2045576*) Leonardo Biason (*2045751*) Günak Yuzak (*2048950*)

***Abstract*—Lip reading is a task that can have various usages as an accessibility feature, but it's also very complex to design: it requires a machine to be able to differentiate between the various words said by a speaker, and also to predict what the speaker said whenever words aren't spelled with a precise motion of the lips. With this paper, we propose a model based on a Spatio-Temporal CNN, capable of reading the words said by a speaker from a video clip of maximum 75 frames.**

I. Introduction

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

II. Implementation

In order to create a lip reading model, the following steps have been undertaken:

1) creation and modeling of the dataset;
2) creation of the model;
3) training of the model;
4) evaluation of the model.

A. Creation and Modeling of the Dataset

A dataset for such task

The dataset comprehends two fundamental parts: the data part and the labels. Since the project aims to recognise the movement of the lips, the data part is composed of multiple clips, of different length, of human speakers saying some words while focusing the video on their lips. The labels contain the phonetic representation of the words said by the speaker.