LIP READING

deep learning project

BY:
MARIAM MAHMOUD
NAWAL SHEHATA
SARAH ELZAHABY

CONTENT

WHY THERE'S A NEED TO HAVE A LIP READING?

01

Communication is a must.

02

listeners with hearing impairments.

03

deep learning analyzes visual patterns of lip movements.

04

Build algorithm to transcribe spoken words.

OBJECTIVE

01

Develop an automatic feature extraction techinque that is able to extract lip geometry information from the mouth region.

02

Design a state-of-art audio-visual speech recognition system using dynamic geometry features from the lip shape.

03

Create powerful tool capable of detecting the face, lips and describe the events of video.

04

Design a model that has mush higher accuracy compared to other.

PROBLEM STATEMENT

Speech recognition may not work well if the user has a loud voice, a strong accent, or a speech condition.

when the user in the public location, a quite library or a private conference, voice recognition may not be possible or desired.

Traditional machine learning models may struggle to handle large amounts of data and exploit its potential due to limitations in scalability and computational efficiency.

Speakers exhibit unique lip movements, making it challenging for traditional machine learning models to adapt to speaker-specific variations.

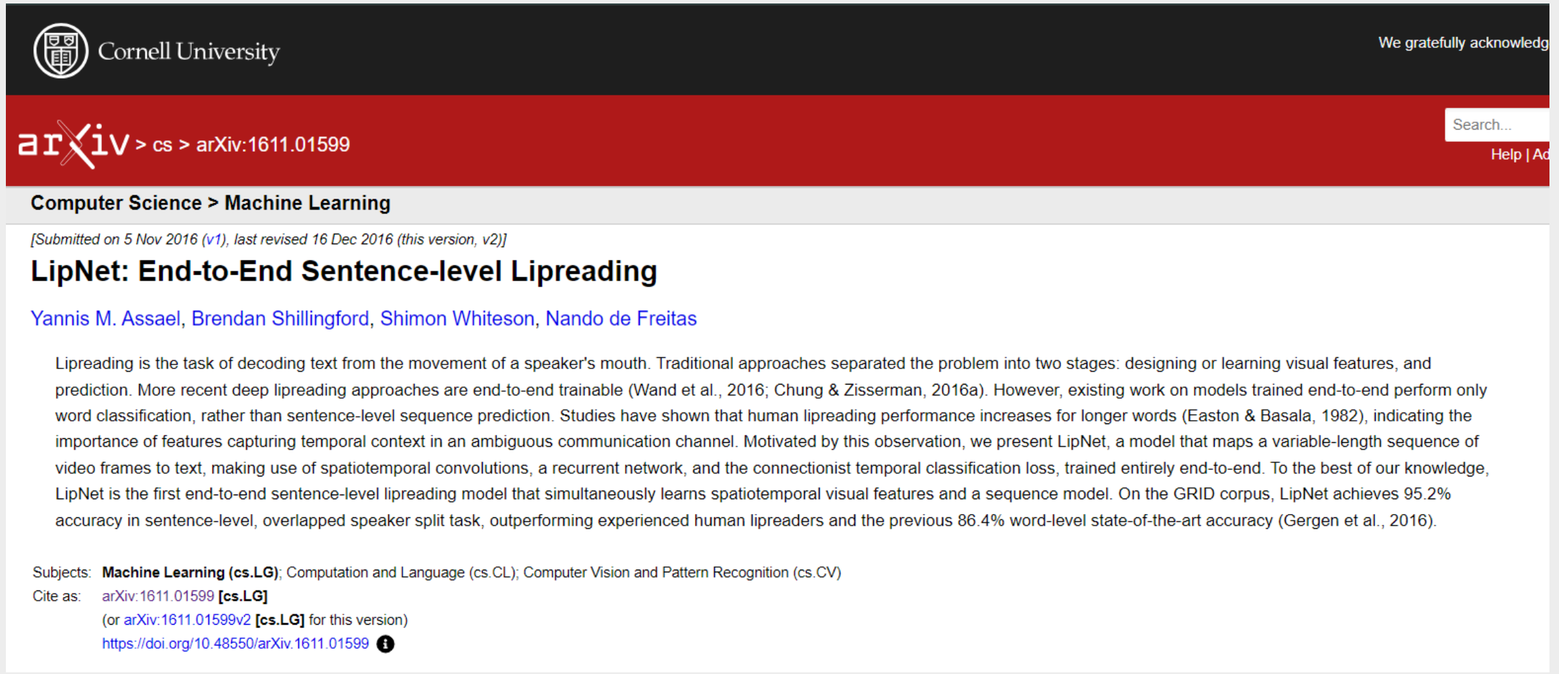
ALL THESE LIPS CAN READ!!

Sl. No.	TITLE of the Paper	AUTHORs, YEAR & PUBLICATION	OBSERVATIONS
1.	Lip Reading Sentences Using Deep Learning with only Visual Cues	Souheil Fenghour, London South Bank University, London 2020	In this Paper, A neural network-based that is Visemes classification, lip reading system has been developed to predict sentences covering a wide range of vocabulary in silent videos from people speaking, they have used BBC LRS2 data set and achieved accuracy of 65%.
2.	Analyzing lower half facial gestures for lip reading applications: Survey on vision techniques	Niranjana Krupa B. Department of ECE, PES University, Bangalore, 560085, India 2022	Lip reading applications discussed in the survey can be of a great helping hand to society by providing security to systems, voice assistants for devices, a hearing aid for deaf people, for generating video text transcriptions, for pronunciation correction or evaluation, aiding forensics with spy cameras, synthesizing voice for unable to talk patients, attempting speech inpainting during noisy video conferencing,

Sl. No.	TITLE of the Paper	AUTHORs, YEAR & PUBLICATION	OBSERVATIONS
3.	Text Extraction through Video Lip Reading Using Deep Learning	S. M. M. H. Chowdhury, M. Rahman, M. T. Oyshi and M. A. Hasan. Moradabad, India, 2021	In this research, a method of converting video data to text data through lip reading has been proposed. The proposed method includes test dataset, image frame analysis and having text output from identified words. They have used HMM Architecture, and used Grid data set the accuracy has been around 76%.
4.	Deep Learning based Lip-Reading Techniques	S. Pujari, S. Sneha, R. Vinusha, P. Bhuvaneshwari and C.Yashaswini. Tirunelveli, India, 2021	In this Paper, convolution Neural Network and Bi-LSTM are used to design Lip reading Model and they have used Ou lu VS2 dataset for training the model they have achieved 82.3% of accuracy.
5.	Lipreading Using Temporal Convolutional Networks	B. Martinez, P. Ma, S. Petridis and M. Pantic Barcelona, Spain, 2020	Firstly, they use Temporal Convolutional Networks (TCN) and used LRW1000 data set. Secondly simplify the training procedure, which allows to train the model in one single stage. Thirdly, to show that the current state-of-the-art methodology produces models that do not generalize well to variations on the sequence length, and addresses this issue by proposing a variable-length augmentation. The

Sl. No.	TITLE of the Paper	AUTHORs, YEAR & PUBLICATION	OBSERVATIONS
6.	Convolutional Neural Network Based Lip Reading System for Hearing Impaired People	Fathima S, C. Jayanthi, N.Sripriya 2020.	Each frame passes through trained CNN architecture and frames are then divided into visemes. Produced visemes go through a thick layer of Long Short Term Memory (LSTM). The result of the LSTM layer turns into the contribution to the following thick layer. Finally, they receive sequence of visemes; classified visemes are labeled by LSTM softmax activation function. Feature extraction of visemes are judged using classifier schema known as visemes to phoneme mapping. Considering the mapping procedure, possible Word is detected using word detector. The accuracy achieved is 83%.
7.	Lip Reading Experiments for Multiple Databases using Conventional Method	T. Shirakata and T. Saitoh. Hiroshima, Japan, 2019	In this paper, not the latest deep learning-based method but the standard recognition method by hidden Markov model (HMM) which mainly used conventionally is applied, and analyzes trends in recognition accuracy. Based-on recognition experiments, it was found that the recognition accuracy was correlated with the number of frames,which is around 91%.

Sl. No.	TITLE of the Paper	AUTHORs, YEAR & PUBLICATION	OBSERVATIONS
8.	Vision based Lip Reading System Using Deep Learning	Fathima S, C. Jayanthi, N.Sripriya 2020.	This paper presents the method for Vision based Lip Reading System that uses convolutional neural network with attention-based Long Shot-Term Memory. The data sets includes video clips pronouncing single digits. They used two pre-trained models namely VGG19 and ResNet50, The system provides 85% accuracy.



Cornell University

We gratefully acknowledg

arXiv > cs > arXiv:1611.01599

Search...

Help | Ad

Computer Science > Machine Learning

[Submitted on 5 Nov 2016 (v1), last revised 16 Dec 2016 (this version, v2)]

LipNet: End-to-End Sentence-level Lipreading

Yannis M. Assael, Brendan Shillingford, Shimon Whiteson, Nando de Freitas

Lipreading is the task of decoding text from the movement of a speaker's mouth. Traditional approaches separated the problem into two stages: designing or learning visual features, and prediction. More recent deep lipreading approaches are end-to-end trainable (Wand et al., 2016; Chung & Zisserman, 2016a). However, existing work on models trained end-to-end perform only word classification, rather than sentence-level sequence prediction. Studies have shown that human lipreading performance increases for longer words (Easton & Basala, 1982), indicating the importance of features capturing temporal context in an ambiguous communication channel. Motivated by this observation, we present LipNet, a model that maps a variable-length sequence of video frames to text, making use of spatiotemporal convolutions, a recurrent network, and the connectionist temporal classification loss, trained entirely end-to-end. To the best of our knowledge, LipNet is the first end-to-end sentence-level lipreading model that simultaneously learns spatiotemporal visual features and a sequence model. On the GRID corpus, LipNet achieves 95.2% accuracy in sentence-level, overlapped speaker split task, outperforming experienced human lipreaders and the previous 86.4% word-level state-of-the-art accuracy (Gergen et al., 2016).

Subjects: **Machine Learning (cs.LG)**; Computation and Language (cs.CL); Computer Vision and Pattern Recognition (cs.CV)
Cite as: arXiv:1611.01599 **[cs.LG]**
(or arXiv:1611.01599v2 **[cs.LG]** for this version)
https://doi.org/10.48550/arXiv.1611.01599

HOW TO MAKE A LIP READ?

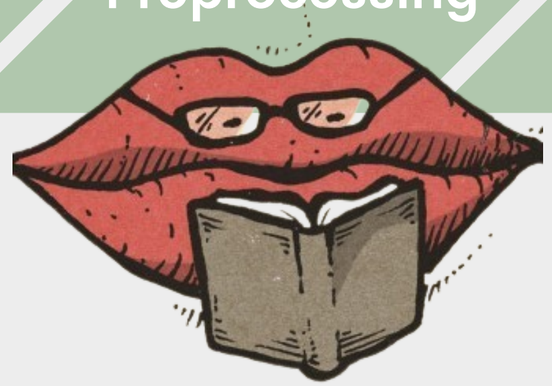
Videos and Alignments Data

Feature Extraction

Data Preprocessing

Model Architecture

Output

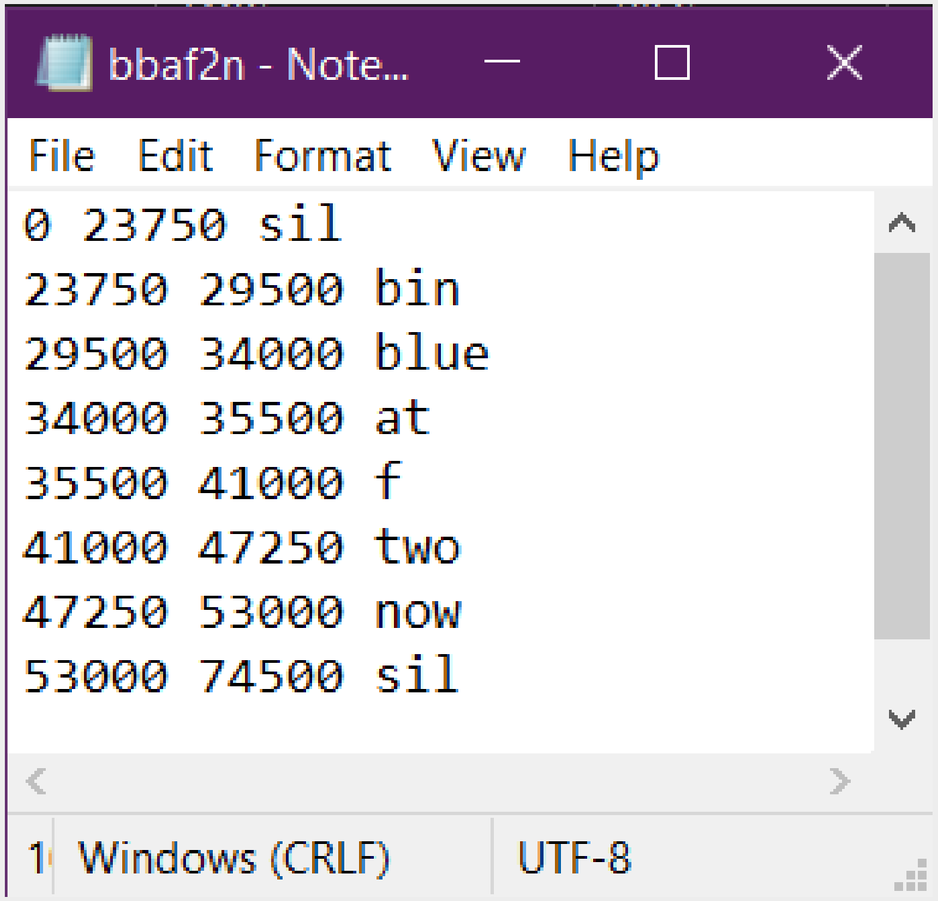


HOW TO MAKE A LIP READ?

Videos and Alignments Data

HOW TO MAKE A LIP READ?

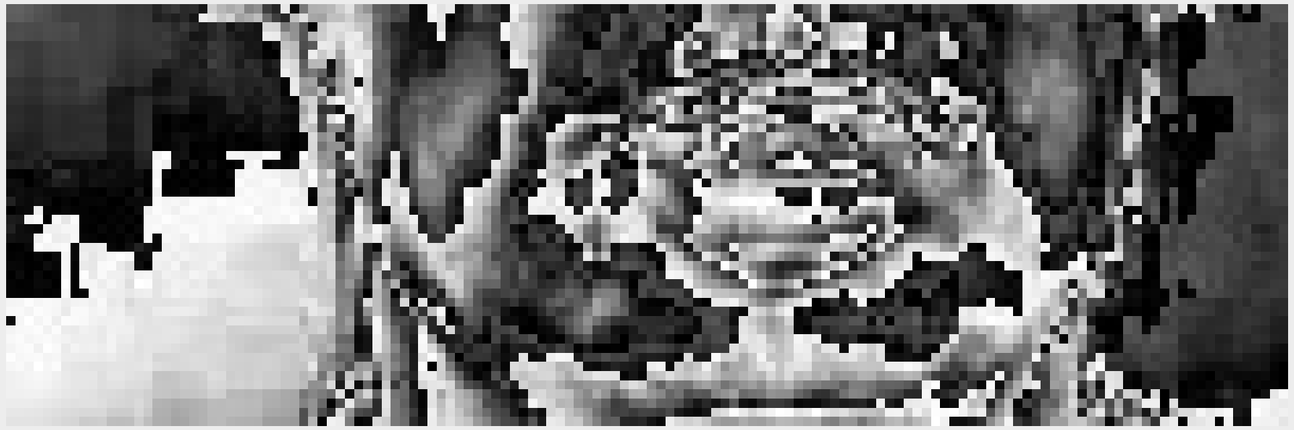
Videos and Alignments Data



HOW TO MAKE A LIP READ?

Videos and Alignments Data

Feature Extraction

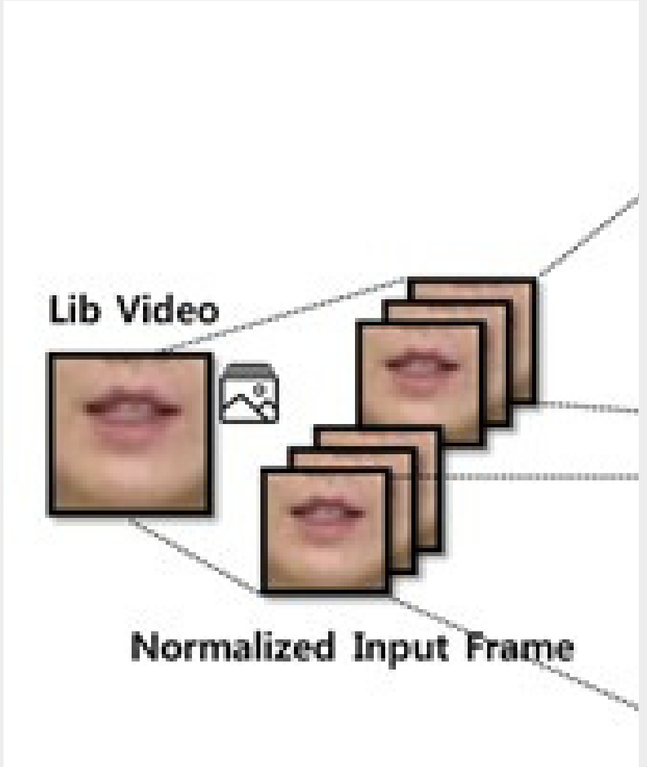


HOW TO MAKE A LIP READ?

Videos and Alignments Data

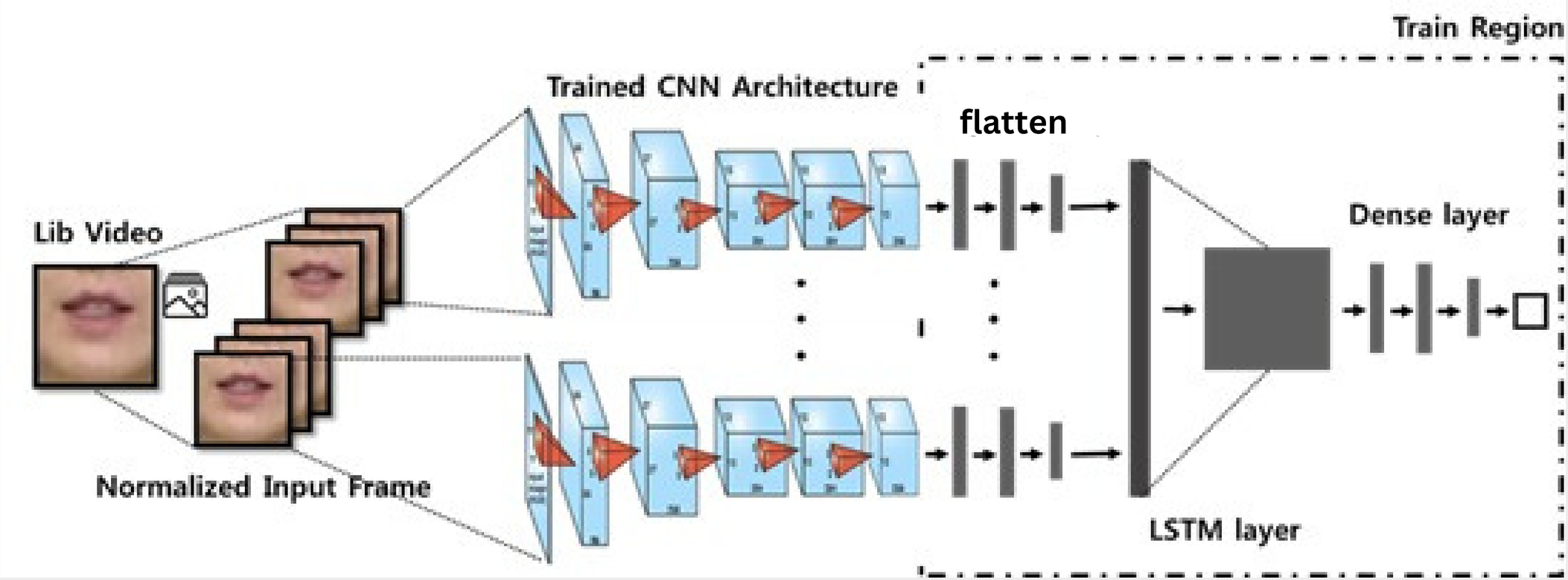
Feature Extraction

Data Preprocessing



HOW TO MAKE A LIP READ?

Model Architecture



HOW TO MAKE A LIP READ?

HOW TO MAKE A LIP READ?

```
[123] yhat = loaded_model.predict(tf.expand_dims(sample, axis=0))

1/1 [==============================] - 8s 8s/step

[124] decoded = tf.keras.backend.ctc_decode(yhat, input_length=[75], greedy=True)[0][0].numpy()

[125] print('~'*100, 'PREDICTIONS')
      [tf.strings.reduce_join([num_to_char(word) for word in sentence]) for sentence in decoded]

~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ PREDICTIONS
[<tf.Tensor: shape=(), dtype=string, numpy=b'bin blue at f two now'>]
```

HOW TO MAKE A LIP READ?

Videos and Alignments Data

Feature Extraction

Data Preprocessing

Model Architecture

Output