LipNet:

A comparative study

Vyom Jain, Srishti Lamba, Shweta Airan

Institute of Technology, Nirma University

Abstract

This paper is an attempt to discuss and explore various approaches in the field of lip reading. With special focus on LipNet, which has served as a landmark paper in this field, the paper discusses the various approaches and architectures proposed in the art. Understanding the differences and the similarities between these architectures will help us to have a deep understanding of the state of lip-reading.

We will witness how the models have changed thus far and how other researchers influence others to create something new or to improve the existing one. We will also this through a timeline starting from 2015, with Recognition of spoken English phrases using visual features extraction and classification, going all the way to 2017, with the latest improvement LipVision.

1. Introduction

Lip reading has not been easy for humans. Accuracy of someone able to comprehend speech based only on the movement of lips has been very low. Deep learning on the other has evolved and is able to understand common patterns in lip movements to judge the actual speech. Audio speech recognition has already evolved to near human accuracy, the same is about to be demonstrated for lip reading.

Applications of this technology are far and widespread. Apart from the surveillance domain, this can help those with hearing difficulties to figure out what other are saying with transcripts for each input as the person is seeing lip movement. Human efficiency is at 52.3 percent, almost all of the papers we discuss have their accuracy more than this. The LipNet has an accuracy of 93.4 percent in some tests.

This paper will first summarise the papers under consideration, and then present a comparative study for all of them.

2. Various Approaches

This section will cover some state-of-the-art approaches and architectures in the field of lip reading. Starting from our main focus, LipNet, we will discuss approaches which were there earlier and which came after LipNet.

2.1 LipNet

This paper has served as a landmark approach for lip reading. The model uses the grid corpus dataset with some augmentation and pre-processing and is able to give accuracy scores well above human capabilities.

This is the first end-to-end sentence level lip reading model. It operates at character level. Uses concepts like spatiotemporal convolutional neural networks (STCNN), Gated Recurrent Unit (GRU) and Connectionist Temporal Classification (CTC)loss to provide accuracy of 95.2% accuracy on the GRID corpus dataset, with 88.6% accuracy on unseen speakers. The most impressive feature is the use of Saliency Visualization techniques, which ensures that the model attends to phonologically important areas in the video. Almost all the erroneous predictions are due to insufficient context for disambiguation.

STCNN is based on the much-known CNN. It convolves across time as well as spatial dimensions.

$$[\mathrm{conv}(\mathbf{x},\mathbf{w})]_{c'ij} = \sum_{c=1}^{C}\sum_{i'=1}^{k_w}\sum_{j'=1}^{k_h} w_{c'ci'j'}x_{c,i+i',j+j'},$$

$$[\mathrm{stconv}(\mathbf{x},\mathbf{w})]_{c'tij} = \sum_{c=1}^{C}\sum_{t'=1}^{k_t}\sum_{i'=1}^{k_w}\sum_{j'=1}^{k_h} w_{c'ct'i'j'}x_{c,t+t',i+i',j+j'}.$$

GRU is a type of Recurrent Neural Network (RNN). It improves upon earlier RNNs by using gates and cells for propagating information over more timestamps.

$$\begin{aligned}
[\mathbf{u}_t,\mathbf{r}_t]^T &= \mathrm{sigm}(\mathbf{W}_z\mathbf{z}_t + \mathbf{W}_h\mathbf{h}_{t-1} + \mathbf{b}_g)\\
\tilde{\mathbf{h}}_t &= \tanh(\mathbf{U}_z\mathbf{z}_t + \mathbf{U}_h(\mathbf{r}_t \odot \mathbf{h}_{t-1}) + \mathbf{b}_h)\\
\mathbf{h}_t &= (\mathbf{1}-\mathbf{u}_t)\odot\mathbf{h}_{t-1} + \mathbf{u}_t\odot\tilde{\mathbf{h}}_t
\end{aligned}$$

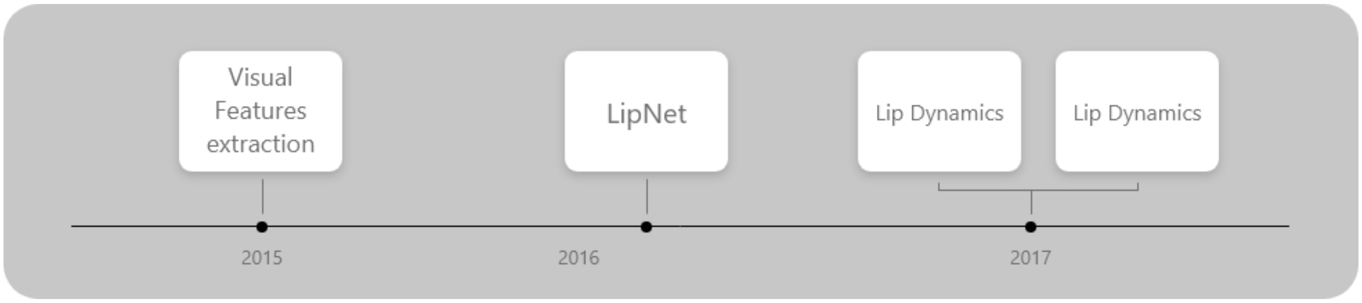

Figure 1: Lip reading a journey through time.

CTC loss is used to eliminate the need for the training data to have input aligned with target outputs. CTC computes the probability of the sequences by marginalising over all sequences that are defined as equivalent to this sequence. This eliminates the need of alignments and addresses variable length sequences.

The resulting architecture has 3 STCNNs, 2 Bi-GRUs, linear transformation is applied at each time-step, followed by SoftMax over vocabulary augmented with CTC blank. Then CTC loss. All the layers use ReLU as the activation function.

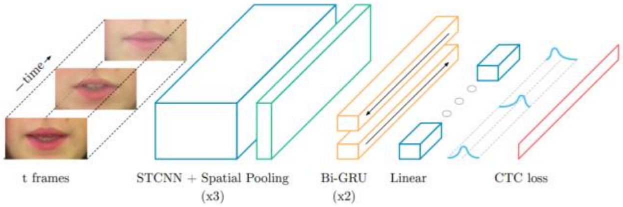

Figure 2: LipNet Architecture.

2.2 LipVision

2.2.1 Face and Mouth Detection-

Face and mouth are detected by using Haar Cascade approach. The cascade function is trained on positive and negative images where the best features are selected after the analysis of error rate. Minimum error rate differentiates between face and non-face images. Adaboost which includes mouth ROI is used to find the best features. The whole process selects a region which will be processed further.

2.2.2 Dataset- Grid Corpus

Here, we use Grid Corpus dataset which has size of 15.6 GB and has 51 distinct words. It has 25 frames per second, resolution of 360x288, bitrate of 1000-1200 kbps, 33 speakers, 3300 videos of 3 second each.

2.2.3 Facial Landmarks

We use Ibug tool to mark facial landmarks. It is used for reading the lips so that points on the lips can be extracted to match them with the points acquired from trained dataset.

2.2.4 Classifier- CNN (Convolutional Neural Network)

Convolutional Neural Networks (CNN) comprises of a perplexing development of cells which respond to little sub-districts of the responsive field. The whole open field is secured by the sub-locales. These divisions locally refine the information space and are fitting to use the solid spatially nearby reliance present in natural images.

OpenCV is used to read images and sends them to multidimensional Tensor to reshape them with respect to the requirements of the system.

2.2.5 Classifier Training – TensorFlow

We will nourish TensorFlow with raw inputs and each of the factors for this information will have a remarkable weight which will at that point be gone through an entirety work. This is then passed through the threshold function to check on the off chance that it may be passed through the following layer.

It assigns value of 1 if the neuron fires, otherwise the value of 0. This passes through many hidden layers which then evaluates the input against the output and communicates back to the system if any error occurs so that the weights can be adjusted.

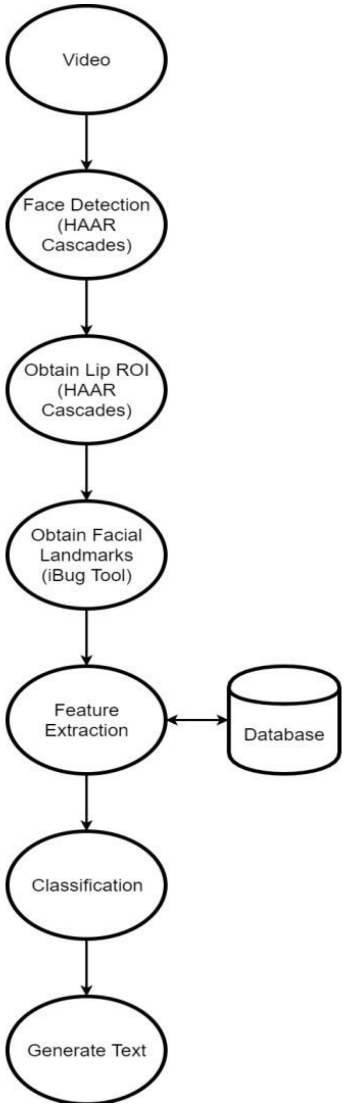
Video
Face Detection (HAAR Cascades)
Obtain Lip ROI (HAAR Cascades)
Obtain Facial Landmarks (iBug Tool)
Feature Extraction
Database
Classification
Generate Text

Figure 3: Flow Chart for Proposed System

2.3 Visual features extraction and classification

This paper makes use of 10 speakers, each speaking ten phrases and each phrase is spoken six times by a particular speaker. Frames are taken out from the video; the first step is to identify the speaker's face using face recognition algorithm. The main aim is to recognize what the speaker is saying, merely by Lip movement. Then comes the most crucial step that is to mark the Lip using four points for left, right, upper and lower lip. Features are extracted on the basis of the change in position of the lips while speaking. The first feature is P1 which is the angle between the left, upper and right point marked on the lips. Similarly, we can find the other three features.

Angle P1 = tan-1 [(m2-m1)/ (1+m1*m2)]

Angle P2 = tan-1 [(m3-m2)/ (1+m2*m3)]

Where

m1= Δ [Left (x, y), Upper (x, y)]

m2= Δ [Left (x, y), Lower (x, y)]

m3 =Δ [Upper (x, y), Right (x, y)]

Feature vectors are trained using SVM (support vector machine). SVM is trained using 300 training dataset and then forming a confusion matrix, we then observe that phrase "welcome" has lowest accuracy. This provides an overall accuracy of 65.6%

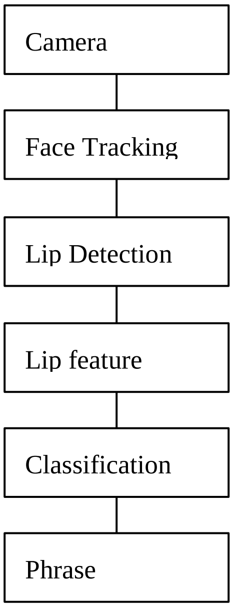

Figure 4: System Diagram.

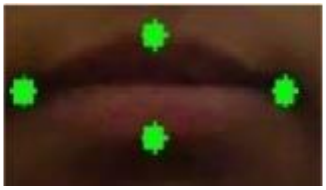

Figure 5: Lip points marked.

2.4 Estimating speech from lip dynamics

2.4.1 Feature Extraction

The first step is to determine the general location of the mouth in the frame. After identifying the initial mask, we convert the image into grayscale and apply active contouring and edge detection. By using K-means algorithm, the lip region is separated from the face.

2.4.2 Extracting Phonemes

In audio speech recognition phonemes (sounds that carry distinction on the basis of language) are detected where as in visual speech recognition only visemes (several sound that look same) are detected. The following table has been used to map phoneme and visemes.

Viseme Number	Viseme Label	Associated Phonemes
1	P	b p m
2	T	d t s z th dh
3	K	g k n l y hh
4	CH	jh ch
5	F	f v
6	W	r w
7	IY	iy ih
8	EH	eh ey ae
9	AA	aa aw ay ah
10	AO	ao oy ow
11	UH	uh uw

Table 1: Phoneme to Viseme Map from Lee and York, 2000, via [5].

2.4.3 Assigning Phonemes

A file which has a collection of words spoken associated with each frame is created. Words are deconstructed and assigned to each frame to form training data.

2.4.4 Classification using HMM

Classification of phoneme to corresponding visemes to create labels for classification algorithm is done. Now, phoneme and visemes are mapped to words with the help of HMM. For phoneme is mapped to a sequence number between 1 and 37, for viseme the sequence number ranges from 1 to 11. Highest accuracy of 87.5% is achieved for the word "bin"

3. Comparison

Name	Architecture used	Accuracy
Lip Net	STCNN, GRU, CTC loss	95.2 on GRID corpus
Visual features extraction and classification	SVM (support vector machine)	10 Different Speakers
Lip Vision	CNN (Convolutional Neural Network)	Not mentioned for the proposed approach
Lip Dynamic	Classification Algorithm and Hidden Markov Model	GRID corpus

Table 2: A comparison of all the discussed papers.

4. Conclusion

Lip reading is the task of comprehending text from the movement of lips of the speaker. Lip reading is an extremely difficult task for humans, especially in the absence of context. The approaches that we discussed in this paper were all an attempt to excel beyond the human performance. Having a comparative study helps us to get familiar with the technologies in the art and also to get a better idea of the problem at hand. We also get to see the evolution of these technologies over the years.

References

[1] Yannis M. Assael, Brendan Shillingford, Shimon Whiteson and Nandode Freitas, "*Lipnet: End-to-end sentence-level lipreading*", arXiv > cs > arXiv:1611.01599, 2016.

[2] Salma Pathan and Archana Ghotkar, "*Recognition of spoken English phrases using visual features extraction and classification*", International Journal of Computer Science and Information Technologies, Vol. 6 (4), 3716- 3719, 2015.

[3] Parth Khetarpal, Riaz Moradian, Shayan Sadar, Sunny Doultani, Salma Pathan, "*LipVision: A Deep Learning Approach*", arXiv > cs > arXiv:1708.01198, 2017.

[4] I. Almajai, S. Cox, R. Harvey, and Y. Lan. "*Improved speaker independent lip reading using speaker adaptive training and deep neural networks*." In IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 2722–2726, 2016.

[5] L. Cappelletta and N. Harte. "*Phoneme-to-Viseme Mapping for Visual Speech*." Department of Electronic and Electrical Engineering, Trinity College Dublin, Ireland. May 2012.