ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal
Regular Issue, Vol. 8 N. 3 (2019), 13-26
eISSN: 2255-2863
DOI: http://dx.doi.org/10.14201/ADCAIJ2019831326

LSTM Based Lip Reading Approach for Devanagiri Script

Mahesh S Patil[a], Satyadhyan Chickerur[b], Anand Meti[a], Priyanka M Nabapure[a], Sunaina Mahindrakar[a], Sonali Naik[a] and Soumya Kanyal[a]

[a] School of Computer Science and Engineering, KLE Technological University, Hubballi, India, 580031
[b] Center for High Performance Computing, KLE Technological University, Hubballi, India, 580031
mahesh_patil@kletech.ac.in, chickerursr@kletech.ac.in, anandsmeti@kletech.ac.in, priyanka09m.n@gmail.com, sunainaa9@gmail.com, naiksonu1998@gmail.com, kanyals98@gmail.com

KEYWORD

Lip-reading; LSTM; Machine learning; Deep Learning; Feature extraction

ABSTRACT

Speech Communication in a noisy environment is a difficult and challenging task. Many professionals work in noisy environments like aviation, constructions, or manufacturing, and find it difficult to communicate orally. Such noisy environments need an automated lip-reading system that could be helpful in communicating some instructions and commands. This paper proposes a novel lip-reading solution, which extracts the geometrical shape of lip movement from the video and predicts the words/sentences spoken. An Indian specific language data set is developed which consists of lip movement information captured from 50 persons. This includes students in the age group of 18 to 20 years and faculty in the age group of 25 to 40 years. All have spoken a paragraph of 58 words within 10 sentences in Hindi (Devanagari, spoken in India) language which was recorded under various conditions. The implementation consists of facial parts detection, along with Long short term memory's. The proposed solution is able to predict the words spoken with 77% and 35% accuracy for data set of 3 and 10 words respectively. The sentences are predicted with 20% accuracy, which is encouraging.

1. Introduction

Artificial Intelligence (AI) is better than humans at lots of tasks, from the obvious of data analysis to the unexpected of composing music, gaming, predicting human behavior, translation, face detection, face recognition, and the list goes on. It is the obvious fact that most of these AI achievements rely on object detection, Computer Vision, and High-performance Computing (HPC). Object detection in humans is a complex process, on which we depend. The human visual system is fast and accurate and can perform complex tasks like identifying multiple objects and detect obstacles with little conscious thought. The human brain can identify the things that we view and compares with the details of the things viewed earlier to make distinctions among the viewed things. With this idea of Recognition followed by Comparison and with the availability of large amounts of data, faster GPU, and better algorithms, we can now easily train computers to detect and classify multiple objects, detect

face within an image with high accuracy. With the availability of data and computer vision, it is possible to predict human behavior, for instance: communication, action, expression, speech, which relies on the fact that people are the creatures with habit.

Lip-reading plays a crucial role in human communication and speech understanding. Lip reading is a technique of understanding speech by visually interpreting the movements of the lips, face, and tongue when normal sound is unavailable, which relies on the fact, that the pronunciation of each word is associated with a set of series of lip movements. Lip-reading is a difficult task for humans, especially in the absence of context. In many situations where it is difficult to hear due to a noisy environment, people find lip-reding helps them to understand more of a conversation; and it may be essential for people with profound deafness. Lip-reading can fill in the gaps in noisy places and help in achieving better communication. Most of us have been lip-reading for years without knowing it in noisy constructions, manufacturing sites, pubs, clubs, and places of work or wherever there is background noise. But to get high accuracy, lip-reading needs continuous concentration and can be very tiring, and regular practice is important. Such a practice will help you to maintain your level of lip-reading skill. Professional Lip Readers achieve an accuracy of only 12-15% even for a limited subset of words. Hence, an important goal is to automate lip-reading by developing an application for lip-reading, by transforming lip-syncing to text and thereby help them in achieving better communication. Therefore, lip-reading can be used for achieving successful delivery of the words by the people, in a vociferous environment for achieving better communication.

A series of fixed lip movements are associated with the successful pronunciation of any word. So, the goal is to capture the series of lip movements associated with the word in the suitable form, either image or numeric values. Our application relies on the series of numeric values, representing the lip movement's pattern of a word. The LSTM sequence classification model, with memory units, is used for training 58 unique words and 10 sentences. Based on the key concept of capturing lip-movement of a word into a numeric series, the proposed lip-reading system with an accuracy of 35% transforms the spoken word to textual form.

2. Literature Survey

Audio to speech conversion systems can be used to assist the communication and which is now available as part of many applications being used in day to day life. This speech recognition can perform better with the help of visual information (Potamianos, Neti, Gravier, & Garg, 2003; Potamianos, Neti, & Luettin, 2004). The use of visual information becomes very much necessary in a noisy environment where the audio to speech conversion systems fails (Erber, 1975; Hilder, Harvey, & Theobald, 2009; Ronquest, Levi, & Pisoni, 2010; Sumby, 1954). In such environments, there is a need to build a speech recognizer that can use both audio and facial visual information. It would also be necessary to have a speech recognizer that would solely use visual information and does not depend on audio input (Fernandez-Lopez & Sukno, 2018). i.e., such speech recognizer should extract features from the facial parts, especially lip movement, and facial expressions. Several attempts have been made to detect speech from the lip movement by (Almajai, Cox, Harvey, & Lan, 2016; Chung, Senior, Vinyals, & Zisserman, 2017; Chung & Zisserman, 2017; Dupont & Luettin, 2000; Petridis & Pantic, 2016; Sui, Bennamoun, & Togneri, 2015; Wand, Koutnik, & Schmidhuber, 2016; Yau, Kumar, & Weghorn, 2007; Zhou, Zhao, Hong, & Pietikäinen, 2014), but the results are much low as compared with audio speech recognizers (Fernandez-Lopez & Sukno, 2018).

The main problem with lip-reading systems is the similar lip movement across the many words spoken by humans (Dupont & Luettin, 2000; Ara V. Nefian, Liang, Pi, Liu, & Murphy, 2002; Zhou et al., 2014). Thus, it is difficult to map a particular pattern of lip movements to one spoken word. But some of the authors argue that a sequence of words appearing in a sentence has a pattern and such sequence of words can resolve the mapping of lip movement patterns to a word (Afouras, Chung, & Zisserman, 2018; Assael, Shillingford, Whiteson, & de Freitas, 2016; Chung et al., 2017; Chung & Zisserman, 2017). Therefore, the autonomous lip-reading system with high accuracy remains to be challenging one (Fernandez-Lopez & Sukno, 2018). The other difficulties involved with the lip-reading system include face view angle, person to person differences in lip movements,

poor video resolution and low frame rates of videos (Buchan, Paré, & Munhall, 2007; Hilder et al., 2009; Ortiz, 2008).

The earlier works have used the Markova module to recognize lip movement patterns. The recent works have used Recurrent neural network/ LSTM to do the same. When their results are compared, the Markova modules perform better than LSTM, and it is mainly due to the unavailability of large data repositories (Fernandez-Lopez & Sukno, 2018). All the earlier research works on the lip-reading system have three steps in common (Fernandez-Lopez & Sukno, 2018).

i. Facial and lip part identification.
ii. Lip/ Facial feature extraction.
iii. Sequence classification.

All the earlier works directly take the images as input and use all the pixels of the image as features(Fernandez-Lopez & Sukno, 2018). Thus, these will end up taking a lot of time to train the module in the presence of large data. But the proposed work in this paper uses dlibs("dlib C++ Library," 2019) facial part detection API's to extract lip portions geometrical shape, which is used as features. This work achieves the accuracy which is near to state of the art results. Earlier works have tried to recognize alphabets/ numbers/ words/ sentences spoken by persons, but the chosen set of such words and sentences for training are easily differentiable in terms of lip movement patterns (Fernandez-Lopez & Sukno, 2018). Also, such sentences used for training aren't meaning full when spoken. But, the proposed work uses a set of words and sentences that are taken from a paragraph and meaningful when spoken. i.e., the sentences are grammatically correct and meaningful. Authors didn't deliberately choose words/ sentences that are easily differentiable in terms of lip movement's pattern.

A large data is very much essential to get better accuracy with any of the deep learning modules. There are many speech audio-video data repositories (Afouras et al., 2018; Bailly-Bailliére et al., 2003; Cooke, Barker, Cunningham, & Shao, 2006; Fernandez-Lopez, Martinez, & Sukno, 2017; Huang, Potamianos, Connell, & Neti, 2004; Matthews, Cootes, Bangham, Cox, & Harvey, 2002; Moll & Daniloff, 1971; Patterson, Gurbuz, & Tufekci, 2002; Petridis, Shen, Cetin, & Pantic, 2018) built by previous researchers for lip-reading and its related problems. Most of these data are captured from a few persons whose count is below 100. Many of these works have captured persons speaking digits and/or alphabets (Fernandez-Lopez & Sukno, 2018). While some of these capture words or sentences being spoken, the numbers of words/ sentences are few. i.e below 100. This unavailability of large data to train the lip-reading system makes it difficult to generalize and achieve high accuracy (Fernandez-Lopez & Sukno, 2018). For example, data repository GRID (Cooke et al., 2006) has data captured from 34 persons speaking 51 classes of phrases. The IBMViaVioce (Moll & Daniloff, 1971) data consists of 10k words spoken by 290 persons, but it is not publicly available. The proposed work builds a lip-reading system for the persons who pose right in front of the camera and speak in Hindi ("Hindi - Wikipedia," 2019) language with continuous meaningful speech. Hence, the authors could not use existing available speech audio-video data, which were unsuitable for the proposed work. Therefore authors built their new data repository consisting of 58 unique words combining to form 10 sentences and being spoken by 50 persons. In the following paragraphs, the paper discusses LipNet and Google's DeepMind as two previous related works.

2.1. LipNet

Lipnet is a deep lip-reading system that is end-to-end trainable. The humans can perform better in lip-reading for longer words are spoken rather than shorter words. Therefore lip-reading is ambiguous and capturing the features plays an important role in lip-reading (Easton & Basala, 1982). Motivated by this observation, LipNet (Assael et al., 2016) a model that can identify text from a sequence of video frames of variable number by using spatio-temporal convolutions, a recurrent network, and the connectionist temporal classification loss, trained entirely end-to-end. LipNet is the first end-to-end sentence level lip-reading model that simultaneously learns spatio-temporal visual features and a sequence model. LipNet can identify the sentences spoken with an accuracy of 95.2% on the GRID corpus, outperforming experienced human lip readers. The

drawback of LipNet is that the grammar in GRID's sentences follows the same pattern, and so is far easier to predict (Assael et al., 2016).

2.2. DeepMind

The AI application, Google's DeepMind ("Google's DeepMind AI can lip-read TV shows better than a pro | New Scientist," 2019), is getting its virtual teeth into the challenge and doing an even better job than humans. Google's DeepMind and university of Oxford have trained deep learning module using a large BBC program's dataset to create a lip-reading system. The AI system was trained using some 5000 hours from six different TV programs such as 'Newsnight', 'BBC Breakfast' and 'Question Time'. These videos consisted of 118 thousand sentences overall. The video was correctly synced to its audio, and a computer system was trained to synchronize the sound with the mouth shape. The trained system was then used to figure out how much the feeds were out of sync when they didn't match up and realigned them. It then automatically processed all 5000 hours of the video and audio ready for the lip-reading challenge. The AI system could synch the audio to mouth shape with an accuracy of 46.8 percent. (Assael et al., 2016).

Google's Deep Mind relies on both audio and video. Hence, it is not supportive in a noisy environment. In order to overcome this, and to achieve better communication in a noisy environment, the author's focus is only on the visual features of the video, ignoring the audio of the video, which will be corrupted in the noisy environment. Authors were successful in taking up this challenge on a spoken paragraph (Fig 1), which consists of 76 words (consisting 58 unique words) and could develop a prototype using the libraries majorly, dlib, cv2, Keras, and LSTM. The proposed prototype outperforms professional lip readers, with an accuracy of 35%.

2.3. Dlib

Dlib is library built using C++, which provides machine learning algorithms to be used as tools to develop AI capable applications. Developers can use this library to build applications in various domains ranging in robotics, commercial apps and HPC softwares. It is being used in both academia and as well as industries. Dlib's open source licensing allows us to use it in any application, free of charge ("dlib C++ Library," 2019).

3. Methodology and Implementation

3.1. Training Dataset

आजकल के समय में निबंध लिखना एक महत्वपूर्ण विषय बन चुका है, खासतौर से छात्रों के लिए। ऐसे कई अवसर आते हैं, जब आपको विभिन्न विषयों पर निबंधों की आवश्यकता होती है। निबंधों के इसी महत्व को ध्यान में रखते हुए हमने इन निबंधों को तैयार किया है। हमारे द्वारा तैयार किये गये निबंध बहुत ही क्रमबद्ध तथा सरल हैं और हमारे वेबसाइट पर छोटे तथा बड़े दोनो प्रकार की शब्द सीमाओं के निबंध उपलब्ध हैं।

Figura 1: Paragraph for the dataset.

A new dataset is generated by the authors for the proposed work. The authors recorded videos of 50 people, where each of them spoke a paragraph in the Hindi language given in Figure 1. The authors could not find any

relevant dataset online, nor could they find any suitable videos from which dataset could be generated. The videos found online are not appropriate as the speaker in the video makes many movements, and the lip detection was challenging there. Hence, the authors collected a new dataset through 50 volunteering persons.

The paragraph spoken by 50 people consists of 58 unique words. Most of the previous works use English language speeches as a data set (Assael et al., 2016), and little work is done using regional language speeches. Hence the authors decided to take up the work in the regional language. Firstly, videos of persons speaking the given paragraph are recorded. Then the timestamps of each spoken word in a video are generated by the authors and volunteers, with the help of media players. Table 1 gives a sample timestamps (ts) for the first few words in paragraph spoken by a person. Unit of ts (time stamp) is 1 kilohertz (1000 per second)

The attributes in the training data set are Person, Person-Id, Word, Word-Id, Start -Timestamp, End- Timestamp, The attribute 'Person', reflects the name of the person. The attribute 'Person -Id', reflects unique Id assigned to the Persons from whom the videos are recorded. The attribute 'Word', reflects words in the paragraph being spoken. The attribute 'Word-Id', reflects the unique Id assigned to the words in the considered paragraph. The attribute 'Start-Time-Stamp' (Start ts), reflects the starting time of that word spoken in the video. The attribute 'End-Time-Stamp' (End ts), reflects the end time of that word spoken in the video.

Tabla1: ts of few initial words spoken in a sample video

Person Id	Word	Word-Id	wIdNum	Start ts	End ts
P01	aajkal	w01	1	0	800
	ke	w29	29	900	1200
	samay	w46	46	1200	1600
	mein	w40	40	1600	1900
	nibhandh	w41	41	1900	2300
	likhana	w36	36	2300	2900
	ek	w17	17	2900	3300
	mahatwapurn	w39	39	3300	4200
	vishay	w55	55	4200	4600
	ban	w10	10	4600	4800
	chuka	w13	13	4800	5200
	hein	w21	21	5200	5800
	khastaur	w30	30	5800	7000
	se	w48	48	7000	7300
	chatron	w11	11	7300	7800
	ke	w29	29	7800	8000
	liye	w37	37	8000	8400
	aise	w04	4	8900	9200
	kahi	w28	28	9200	9600

Subsection 3.3 explains dlib, used to detect lip part from each frame of the video, and then generate features from the lip portion. From each frame, the height(feature) of the lip is obtained as a sequence. Later to train, the sequence of heights are split as per the timestamps of words being spoken in a video. The training of data is explained in subsection 3.3 and 3.4.

Lip Reading System is made up of following functional blocks:

a) Video to Frame Converter
b) Fetching Facial Landmark Points
c) Training Data
d) Model.

3.2. Video to Frame Converter

Figure 2: Frames with cropped lip portion.

A given video is converted to frames using the cv2 library and its inbuilt functions. The frame rate of the video is of HD quality i.e 1920x1080 pixels with 30 frames per second. The video is converted to frames, in order to capture each lip movement associated with the words being spoken in the video. The generated frames (JPG file format) are given as input to the Bounding Box Algorithm of dlib, used for lip detection. Sample frames are given in Figure 2.

3.3. Fetching Facial Landmark Points

3.3.1. *Lip Detector*

The generated frames by the 'video to frame converter' are the input for Lip Detector. Firstly, the Lip Detector detects the face of the person in the frame. Secondly, the lip in the detected face is detected, as shown in Figure 3. The facial landmark detector of dlib is used for detecting the face and lip portion in a given frame. The supplementary libraries used are imutils, numpy, argparse, cv2, and glob. The input frame is first converted to a grayscale image using the cv2 library. The facial feature edges are clearer to visualize in the grayscale image. The facial landmark detector of dlib produces 68(x, y)-coordinates on the grayscale image as shown in Figure 4, that map to specific facial structures. This grayscale image and facial landmark points coordinates are given as an input to the lip height calculator explained in below subsection 3.3.2.

Figure 3: Lip Detection.

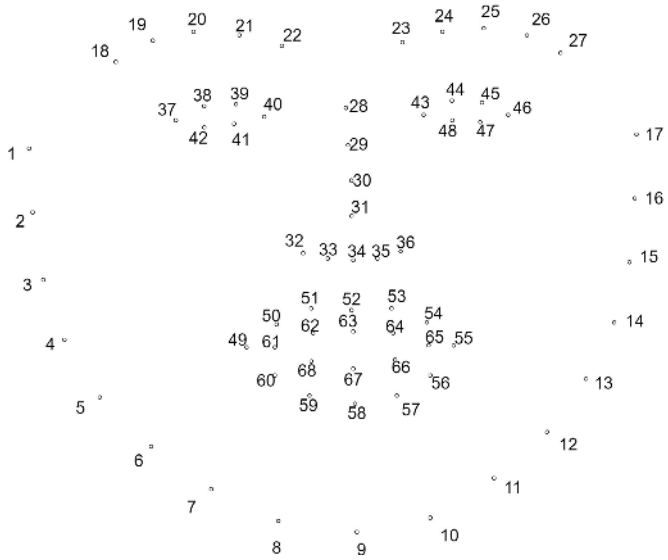

Figure 4: Visualization of 68 facial landmark points.

3.3.2. *Lip Height Calculator*

The grayscale image and 68 facial landmarks coordinates are the input to the lip height calculator. The facial landmark points 49-68 shown in Figure 4 corresponds to lip-region, the region of interest in this scenario. Firstly, the 'lip height calculator' loops to these 49-68 facial landmark points and then captures the region as a set of coordinates. Secondly, it calculates the relative x, y coordinates of these points with respect to the captured lip-region. This is done for all the frames of each word. The set of coordinates depicts the geometrical shape of the lip portion from which heights or/and width of the lip can be extracted and further used as features to train the lip-reading module. In the proposed work, one or two different heights from the geometrical shape of the lip are calculated, which are explained in below subsection 3.3.2.1 and 3.3.2.2.

3.3.2.1. *One-Height Feature Extraction method*

The input for this method is a set of coordinates i.e. lip with 49-68 facial landmarks plotted on it, as shown in Figure 5 and Figure 7. As mentioned, the pronunciation of any word is associated with a series of lip movements. It is clear that all of these lip movements are not similar. i.e., the series of lip heights have patterns. In order to capture this pattern, there is a need to calculate the lip height from the geometrical shape of the lip from every frame.

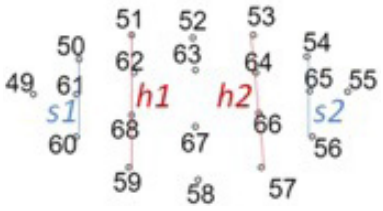

Figura 5: Distances between the landmark points of lip region of a frame.

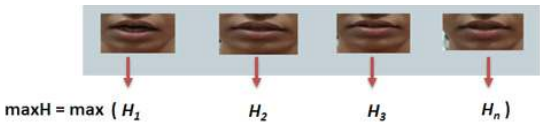

Figura 6: Max height from the sequence of heights.

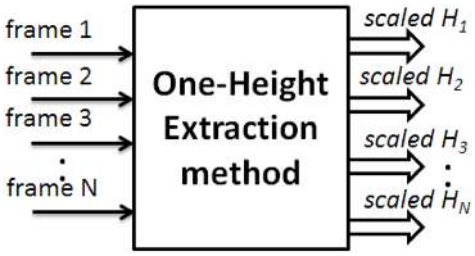

Figura 7: One-Height Extraction method.

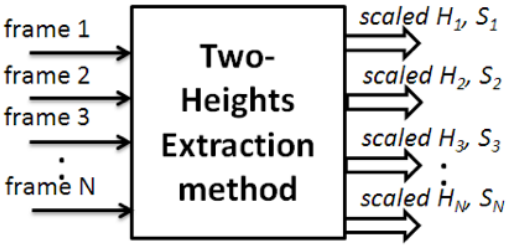

Figura 8: Two-Heights Extraction method.

The height of the lip region in each of the frames is determined using the facial landmark points(coordinates) on the lip region of that frame. The height is calculated as per equation 1, given below.

Let x-y coordinates of facial landmark point i be x_i *and* y_i*. Example:* x_{51}, y_{51}

$$h1 = \sqrt{(x_{51} - x_{59})^2 + (y_{51} - y_{59})^2} \quad \textit{i.e. distance between points 51 and 59}$$

$$h2 = \sqrt{(x_{53} - x_{57})^2 + (y_{53} - y_{57})^2}$$

let j be the frame number in the video

$$H_j = max\ (h1, h2) \qquad (1)$$

Thus a question might arise why the distance between the landmark points 51-59, *h1,* and 53-57, *h2* are considered (Figure 5)? This is because the variation of the orientation of these points is majorly responsible for the lip movement in the frame. When the person speaking in the video is aligned straight to the camera, the *h1* and *h2* are equal. When the person speaking in the video is aligned towards left, the *h1* is greater than *h2*. When the person speaking in the video is aligned towards the right, *h1* is less than *h2*. So, in order to overcome this variation, max of the *h1* and *h2* is considered, i.e. lip Height H_j in equation (1). So, lip movement in j^{th} frame is represented by this height, H_j. A series of such Heights, H_j, is generated for every frame and every word spoken by the person from the input video. Hence, a series of heights from each frame, say H_1, H_2, H_3....H_j is obtained for each word as shown in Figure 7. Next, the scaling of these heights is performed. For scaling of heights, the maxH, from all the frames of a person's video is calculated i.e. max of (H1, H2, H3, …Hn). To get scaled height for every frame, equation (2) is applied. This Procedure is pictorially represented in Figure 6 and Figure 7.

$$maxH = max\ (H_1, H_2, H_3....H_j)$$

$$Scaled\ H_j = ((H_j)/maxH) * scale \qquad (2)$$

Hence, by One-Height Feature Extraction method, a series of scaled heights are obtained for each word of each video and are used for training the model.

3.3.2.2. *Two-Heights Feature Extraction method*

This method focuses on obtaining more detailed data of the lip movement from each of the frames. Therefore, it considers two heights H_n and S_n from a frame as shown in Figure 5. The heights S_n is obtained by the below-mentioned equations (3). In this method, H_n and S_n from all the frames represent the lip movement sequence of words spoken in the given input video.

Let x-y coordinates of facial landmark point i be x_i *and* y_i*. Example:* x_{s1}, y_{s1}

$$s1 = \sqrt{(x_{50} - x_{60})^2 + (y_{50} - y_{60})^2} \quad i.e.\ distance\ between\ points\ 50\ and\ 60$$

$$s2 = \sqrt{(x_{54} - x_{56})^2 + (y_{54} - y_{56})^2}$$

let n be the frame number in the video

$$S_n = max\ (s1, s2) \tag{3}$$

$$Scaled\ S_n = ((S_n)/maxH) * scale \tag{4}$$

Hence, the sequence of heights H_n and S_n are obtained for each word spoken from all videos. Similar to 'one-height feature extraction' method, and these heights are scaled as given in equation (1)-(4). This is depicted in Figure 8. The sequences of scaled heights for spoken words are used for training the model.

3.4. Model

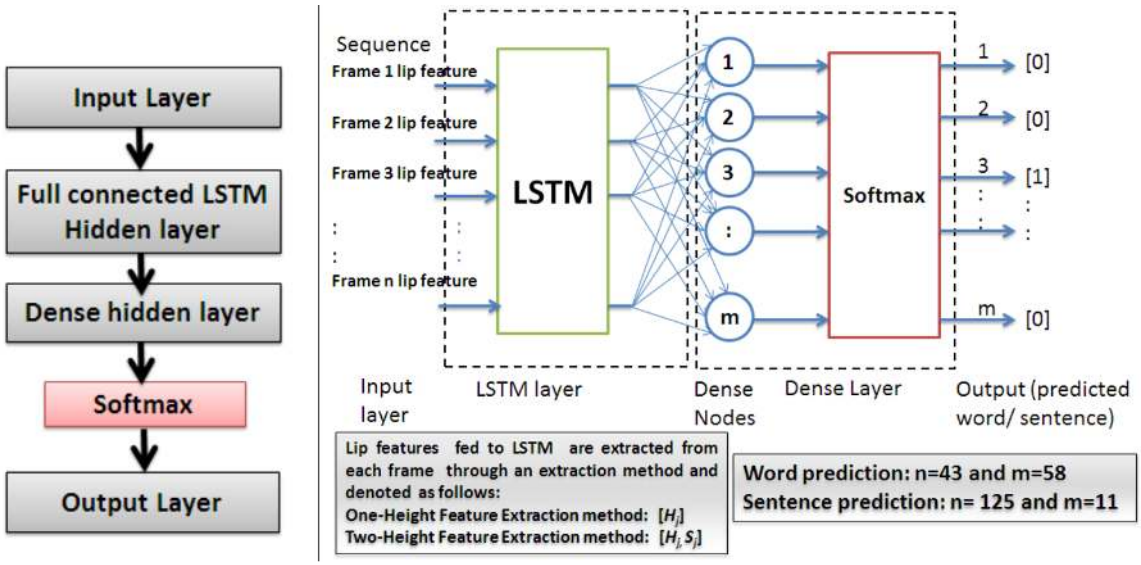

Figura 9: Lip reading's deep learning architecture.

Lip-Reading system relies on the type of Recurrent Neural Network, the Long Short Term Memory (LSTM). The architecture in this scenario consists of three hidden layers, as shown in Figure 9. The First Layer, Sequential Layer, consists of n nodes where n is the maximum length of a sequence in the input data. Each height, H_j, in the lip height Sequence of the word will be input to each node in case of 'one-height feature extraction method'. This layer learns the pattern in this sequence, based on the feedback mechanism. The Second Layer, Dense Layer, consists of m nodes, each of which corresponds to unique words/sentence in the considered paragraph. The output of the first layer is fed as input to each of the nodes in the Dense Layer. The third layer is a softmax

layer. The model is trained on the training data set. Each node in the Dense Layer, representing each word/ sentence, matches its pattern with its input. If the pattern matches, it produces output 1. Only one node among the *m* nodes produces the output 1, and the remaining nodes give 0 as an output. The input word/ sentence sequence corresponds to the word/ sentence reflected by the node, which produces 1 as an output.

4. Results

4.1. Word Recognition

The Lip-Reading training is performed using both one-height and two-heights feature extraction methods, which are explained in subsection 3.3.2.1 and 3.3.2.2.

In training with one-height feature extraction, the proposed model gives an accuracy of 77% when trained data of 3 words spoken by 9 people with 37 epoch. The model gives an accuracy of 8% when the number of words in the data is increased to 58 words spoken by 30 people even after 100 epochs. The accuracy can be increased if more features from the lip geometrical shape are extracted. Therefore, the authors used $S_{j,}$ a second height from the geometrical shape of the lip while training with two-heights feature extraction method (Section 3.3.2.2). The two heights(H_j and S_j) taken from the geometrical shape of the lip of each frame should increase the accuracy, and the same is depicted in the comparison of experiment number II and IV given in Table 2.

Tabla 2: Results- word recognition

Experiment Number	Feature extraction Method	Epoch Value	Number of Words	Number of People	Accuracy
I	One-Height Feature Extraction	37	3	9	77.02%
II	One-Height Feature Extraction	100	58	30	08.10%
III	Two-Heights Feature Extraction	70	10	30	35.60%
IV	Two-Heights Feature Extraction	100	58	30	12.32%

In training with a two-height feature extraction method, the model gives an accuracy of 35% when trained for data of 10 words spoken by 30 people with 70 epochs. When the number of words in training is increased to 58 with 100 epochs, the model gives an accuracy of 12%. The authors tried to extract more features from the geometrical shape of lip landmark points to train the model, like lip width taken as the distance between the facial landmark points 49 and 55(Figure 3). But there is no significant increase in accuracy. The overall analysis of these results is given in the following paragraph.

In training with a two-height feature extraction method, the model gives the best overall accuracy of 35% for limited 10 words. Such accuracy is obtained due to two lip heights used as features that reflect detailed lip movement. Using two lip heights helps the LSTM to learn better patterns which leads to the overall increase in the accuracy of the model.

The authors first trained the LSTM model with only 3 words (experiment I in Table 2) and obtained an accuracy of 77%. This motivated the authors to proceed further with more experiments and train the model with more words. Later, in subsequent efforts to train the model with more data, the authors could achieve much less accuracy than the first experiment, but it is comparable with state of the art (Noda, Yamaguchi, Nakadai, Okuno, & Ogata, 2015) and (Lucey, Sridharan, & Dean, 2008) results which are using data of similar sizes. The low accuracy can be attributed to the following reasons.

i. **Dataset size**: The size of the dataset is small. The authors could record videos of 50 persons, and only 35 person's data could be used for training. This is because the lip detection program could not detect lip region accurately for the persons with mustache and beards, as shown in Figure 10, i.e. lip landmark

points are not mapping the edges of lip. Since recording and capturing data of lip movement is a hectic task, authors had to manage with the available data of only 35 persons. But the authors believe that the accuracy could be increased by training the model with a large dataset. Therefore capturing video and generating lip movement features can be a future research question to be addressed.

ii. **Frame Rate**: The frame rate of the video is 30, and only limited snapshots of lip movement could be captured. A person can speak up to 4 words per second, and therefore video capturing the lip movement should have a much higher frame rate to capture sufficient lip movements. The authors approximately suggest that the frame rate of the video should be 100 or more. Therefore, the authors believe that a better frame rate of the video could capture more information about lip movement and hence better can be the accuracy. However, the videos captured in many of the real-world applications like CCTV cameras are always of low frame rate.

iii. **Training Module**: The architecture of the training model is shown in Figure 9 which consists of many-to-many LSTM and a hidden layer with a SoftMax layer. This module can give higher accuracy with large data set since similar modules are used for text processing ("Long short- term memory - Wikipedia," 2019).

iv. **Image Processing and Facial part detection**: The dlib library ("dlib C++ Library," 2019) is used for lip detection from the video frames, as explained in subsection 3.3.1. The dlib does an excellent job of detecting the lip and provide 20 coordinates that landmark the edges of the lip. However, the library doesn't do well for persons having mustaches and beards, as shown in Figure 10.

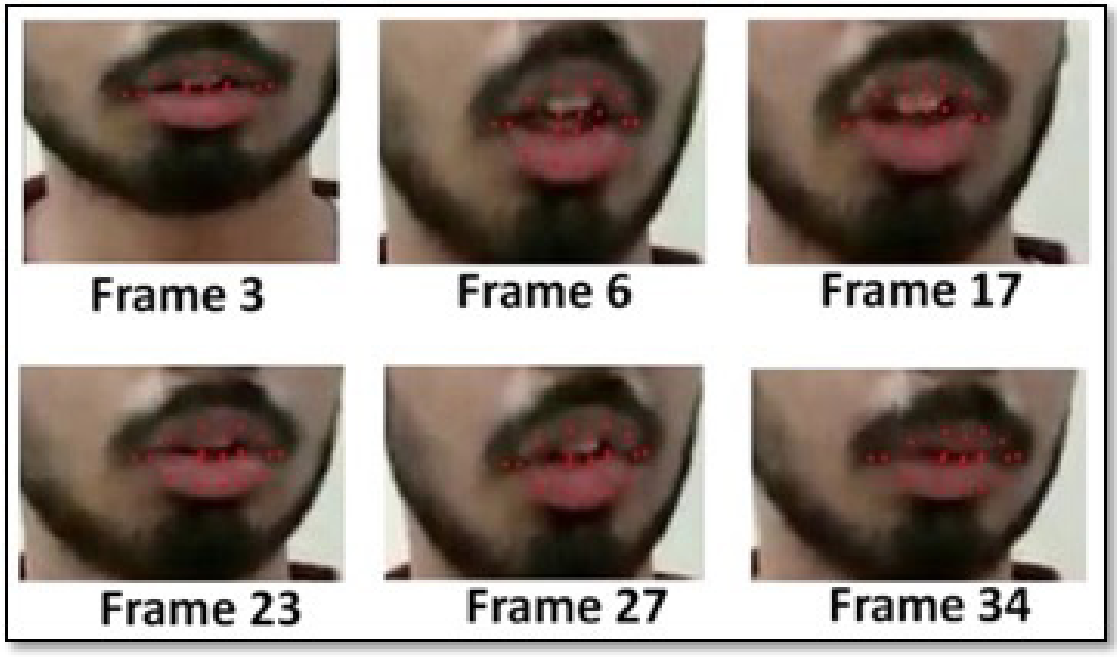

Figura 10: Problems with lip detection.

4.2. Sentence Recognition

The paragraph spoken by the person is split into 10 sentences, and the proposed lip-reading is trained to recognize these sentences. Here, the two-heights extraction method is used for extracting the features of lip movements. After training with 10 sentences spoken by 18 persons, an accuracy of 20% is achieved. This result is

near to the performance of previous similar works (Afouras et al., 2018) and (Sterpu & Harte, 2018) which have a much more complex learning module than the proposed learning module.

Tabla 3: Results- Sentence Recognition

Sl. No	Feature Extraction Method	Epoch Value	Number of Sentences	Number of People	Accuracy
I	One-Height Extraction	73	10	18	20.32%

5. Conclusions

A novel solution of capturing features from the geometrical shape of lip movement from the frames of video and then use it for training is proposed in this paper. An accuracy of 77% and 35% is achieved when trained on the data set of 3 words and 10 words, respectively. Also, an accuracy of 20% is achieved in recognizing sentences. These results are comparable with state of the art performances (Fernandez-Lopez & Sukno, 2018) which have used either similar size of data or much complex learning modules for training when compared to proposed work. The proposed work uses two features from the lip's geometrical shape and a learning model with a small number of neurons when compared to state of the art. It is also worthy to note that earlier works (Fernandez-Lopez & Sukno, 2018; Cooke et al., 2006) have used words/sentences that are easily differentiable in lip movement patterns. In most cases, sentences being used are not meaningful when spoken. But the proposed work uses sentences/words that are continuous and meaning full when spoken. Section 4 analyzes results and provides hints to improve the accuracy of the proposed solution for future research. Generating lip movement features from low frame rate videos is a research question which may be adressed in the future. The authors also believe that the learning model might be improved by using other approaches coupled with LSTM .

Overall the automated lip-reading is required for communication in noisy environments, and very little research has been conducted to provide solutions that address such a research question. It is believed that the solution provided in this paper could be of real use in research and building of the automated lip-reading application.

6. Acknowledgment

The authors wish to express special thanks to the volunteers (KLETech University 2nd year students of computer science & engineering) who supported us in developing the training data set. The authors would also like to thank the management of KLE Technological University for motivating and providing resources to do research.

7. References

Afouras, T., Chung, J. S., & Zisserman, A. (2018). Deep lip reading: A comparison of models and an online application. In *Interspeech 2018* (pp. 3514-3518). ISCA: ISCA. https://doi.org/10.21437/Interspeech.2018-1943.

Almajai, I., Cox, S., Harvey, R., & Lan, Y. (2016). Improved speaker independent lip reading using speaker adaptive training and deep neural networks. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 2722-2726). IEEE. https://doi.org/10.1109/ICASSP.2016.7472172.

Assael, Y. M., Shillingford, B., Whiteson, S., & de Freitas, N. (2016). LipNet: End-to-End Sentence-level Lipreading.

Bailly-Bailliére, E., Bengio, S., Bimbot, F., Hamouz, M., Kittler, J., Mariéthoz, J., … Thiran, J.-P. (2003). The BANCA database and evaluation protocol. In J. Kittler & M. S. Nixon (Eds.), *Audio- and Video-Based*

Biometric Person Authentication (Vol. 2688, pp. 625-638). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/3-540-44887-X_74.

Buchan, J. N., Paré, M., & Munhall, K. G. (2007). Spatial statistics of gaze fixations during dynamic face processing. *Social Neuroscience*, *2*(1), 1-13. https://doi.org/10.1080/17470910601043644.

Chung, J. S., Senior, A., Vinyals, O., & Zisserman, A. (2017). Lip reading sentences in the wild. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 3444-3453). IEEE. https://doi.org/10.1109/CVPR.2017.367.

Chung, J. S., & Zisserman, A. (2017). Lip reading in the wild. In S.-H. Lai, V. Lepetit, K. Nishino, & Y. Sato (Eds.), *Computer vision - ACCV 2016* (Vol. 10112, pp. 87-103). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-54184-6_6.

Cooke, M., Barker, J., Cunningham, S., & Shao, X. (2006). An audio-visual corpus for speech perception and automatic speech recognition. *The Journal of the Acoustical Society of America*, *120*(5), 2421-2424. https://doi.org/10.1121/1.2229005.

dlib C++ Library. (n.d.). Retrieved July 23, 2019, from http://dlib.net/.

Dupont, S., & Luettin, J. (2000). Audio-visual speech modeling for continuous speech recognition. *IEEE Transactions on Multimedia*, *2*(3), 141-151. https://doi.org/10.1109/6046.865479.

Easton, R. D., & Basala, M. (1982). Perceptual dominance during lipreading. *Perception & Psychophysics*, *32*(6), 562-570.

Erber, N. P. (1975). Auditory-visual perception of speech. *The Journal of Speech and Hearing Disorders*, *40*(4), 481-492.

Fernandez-Lopez, A., Martinez, O., & Sukno, F. M. (2017). Towards Estimating the Upper Bound of Visual-Speech Recognition: The Visual Lip-Reading Feasibility Database. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)* (pp. 208-215). IEEE. https://doi.org/10.1109/FG.2017.34.

Fernandez-Lopez, A., & Sukno, F. M. (2018). Survey on automatic lip-reading in the era of deep learning. *Image and Vision Computing*, *78*, 53-72. https://doi.org/10.1016/j.imavis.2018.07.002.

Google's DeepMind AI can lip-read TV shows better than a pro | New Scientist. (n.d.). Retrieved July 23, 2019, from https://www.newscientist.com/article/2113299-googles-deepmind-ai-can-lip-read-tv-shows-better-than-a-pro/.

Hilder, S., Harvey, R. W., & Theobald, B. J. (2009). Comparison of human and machine-based lip-reading. *AVSP*.

Hindi - Wikipedia. (n.d.). Retrieved July 24, 2019, from https://en.wikipedia.org/wiki/Hindi.

Home - Keras Documentation. (n.d.). Retrieved July 23, 2019, from https://keras.io/.

Huang, J., Potamianos, G., Connell, J., & Neti, C. (2004). Audio-visual speech recognition using an infrared headset. *Speech Communication*, *44*(1-4), 83-96. https://doi.org/10.1016/j.specom.2004.10.007.

Long short- term memory - Wikipedia. (n.d.). Retrieved July 23, 2019, from https://en.wikipedia.org/wiki/Long_short-_term_memory

Lucey, P. J., Sridharan, S., & Dean, D. B. (2008). Continuous pose-invariant lipreading.

Matthews, I., Cootes, T. F., Bangham, J. A., Cox, S., & Harvey, R. (2002). Extraction of visual features for lipreading. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *24*(2), 198-213. https://doi.org/10.1109/34.982900.

Moll, K. L., & Daniloff, R. G. (1971). Investigation of the Timing of Velar Movements during Speech. *The Journal of the Acoustical Society of America*, *50*(2B), 678-684. https://doi.org/10.1121/1.1912683.

Nefian, A V, Liang, L., Pi, X., & Xiaoxiang, L. (n.d.). A coupled HMM for audio-visual speech recognition. …

Nefian, Ara V., Liang, L., Pi, X., Liu, X., & Murphy, K. (2002). Dynamic Bayesian Networks for Audio-Visual Speech Recognition. *EURASIP Journal on Advances in Signal Processing*, *2002*(11), 783042. https://doi.org/10.1155/S1110865702206083.

Noda, K., Yamaguchi, Y., Nakadai, K., Okuno, H. G., & Ogata, T. (2015). Audio-visual speech recognition using deep learning. *Applied Intelligence*, *42*(4), 722-737. https://doi.org/10.1007/s10489-014-0629-7.

Ortiz, I. de los R. R. (2008). Lipreading in the Prelingually Deaf: What makes a Skilled Speechreader? *The Spanish Journal of Psychology*, *11*(2), 488-502. https://doi.org/10.1017/S1138741600004492.

Patterson, E. K., Gurbuz, S., & Tufekci, Z. (2002). CUAVE: A new audio-visual database for multimodal human-computer interface research.

Petridis, S., & Pantic, M. (2016). Deep complementary bottleneck features for visual speech recognition. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 2304-2308). IEEE. https://doi.org/10.1109/ICASSP.2016.7472088.

Petridis, S., Shen, J., Cetin, D., & Pantic, M. (2018). Visual-Only Recognition of Normal, Whispered and Silent Speech. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6219-6223). IEEE. https://doi.org/10.1109/ICASSP.2018.8461596.

Potamianos, G., Neti, C., Gravier, G., & Garg, A. (2003). Recent advances in the automatic recognition of audiovisual speech. *Proceedings of the ...*

Potamianos, G., Neti, C., & Luettin, J. (2004). Audio-visual automatic speech recognition: An overview. *Issues in Visual and Audio ...*

Ronquest, R. E., Levi, S. V., & Pisoni, D. B. (2010). Language identification from visual-only speech signals. *Attention, Perception, & Psychophysics*, *72*(6), 1601-1613. https://doi.org/10.3758/APP.72.6.1601.

Sterpu, G., & Harte, N. (n.d.). Towards lipreading sentences using active appearance models.

Sui, C., Bennamoun, M., & Togneri, R. (2015). Listening with Your Eyes: Towards a Practical Visual Speech Recognition System Using Deep Boltzmann Machines. In *2015 IEEE International Conference on Computer Vision (ICCV)* (pp. 154-162). IEEE. https://doi.org/10.1109/ICCV.2015.26.

Sumby, W. H. (1954). Visual Contribution to Speech Intelligibility in Noise. *The Journal of the Acoustical Society of America*, *26*(2), 212. https://doi.org/10.1121/1.1907309.

Wand, M., Koutnik, J., & Schmidhuber, J. (2016). Lipreading with long short-term memory. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6115-6119). IEEE. https://doi.org/10.1109/ICASSP.2016.7472852.

Yau, W. C., Kumar, D. K., & Weghorn, H. (2007). Visual speech recognition using motion features and hidden markov models. In W. G. Kropatsch, M. Kampel, & A. Hanbury (Eds.), *Computer analysis of images and patterns* (pp. 832-839). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-74272-2_103.

Zhou, Z., Zhao, G., Hong, X., & Pietikäinen, M. (2014). A review of recent advances in visual speech decoding. *Image and Vision Computing*, *32*(9), 590-605. https://doi.org/10.1016/j.imavis.2014.06.004.