ABSTRACT

In last three decades progress in Artificial intelligence and Machine learning has made tremendous development in the field of computer vision, Natural language processing (NLP), Automatic Speech Recognition (ASR) and many more. ASR also known as speech to text or Lip Reading, is a visual speech recognition technology, which interprets the speech by observing the lip, tongue, and teeth movements of the speaker without speech signals. Lip movements are, especially, used by partially deaf people to perceive the meaning of speech. Deaf people communicate with each other using sign language. However, it becomes difficult for deaf people to communicate with people who do not know sign language. Through technological innovations and Lip reading techniques, this deficit can be filled. Using Lip reading, in their childhood only, deaf children can be taught their mother tongue easily. Also, the lip expressions can be converted into sentences by a Lip reading system which can be displayed on a deaf person's mobile screen which can be easily read. According to research, skills acquired in childhood help a child to learn other languages in a better way. Most of the Lip reading work is carried out for the English language and other foreign language. Our work is to design and develop an algorithm for Lip detection and extraction algorithm, create a dataset for Gujarati alphabets and alphabet recognition using CNN-LSTM model. We have created an algorithm named ViLiDEX algorithm to remove extra frames, a dataset named GVarna for 34 consonants of Gujarati language. We have used mobile net for class wise (Guttural, Palatal, Retroflex, Dental and Labial) recognition and alphabet classification.

TABLE OF CONTENTS

1. INTRODUCTION

1.1 What is Lip Reading?

Lip reading means to interpret the speech by observing the lip, tongue and teeth movements of speaker without speech signals. Three basic strengths of humans are vision, speech, and hearing capability which are used for communication. If there is a deficiency in any of the capabilities, communication becomes difficult. As hearing is necessary for correct speech and language development, it becomes almost impossible when hearing capability is not fully functioning. Though speech, vision, and hearing skills develop at different ages, hearing skill is much more important for the proper development of speech and language skill. A child who is born deaf or becomes deaf before he begins to speak cannot communicate and become dumb, although they don't have any speaking problems [1]. Deaf people use sign language to communicate. Sign language is an easy and simple way for deaf people to communicate with each other. Using sign language they can explain their feeling to normal people with some difficulty. Deaf people will face a problem when they have to understand any message from normal people. Normal people don't know to communicate in sign language and fail to deliver the message. This could be serious when deaf people are in danger and don't get help from others. In this situation, Lip reading can be useful to communicate with normal people [2]. In childhood, deaf people could be trained to learn their mother tongue using Lip Reading and additional resources.

1.2 Automatic Speech Recognition

Technology has helped whenever strengths of human being became weak. Evolution in Automatic speech recognition (ASR) from ADALINE to Alexa happened to strengthen the human being. In this evolution Visual speech recognition or Automatic Lip reading needs more focus. Speech signals carry more information than visual signals so it's easier to get better results with speech signals in ASR. In visual speech recognition, lip movements characteristics are used to recognize the speech content of the speaker without a speech signal. Information collected by visual channel is two-dimensional and so contain more redundant data than one-dimensional voice information. Speech signals easily interfere in a noisy environment while visual signals used for ALR will not be affected and could improve recognition rate in a noisy environment [3][4]. Sign language and Lip reading methods combined can help to minimize the communication issues of deaf people [5][6].

1.2.1 History of Automatic Lip Reading

The first ASR system was designed and developed at Bell laboratories in 1952. The system "Digit Recognizer" was based on sound signals and could recognize ten digits spoken in English language [7]. Sumby and Pollack [9] gave the first theory of Automated Lip Reading and suggested that visual observations of speaker can improve oral speech clarity. As digital image processing techniques were not introduced, no lip reading model was designed yet. Lately after 20 years of digital image processing techniques were developed, in 1984 the first Audio-Visual Automatic Speech recognition (AV-ASR) system was designed by Petajan [10]. Goldschen [11] have used statistical model to design a visual-only sentences-level lip reading method using Hidden Markov Model.

Rise in Artificial Intelligence made 21st century a stepping stone for technology. Innovation in machine learning and deep learning methods gave a new direction to the lip reading work. Ngiam [15] has proposed an AV-ASR approach based on deep auto-encoder and restricted Boltzmann machines in 2011. In 2014 Noda has used Convolution Neural Network for feature extraction from images and showed that CNN is significantly better than image processing based methods [16]. In 2016 Wand have used Long Short-Term Method for lip reading on GRID dataset [17]. First end-to-end sentence-level lip reading model named LipNet was presented by Assael [18]. LipNet is based on Spatio-temporal Convolution Neural Network and Recurrent Neural Network. Chung et al. [19] has proposed a WLAS Network based on CNN and LSTM on LRS Dataset.

In last decade, Lip Reading work carried out for Indian languages also. Nandini[143] have implemented Lip reading for Kannada language and Patil [144] have given LSTM based Lip reading approach for Devanagari Script.

Google dictionary feature "learn to pronounce" added in December 2018 is a lip movement of words in different accents (see Figure 1). It helps hearing of hard people to pronounce the word.

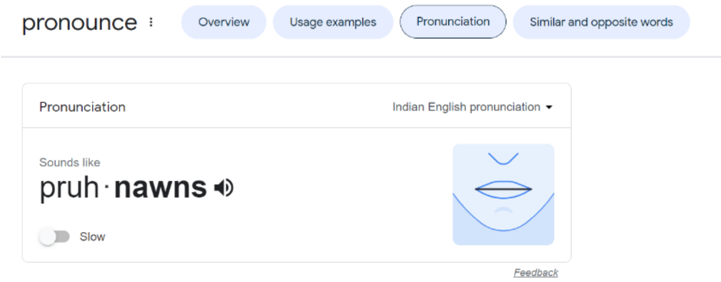

Figure 1. "Learn to pronounce" feature of Google Dictionary (courtesy: google.com)

1.2.2 Automatic Lip Reading Methods

Automatic Lip Reading Methods involves series of actions. These actions are

1. Lip Detection and Extraction: In this step lip region is extracted from raw video. Speech content are recognized through visual information of lips, the quality of extracted lip region (ROI) is very important. Traditional methods of lip detection are colour information-based method, face structure-based method and model based method. Deep learning methods of lip detection method use pre-trained models to extract lip ROI.
2. Feature extraction: the next step after Lip detection and extraction is feature extraction for subsequent classification. Traditional methods for feature extraction are pixel-based, shape-based and mixed feature extraction methods. Deep Neural network structures like Feedforward Neural Network [17], [56], [57] Autoencoder [58], Boltzoman Machine [59], Convolution Neural Network (CNN) are used for feature extraction.
3. Feature transformation and classification**:** Classification methods like Template Matching method, Artificial Neural Network and Hidden Markov Model were used. Now a days RNN, LSTM [60], GRU [61], 2D CNN, 3D CNN, 2D+ 3D CNN, Bi-LSTM and many more deep learning based models are used for classification.

1.3 Motivation

The biggest motivation behind this work is to make Lip reading and understanding easier for deaf people for Gujarati Language and be helpful to them through technology. Lip reading relies on the kind of language of the dataset, and Gujarati is our mother tongue, we chose Gujarati Language to implement Lip Reading.

1.4 Problem statement, objectives and applications

1.4.1 Problem statement: Design and Development of Lip Extraction Algorithm and Dataset Creation for Gujarati Alphabets Recognition via Lip Movement using Deep Learning

1.4.2 Objectives:

Objective of this research is to design a frame removal algorithm, design a dataset and alphabet recognition using pre-trained deep learning model. Further research in this direction can be helpful to hearing of hard people to learn mother tongue in early ages. Hearing of hard people are trained with sign language, but it cannot be helpful for communication with normal people. Using lip reading they can learn pronunciation, and so, can communicate with others. This research can be helpful to the people of rural area of Gujarat.

1.4.3 Research Contribution

a. Dataset for Gujarati Language is not available for Automatic speech recognition, so we have created dataset from scratch, named GVarna. This will be helpful for further research in lip reading for Gujarati language.
b. Different speakers have different span for alphabet utterance, so we have designed a ViLiDEx algorithm to remove extra frames and store key frames.
c. After creating dataset, we have used Mobile Net-a pre-trained model for alphabet classification and recognition based on 5 classes (Guttural, Palatal, Retroflex, Dental, and Labial).

1.4.4 Applications:

There are many applications of automatic lip reading. India is a country with many local languages. Many applications with English language can be implemented with local languages.

1. This work will helpful for further research in Gujarati Language.
2. Automatic Lip reading will help to hearing of hard people to learn the language from childhood.
3. It will help hearing of hard people to learn pronunciation of different languages.
4. In rural area, cellphone authentication with lip movement can be implemented with local languages.

2. LITERATURE STUDY

2.1 Literature study of Automatic Lip reading

Automatic lip reading is based on language, dataset size, image quality, image size and number of speakers involved in dataset. Here we have discussed about lip reading in different languages and technology used for implementation. Initially the dataset was limited, but gradually the dataset size become more complex, as number of speakers, diversity in posture, illumination conditions, background environment changes. Table 1 summarizes lip reading implemented for different languages. Table 2 summarizes lip reading datasets for alphabet and digits. Table 3 summarizes different models used for dataset for alphabet and digits.

Table 1 Lip Reading in different languages

Sr. No	**Title and year**	**Language**	Model	Accuracy
1	A PCA based visual DCT feature extraction method for lip-reading (2006)	Chinese	PCA based Traditional Method	67%
2	A Lip Reading Application on MS Kinect Camera (2013)	Turkish	KNN+HMM	72.44% to 78.22%
3	Lipnet: end-to-end sentence-level lipreading (2016)	English	SpatioTemporal CNN +Bi-Gated Recurrent Unit+ Connectionist Temporal Classification	95.2 % (s) 86.4% (w)
4	Designing and Implementing a System for Automatic Recognition of Persian letters by Lip reading Using Image Processing Methods (2019)	Persian	Back propagation ANN and Radial basis function	-
5	Automatic lip-reading of hearing impaired people(2019)	Japanese	HAAR+AAM	-
6	Marathi digit recognition using lip geometric shape features and dynamic time warping(2017)	Marathi	Traditional Model	63%
7	Deep weighted feature descriptors for lip reading of Kannada language (2019)	Kannada	Deep learning model	84.82%
8	LSTM model for visual speech recognition through facial expressions (2023)	Malayalam	CNN+LSTM	-

Table 2 Lip reading datasets

Name	Language	Task	Speakers	Best Accuracy
Tulips (1995) [102]	English	Digits	96	89.53%
M2VTS (1999) [103]	French	Digits	2920	76.60%
AVLetters (2002) [97]	English	Alphabet	780	69.60%
CUAVE (2002) [105]	English	Digits	7000	83.00%
BANCA (2003) [132]	Multiple	Digits	29,952	-
AV@CAR (2004) [101]	Spanish	Digits, Alphabet	800,600	23.00%
AVICAR (2004) [98]	English	Digits, Alphabet	59,000	37.87%
VALID (2005) [133]	English	Digits	1590	63.21%
AVLetter2 (2008) [100]	English	Alphabet	910	91.80%
IBMSR (2008) [134]	English	Digits	1661	68.58%
CENSREC-1-AV (2010) [109]	Japanese	Digits	3234	39.30%
NDUTAVSC (2010) [108]	German	Digits	6907	84.24%
AGH AV (2012) [110]	Polish	Digits	-	-
AusTalk (2014) [131]	English	Digits	24,000	69.18%
OuluVS2 (2015) [106]	English	Digits	1590	96.90%
AV Digits (2018) [107]	English	Digits	795	68.00%

Table 3 Models used for Alphabet and Digit datasets

Name	Year	Model		Accuracy
		Front end	Back end	
AVLetters (Alphabet)	2002	LBP-TOP	SVM	62.80%
	2013	RFMA		69.60%
	2016	DBNFs + DCT	LSTM	58.10%
	2016	RMRBM		64.63%
CAUVE (digit)	2009	AAM	HMM	83%
	2011	Autoencoder + RBMs		68.70%
	2017	DBNF	GMM-HMM	63.40%
	2017	Autoencoder+LSTM	Bi-LSTM	78.60%

2.2 Difficulties and Challenges of Automatic Lip reading

Automatic Lip reading is a challenging task because its input is a video or image sequences and most of the image contents are similar or unchanged. The primary distinction is the change in lip movement, but for alphabets, this change is very minute. Main challenges of lip reading task are described as follows:

External Factors: Different factors like illumination, skin colour, beards, and wrinkles on the skin affect the process of feature extraction. To overcome this problem traditional lip reading methods use shape-based methods. In Shape-based methods extracted features only include the shape of the lips [44],[45] and other external factors like illumination, skin colour, and beard will be discarded. In deep learning-based methods, various methods are used to extract spatial and temporal features of lip movement.

Visual Ambiguity: In alphabet pronunciations, different phonemes have same mouth shape. Such visemes are difficult to distinguish without context. Speakers' ascent will add more complexity to the feature extraction task. Phoneme-to-viseme mappings [38], [153], [68] and adjacent character/words phenomena [19], [141], [87] will solve the problem of visible ambiguity at some extent.

Speakers' Pose: when data are collected from TV shows or online sites, position of speakers' head may vary. Different postures of speakers with different angles make feature extraction task difficult. The multi-view datasets like LRW [113], LRW-1000[86], LRS2-BBC [126], OuluVS2[106] are very helpful to solve this problem.

Speaker dependent: Performance of lip reading task is very much dependent on number of speakers. People from different region have different styles, ascent, pronunciation and habit of speaking. This may also affect the performance of lip reading task. If large scale of dataset is available with a greater number of speakers, influence of speakers' dependency can be reduced.

Database parameters: Databases with limited number of speakers, corpus and samples also affects the performance of lip reading task. Databases collected from TV shows, the background, illumination, environment and other parameters are similar as well language content is also limited. A large-scale of datasets with a greater number of speakers from different regions and different postural background give more fruitful results for lip reading task.

3. RESEARCH CONTRIBUTION

3.1 Dataset creation

We are implementing Lip Reading using machine learning for Gujarati language. There are total 36 consonants and 12 vowels in Gujarati Alphabet (see Figure 2). 36 consonants are classified in 5 sub classes named Guttural, Palatal, Retroflex, Dental and Labial (see Figure 3).

ક	ખ	ગ	ઘ	ઙ	ચ	છ	જ	ઝ	ઞ
ka	kha	ga	gha	ṅa	ca	cha	ja	jha	ña
[kə]	[kʰə]	[gə]	[gʱə]	[ŋə]	[ʧə]	[ʧʰə]	[ʤə]	[ʤʱə]	[ɲə]
ટ	ઠ	ડ	ઢ	ણ	ત	થ	દ	ધ	ન
ṭa	ṭha	ḍa	ḍha	ṇa	ta	tha	da	dha	na
[ʈə]	[ʈʰə]	[ɖə]	[ɖʱə]	[ɳə]	[t̪ə]	[t̪ʰə]	[d̪ə]	[d̪ʱə]	[nə]
પ	ફ	બ	ભ	મ	ય	ર	લ	વ	
pa	pha	ba	bha	ma	ya	ra	la	va	
[pə]	[fə]	[bə]	[bʱə]	[mə]	[jə]	[rə]	[lə]	[ʋə]	
શ	ષ	સ	હ	ળ	ક્ષ	જ્ઞ			
śa	ṣa	sa	ha	ḷa	kṣa	gña			
[ʃə]	[ʃə]	[sə]	[ɦə]	[ɭə]	[kʃə]	[gnə]			

Figure 2. Gujarati Alphabet (courtesy: omniglot.com)

	sparśa (Stop)								*anunāsika* (Nasal)		*antastha* (Approximant)		*ūṣma/saṃghashrī* (Fricative)			
Voicing →	*aghoṣa*				*ghoṣa*								*aghoṣa*		*ghoṣa*	
Aspiration →	*alpaprāṇa*		*mahāprāṇa*		*alpaprāṇa*		*mahāprāṇa*		*alpaprāṇa*				*mahāprāṇa*			
kaṇṭhya (Guttural)	क	ka /k/	ख	kha /kʰ/	ग	ga /g/	घ	gha /gʱ/	ङ	ṅa /ŋ/					ह	ha /h/
tālavya (Palatal)	च	ca /c, t͡ʃ/	छ	cha /cʰ, t͡ʃʰ/	ज	ja /ɟ, d͡ʒ/	झ	jha /ɟʱ, d͡ʒʱ/	ञ	ña /ɲ/	य	ya /j/	श	śa /ɕ, ʃ/		
mūrdhanya (Retroflex)	ट	ṭa /ʈ/	ठ	ṭha /ʈʰ/	ड	ḍa /ɖ/	ढ	ḍha /ɖʱ/	ण	ṇa /ɳ/	र	ra /r/	ष	ṣa /ʂ/		
dantya (Dental)	त	ta /t̪/	थ	tha /t̪ʰ/	द	da /d̪/	ध	dha /d̪ʱ/	न	na /n/	ल	la /l/	स	sa /s/		
oṣṭhya (Labial)	प	pa /p/	फ	pha /pʰ/	ब	ba /b/	भ	bha /bʱ/	म	ma /m/	व	va /ʋ/				

Figure 3. Devanagari alphabet classification (Courtesy: https://bhashabodha.blogspot.com/)

For consonant classification purpose we have considered 34 consonants only. No dataset for Gujarati language is available for Automatic speech recognition, so we have created this dataset from scratch, named GVarna. GVarna is a 2D image data with depth. Which included 34 consonants of Gujarati language. We have recorded videos using Nikon D 5300 camera with 1920 X 1080 full HD resolutions and 30 frames/second. Our family members, friends, relatives and students who know Gujarati language were speaking 34 consonants in one continuous video. Recording is performed at one place to avoid the difference of illumination, light and noise. As speakers have different speed and accent one character span is 1 or 2 seconds. We have recorded such 3 shots of 24 speakers for 34 consonants. Total 72 video files are recorded and each consonant separated using "Movies and TV" application on Windows 10 operating system.

3.2 ViLiDEx algorithm for Lip detection and Extraction

We have designed lip detection and extraction algorithm based on Facial landmark pre-trained model of Dlib. Facial landmark using Dlib gives total of 68 landmarks of face, among them landmarks from 49-68 which are for lip area are cut down and given as an input for next level (see Figure 4). Existing algorithm keeps first even/odd number of frames from total number of frames and discard remaining frames. Key frames are distributed throughout the video. Alphabet utterance of different speaker may vary in time. For long utterance, total number of frames are more than short utterance, and if first odd/even frames will be kept, key frames may be discarded. To overcome this problem, we proposed an algorithm ViLiDEx.

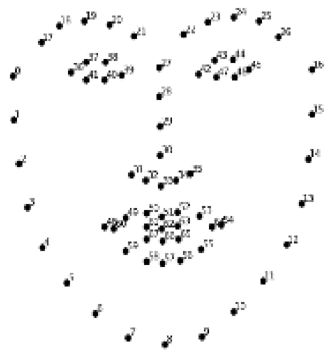

Figure 4. Face landmark points

3.2.1 Working of ViLiDEx algorithm

This algorithm takes a video as an input, count total frames, for each frame detects lip area and extract and save the new frame. If the total number of frames is more than the limit (20/25), extra frames will be removed. Frame removal is based on frame numbers. Frame numbers divided by following numbers (2, 3, 5, 7, 11, 17…up to Total number of frames) will be removed.

This algorithm calculates total frames of input video of alphabet. If total frames are multiple of 20 (20*1, 40(20*2), 60(20*3), 80(20*4) … and so on), then frame number divisible by multiplicand (1, 2, 3, 4…) will be kept and others will be discarded as extra frames. If total frames are not multiple of 20 then Frame difference will be calculated. Prime numbers and total numbers divisible by these prime numbers up to total frames are listed. Prime numbers whose count is equal to frame difference will be searched and frame numbers divisible by these prime numbers will be discarded (See Table 1). For the remaining 20 frames, using Face landmark points 49-68, lip area will be extracted and stored. Time complexity of this algorithm is *O(m*n*p)*, where *m*n* is the resolution of image in the frame and *p* is total number of frames. Steps of ViLiDEx algorithm are as follows.

3.2.2 ViLiDEx algorithm:

1. *Read input video.*
2. *Count Total number of Frames.*
3. *Calculate Frame difference = Total Frames- 20*
4. *If frame difference = 0*
    Density = 'E'
    Divisor =1
   Else if Frame difference % 20 = 0
    Density = 'M'
    Divisor = int (Total Frames / 20)
   Else
    Density='S'
    List Prime numbers from 3 to Total Frames
    Count total numbers (1 to Total Frames) divisible by each prime number listed above
    Search for the counts whose total is equal to frame difference
    Corresponding numbers in list of primes are List of Divisors for Extra frames
5. *Set the path to store dataset*
6. *For each frame in input video*
   If Density = 'E'
    Crop lip area from each frame and store
   Else if Density = 'M'
    Crop the frames whose number divisible by Divisor and store
    Discard other frames
   Else
    Crop the frames whose number divisible by List of Divisors and store
    Discard other frames
7. *Close input video*

Table 4. Working of ViLiDEx algorithm

Total Frames	**Frame Difference**	**Divisor /List of Primes**	**Density**	**Prime Nos Needed**	**Count total numbers divisible by each prime**
20	0	1	'E' for Equal	-	-
40	20	40/2=2	'M' for Multiple of 20	-	-
30	10	[3]	'L' for in List of Primes	[**3**, 5, 7, 11, 13, 17, 19, 23, 29]	[**10**, 6, 4, 2, 2, 1, 1, 1, 1]

39	19	[3, 7, 17]	‘S’ for search in List of Primes	[**3**, 5, **7**, 11, 13, **17**, 19, 23, 29, 31, 37]	[**13**, 7, **5**, 3, 3, **2**, 2, 1, 1, 1, 1] 13 + 5 + 2 -1(remove frame no 21 common for 3 and 7) = 19
38	18	[3, 7, 11]	‘S’ for search in List of Primes	[**3**, 5, **7**, **11**, 13, 17, 19, 23, 29, 31, 37]	[**12**, 7, **5**, 3, 3, 2, 2, **1**, 1, 1, 1] 12+5-1+3-1

3.3 Alphabet Recognition/Classification

For alphabet classification we are using CNN-LSTM model. CNN identifies lip shape and then each of the output of CNN become transform to sequence. LSTM learns pattern the sequence that contain the change of lip shape. We are using well trained CNN model MobileNet.

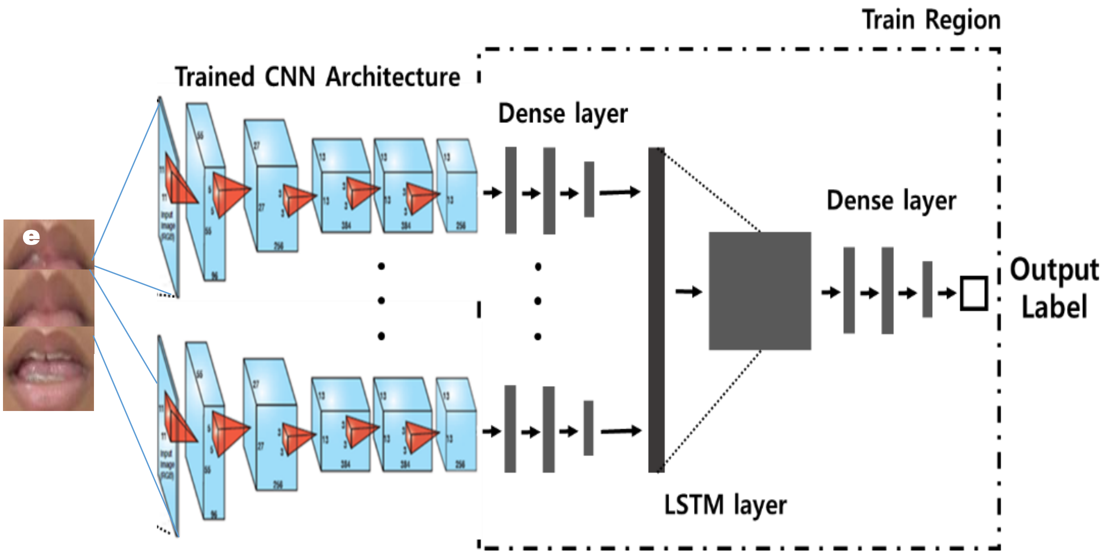

Figure 5. CNN LSTM Model for classification

Here we have used 8800 images for training and 4400 images for testing. We have used two learning rates (0.0001 and 0.0002) and different epochs (5,10,20,30,40,50,100). We have applied cross validation with 3 set of overlapping datasets. Steps for training and testing using Mobile Net are given below.

Training of GVarna dataset using Mobile Net:

1. Set Parameters like timestamp, labels, learning rate, batch size, epochs etc.
2. Create empty dataset and load training dataset
3. Build a model

4. Compile the model
5. Train the model
6. Save the model

Testing of GVarna dataset using Mobile Net:

1. Load the saved model
2. Load testing dataset
3. Test dataset
4. Print result

4. RESULTS AND OBSERVATIONS

4.1 Results

GVarna dataset is trained using CNN-LSTM based MobileNet model. Total 8800 images are trained and 4400 images are tested (66:33 ratio). We have used two learning rates 0.0001 and 0.0002 for 5, 10, 20, 30, 40, 50, 60, and 100 epochs (table 5 and table 6).

Table 5 Precision, Recall and F1 score for learning rate 0.0001 and different epochs

		Learning Rate 0.0001							
Epochs		**5**	**10**	**20**	**30**	**40**	**50**	**60**	**100**
Precision	**Guttural**	0.21	0.21	0.19	0.17	0.15	0.22	0.19	0.16
	Palatal	0.00	0.00	0.29	0.40	0.33	0.23	0.20	0.24
	Retroflex	0.00	0.00	0.23	0.21	0.20	0.25	0.26	0.29
	Dental	0.20	0.27	0.15	0.12	0.23	0.19	0.16	0.08
	Labial	0.00	0.00	0.38	0.33	0.33	0.40	0.36	0.31
	Overall	**0.20**	**0.21**	**0.24**	**0.24**	**0.23**	**0.23**	**0.23**	**0.21**
Recall	**Guttural**	0.93	0.95	0.09	0.18	0.14	0.25	0.18	0.20
	Palatal	0.00	0.00	0.52	0.18	0.30	0.41	0.41	0.32
	Retroflex	0.00	0.00	0.34	0.41	0.25	0.32	0.11	0.18
	Dental	0.09	0.09	0.16	0.05	0.34	0.14	0.11	0.07
	Labial	0.00	0.00	0.07	0.36	0.14	0.05	0.32	0.30
	Overall	**0.20**	**0.21**	**0.24**	**0.24**	**0.23**	**0.23**	**0.23**	**0.21**
F1 score	**Guttural**	0.34	0.34	0.12	0.17	0.14	0.24	0.18	0.18
	Palatal	0.00	0.00	0.38	0.25	0.31	0.29	0.27	0.27
	Retroflex	0.00	0.00	0.28	0.28	0.22	0.28	0.16	0.22
	Dental	0.13	0.14	0.15	0.07	0.27	0.16	0.13	0.08
	Labial	0.00	0.00	0.12	0.34	0.19	0.08	0.34	0.30
	Overall	**0.09**	**0.09**	**0.21**	**0.22**	**0.23**	**0.21**	**0.22**	**0.21**

Table 6 Precision, Recall and F1 score for learning rate 0.0002 and different epochs

		Learning Rate 0.0002							
Epochs		**5**	**10**	**20**	**30**	**40**	**50**	**60**	**100**
Precision	**Guttural**	0.20	0.21	0.14	0.20	0.19	0.16	0.19	0.19
	Palatal	0.00	0.00	0.27	0.22	0.29	0.23	0.19	0.22
	Retroflex	0.25	0.57	0.00	0.29	0.25	0.23	0.29	0.31
	Dental	0.00	0.25	0.17	0.20	0.12	0.10	0.13	0.12
	Labial	0.00	0.00	0.31	0.34	0.32	0.34	0.30	0.33
	Overall	**0.09**	**0.21**	**0.18**	**0.25**	**0.23**	**0.21**	**0.22**	**0.23**
Recall	**Guttural**	1.00	0.64	0.16	0.25	0.36	0.16	0.20	0.16
	Palatal	0.00	0.00	0.48	0.25	0.45	0.64	0.16	0.48
	Retroflex	0.02	0.09	0.00	0.27	0.05	0.07	0.27	0.25

	Dental	0.00	0.45	0.25	0.18	0.07	0.02	0.11	0.11
	Labial	0.00	0.00	0.18	0.27	0.23	0.25	0.39	0.07
	Overall	**0.20**	**0.24**	**0.21**	**0.25**	**0.23**	**0.23**	**0.23**	**0.21**
F1 score	**Guttural**	0.34	0.32	0.15	0.22	0.25	0.16	0.20	0.18
	Palatal	0.00	0.00	0.34	0.23	0.35	0.34	0.18	0.30
	Retroflex	0.04	0.16	0.00	0.28	0.08	0.11	0.28	0.28
	Dental	0.00	0.33	0.21	0.19	0.09	0.04	0.12	0.11
	Labial	0.00	0.00	0.23	0.30	0.27	0.29	0.34	0.11
	Overall	**0.08**	**0.16**	**0.18**	**0.25**	**0.21**	**0.19**	**0.22**	**0.20**

We have performed cross-validation with three overlapping testing datasets DS0, DS1, and DS2 (table 7). Precision, Recall and F1 score value for 5 classes and two learning rate is shown in table 8, 9, 10, 11, 12.

Table 7 Cross validation

Learning Rate		**0.0002**			**0.0001**		
		Overall Accuracy					
Epochs	**Class**	**DS0**	**DS1**	**DS2**	**DS0**	**DS1**	**DS2**
Precision	**30**	**0.23**	**0.25**	**0.24**	**0.25**	**0.24**	**0.20**
	40	0.19	0.23	0.22	0.25	0.23	0.25
	50	0.25	0.21	0.20	0.25	0.23	0.23
Recall	**30**	**0.25**	**0.25**	**0.25**	**0.25**	**0.24**	**0.20**
	40	0.24	0.23	0.23	0.25	0.23	0.25
	50	0.25	0.23	0.21	0.25	0.23	0.23
F1 score	**30**	**0.23**	**0.25**	**0.22**	**0.24**	**0.22**	**0.16**
	40	0.21	0.21	0.21	0.22	0.23	0.24
	50	0.24	0.19	0.20	0.23	0.21	0.21

Table 8 Accuracy for Guttural Class

Guttural class		**Precision**		**Recall**		**F1 score**	
Learning Rate -->		**0.0001**	**0.0002**	**0.0001**	**0.0002**	**0.0001**	**0.0002**
Epochs	**Dataset**						
30 Epochs	**DS0**	0.21	0.27	0.39	0.20	0.27	0.23
	DS1	0.17	0.20	0.18	0.25	0.17	0.22
	DS2	0.09	0.25	0.02	0.32	0.04	0.28
40 Epochs	**DS0**	0.25	0.25	0.18	0.23	0.21	0.24
	DS1	0.15	0.19	0.14	0.36	0.14	0.25
	DS2	0.25	0.31	0.20	0.18	0.23	0.23
50 Epochs	**DS0**	0.26	0.24	0.14	0.45	0.18	0.31
	DS1	0.22	0.16	0.25	0.16	0.24	0.16
	DS2	0.17	0.22	0.07	0.25	0.10	0.23

Table 9 Accuracy for Palatal Class

Palatal class		Precision		Recall		F1 score	
Learning Rate -->		0.0001	0.0002	0.0001	0.0002	0.0001	0.0002
Epochs	Dataset						
30 Epochs	DS0	0.20	0.05	0.09	0.02	0.13	0.03
	DS1	0.40	0.22	0.18	0.25	0.25	0.23
	DS2	0.00	0.07	0.00	0.02	0.00	0.03
40 Epochs	DS0	0.06	0.00	0.02	0.00	0.03	0.00
	DS1	0.33	0.29	0.30	0.45	0.31	0.35
	DS2	0.30	0.07	0.25	0.02	0.27	0.03
50 Epochs	DS0	0.10	0.13	0.07	0.07	0.08	0.09
	DS1	0.23	0.23	0.41	0.64	0.29	0.34
	DS2	0.19	0.18	0.11	0.07	0.14	0.10

Table 10 Accuracy for Retroflex class

Retroflex class		Precision		Recall		F1 score	
Learning Rate -->		0.0001	0.0002	0.0001	0.0002	0.0001	0.0002
Epochs	Dataset						
30 Epochs	DS0	0.22	0.30	0.39	0.41	0.28	0.35
	DS1	0.21	0.29	0.41	0.27	0.28	0.28
	DS2	0.21	0.28	0.48	0.34	0.30	0.31
40 Epochs	DS0	0.24	0.29	0.61	0.43	0.43	0.35
	DS1	0.20	0.25	0.25	0.05	0.22	0.08
	DS2	0.30	0.30	0.18	0.25	0.23	0.27
50 Epochs	DS0	0.28	0.33	0.48	0.39	0.35	0.35
	DS1	0.25	0.23	0.32	0.07	0.28	0.11
	DS2	0.27	0.18	0.27	0.20	0.27	0.19

Table 11 Accuracy for Dental class

Dental class		Precision		Recall		F1 score	
Learning Rate -->		0.0001	0.0002	0.0001	0.0002	0.0001	0.0002
Epochs	Dataset						
30 Epochs	DS0	0.33	0.19	0.11	0.25	0.17	0.22
	DS1	0.12	0.20	0.05	0.18	0.07	0.19
	DS2	0.08	0.23	0.07	0.43	0.07	0.30
40 Epochs	DS0	0.25	0.15	0.11	0.16	0.16	0.16
	DS1	0.23	0.12	0.34	0.07	0.27	0.09
	DS2	0.16	0.16	0.18	0.23	0.17	0.19
50 Epochs	DS0	0.18	0.19	0.14	0.14	0.16	0.16
	DS1	0.19	0.10	0.14	0.02	0.16	0.04
	DS2	0.22	0.15	0.30	0.16	0.25	0.15

Table 12 Accuracy for Labial class

Labial class		Precision		Recall		F1 score	
Learning Rate -->		0.0001	0.0002	0.0001	0.0002	0.0001	0.0002
Epochs	Dataset						
30 Epochs	DS0	0.41	0.32	0.27	0.36	0.33	0.34
	DS1	0.33	0.34	0.36	0.27	0.34	0.30
	DS2	0.31	0.38	0.45	0.11	0.37	0.18
40 Epochs	DS0	0.37	0.25	0.32	0.39	0.34	0.31
	DS1	0.33	0.32	0.14	0.23	0.19	0.27
	DS2	0.25	0.24	0.41	0.45	0.31	0.32
50 Epochs	DS0	0.34	0.34	0.45	0.23	0.39	0.27
	DS1	0.40	0.34	0.05	0.25	0.08	0.29
	DS2	0.25	0.30	0.41	0.36	0.31	0.33

4.2 Observations

- Accuracy is better for LR 0.0002 than 0.0001
- Accuracy is increases as number of Epochs are increasing
- In Guttural (ka, kha, ga,gha), Palatal (cha, chha, ja, jha) and Labial (pa, pha, ba, bha) class accuracy is increasing with LR and no of Epochs.
    - This is because more lip movement compared to other class.
- For Dental (ta, tha, da, dha) accuracy is increasing as no of epochs increasing but effect of LR is changing.
- For Retroflex (tta, thha, dda, ddha) 5 epochs give best accuracy.
- Epochs from 30 to 60 epochs give overall better accuracy for all classes.
- Higher learning rate also gives good accuracy.

5. CONCLUSION

- Accuracy for Alphabet depends on
    - Learning rate
    - No of Epochs
    - Class
- Higher the LR and No of Epochs, accuracy is more.
- Large Number of Epochs are needed because we are using 2D image with depth.
- Accuracy depends on speaking style of speakers and dataset.

6. REFERENCES

1. Werner, D. (1987). Disabled village children. The Hesperian Fdn, Palo Alto.
2. Muljono, M., Saraswati, G., Winarsih, N., Rokhman, N., Supriyanto, C., & Pujiono, P. (2019). Developing BacaBicara: An Indonesian Lipreading System as an Independent Communication Learning for the Deaf and Hard-of-Hearing. International Journal of Emerging Technologies in Learning (iJET), 14(4), 44-57.
3. Juang, B. H., & Rabiner, L. R. (2005). Automatic speech recognition–a brief history of the technology development. Georgia Institute of Technology. Atlanta Rutgers University and the University of California. Santa Barbara, 1, 67.
4. Kyle, F. E., & Harris, M. (2006). Concurrent correlates and predictors of reading and spelling achievement in deaf and hearing school children. *The Journal of Deaf Studies and Deaf Education*, *11*(3), 273-288
5. Koller, O., Camgoz, N. C., Ney, H., & Bowden, R. (2019). Weakly supervised learning with multi-stream CNN-LSTM-HMMs to discover sequential parallelism in sign language videos. IEEE transactions on pattern analysis and machine intelligence, 42(9), 2306-2320.
6. Zhou, H., Zhou, W., Zhou, Y., & Li, H. (2021). Spatial-temporal multi-cue network for sign language recognition and translation. IEEE Transactions on Multimedia, 24, 768-779.
7. Davis, K. H., Biddulph, R., & Balashek, S. (1952). Automatic recognition of spoken digits. The Journal of the Acoustical Society of America, 24(6), 637-642.
8. Olson, H. F., & Belar, H. (1956). Phonetic typewriter. *The Journal of the Acoustical Society of America*, *28*(6), 1072-1081.
9. Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. The journal of the acoustical society of america, 26(2), 212-215.
10. Petajan, E. D. (1984). Automatic lipreading to enhance speech recognition (speech reading) (Doctoral dissertation, University of Illinois at Urbana-Champaign).
11. Goldschen, A. J., Garcia, O. N., & Petajan, E. (1994, November). Continuous optical automatic speech recognition by lipreading. In Proceedings of 1994 28th Asilomar Conference on Signals, Systems and Computers (Vol. 1, pp. 572-577). IEEE.
12. Neti, C., Potamianos, G., Luettin, J., Matthews, I., Glotin, H., Vergyri, D., ... & Mashari, A. (2000). Audio visual speech recognition (No. REP_WORK). IDIAP.
13. Potamianos, G., Neti, C., Gravier, G., Garg, A., & Senior, A. W. (2003). Recent advances in the automatic recognition of audiovisual speech. Proceedings of the IEEE, 91(9), 1306-1326.
14. Zhao, G., Barnard, M., & Pietikainen, M. (2009). Lipreading with local spatiotemporal descriptors. IEEE Transactions on Multimedia, 11(7), 1254-1265.
15. Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., & Ng, A. Y. (2011, January). Multimodal deep learning. In ICML.
16. Noda, K., Yamaguchi, Y., Nakadai, K., Okuno, H. G., & Ogata, T. (2014). Lipreading using convolutional neural network. In fifteenth annual conference of the international speech communication association.
17. Wand, M., Koutník, J., & Schmidhuber, J. (2016, March). Lipreading with long short-term memory. In 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 6115-6119). IEEE.
18. Assael, Y. M., Shillingford, B., Whiteson, S., & De Freitas, N. (2016). Lipnet: End-to-end sentence-level lipreading. arXiv preprint arXiv:1611.01599.
19. Chung, J. S., Senior, A., Vinyals, O., & Zisserman, A. (2017, July). Lip reading sentences in the wild. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 3444-3453). IEEE.
20. Hao, M., Mamut, M., Yadikar, N., Aysa, A., & Ubul, K. (2020). A Survey of Research on Lipreading Technology. IEEE Access.
21. Wark, T., Sridharan, S., & Chandran, V. (1998, August). An approach to statistical lip modelling for speaker identification via chromatic feature extraction. In Proceedings. Fourteenth International Conference on Pattern Recognition (Cat. No. 98EX170) (Vol. 1, pp. 123-125). IEEE.
22. Lewis, T. W., & Powers, D. M. (2000, December). Lip feature extraction using red exclusion. In Selected papers from the Pan-Sydney workshop on Visualisation-Volume 2 (pp. 61-67).
23. Skodras, E., & Fakotakis, N. (2011, May). An unconstrained method for lip detection in color images. In 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 1013-1016). IEEE.
24. Ghaleh, V. E. C., & Behrad, A. (2010, September). Lip contour extraction using RGB color space and fuzzy c-means clustering. In 2010 IEEE 9th International Conference on Cyberntic Intelligent Systems (pp. 1-4). IEEE.
25. Gritzman, A. D., Rubin, D. M., & Pantanowitz, A. (2015). Comparison of colour transforms used in lip segmentation algorithms. Signal, Image and Video Processing, 9(4), 947-957.
26. Fan, X., Zhang, F., Wang, H., & Lu, X. (2012, May). The system of face detection based on OpenCV. In 2012 24th Chinese Control and Decision Conference (CCDC) (pp. 648-651). IEEE.
27. Puviarasan, N., & Palanivel, S. (2011). Lip reading of hearing impaired persons using HMM. Expert Systems with Applications, 38(4), 4477-4481.
28. Kass, M., Witkin, A., & Terzopoulos, D. (1988). Snakes: Active contour models. International journal of computer vision, 1(4), 321-331.
29. Nguyen, Q. D., & Milgram, M. (2008, August). Multi features active shape models for lip contours detection. In 2008 International Conference on Wavelet Analysis and Pattern Recognition (Vol. 1, pp. 172-176). IEEE.
30. Rothkrantz, L. (2017, May). Lip-reading by surveillance cameras. In 2017 Smart City Symposium Prague (SCSP) (pp. 1-6). IEEE.
31. Lee, C. G., Lee, E. S., Jung, S. T., & Lee, S. S. (2004). Design and implementation of a real-time lipreading system using PCA and HMM. Journal of Korea Multimedia Society, 7(11), 1597-1609.
32. Yao, J., & Kaifeng, Z. (2016, April). Evaluation model of the artist based on fuzzy membership to improve the principal component analysis of robust kernel. In 2016 IEEE 2nd International Conference on Big Data Security on Cloud (BigDataSecurity), IEEE International Conference on High Performance and Smart Computing (HPSC), and IEEE International Conference on Intelligent Data and Security (IDS) (pp. 322-326). IEEE.
33. Sterpu, G., & Harte, N. (2017). Towards Lipreading Sentences with Active Appearance Models. In AVSP (pp. 70-75).

34. Matthews, I., Potamianos, G., Neti, C., & Luettin, J. (2001, August). A comparison of model and transform-based visual features for audio-visual LVCSR. In IEEE International Conference on Multimedia and Expo, 2001. ICME 2001. (pp. 210-210). IEEE Computer Society.
35. Morade, S. S., & Patnaik, S. (2014). Lip reading by using 3-D discrete wavelet transform with dmey wavelet. International Journal of Image Processing (IJIP), 8(5), 384.
36. He, J., Zhang, H., & Liu, J. Z. (2009). LDA based feature extraction method in DCT domain in lipreading. Computer Engineering and Applications, 45(32), 150-155.
37. Liang, Y., Yao, W., & Du, M. (2010, October). Feature extraction based on lsda for lipreading. In 2010 International Conference on Multimedia Technology (pp. 1-4). IEEE.
38. Almajai, I., Cox, S., Harvey, R., & Lan, Y. (2016, March). 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.
39. Potamianos, G., Luettin, J., & Neti, C. (2001, May). Hierarchical discriminant features for audio-visual LVCSR. In 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221) (Vol. 1, pp. 165-168). IEEE.
40. Zhou, Z., Zhao, G., & Pietikäinen, M. (2011, June). Towards a practical lipreading system. In CVPR 2011 (pp. 137-144). IEEE.
41. Zhao, G., & Pietikainen, M. (2007). Dynamic texture recognition using local binary patterns with an application to facial expressions. IEEE transactions on pattern analysis and machine intelligence, 29(6), 915-928.
42. Shaikh, A. A., Kumar, D. K., Yau, W. C., Azemin, M. C., & Gubbi, J. (2010, October). Lip reading using optical flow and support vector machines. In 2010 3Rd international congress on image and signal processing (Vol. 1, pp. 327-330). IEEE.
43. Cappelletta, L., & Harte, N. (2011, August). Viseme definitions comparison for visual-only speech recognition. In 2011 19th European Signal Processing Conference (pp. 2109-2113). IEEE.
44. X. Ma, L. Yan, and Q. Zhong, ‘‘Lip feature extraction based on improved jumping-snake model,’’ in Proc. 35th Chin. Control Conf. (CCC), Jul. 2016, pp. 6928–6933.
45. Luettin, J., & Thacker, N. A. (1997). Speechreading using probabilistic models. Computer vision and image understanding, 65(2), 163-178.
46. Cootes, T. F., Edwards, G. J., & Taylor, C. J. (2001). Active appearance models. IEEE Transactions on pattern analysis and machine intelligence, 23(6), 681-685.
47. Watanabe, T., Katsurada, K., & Kanazawa, Y. (2016, November). Lip reading from multi view facial images using 3D-AAM. In Asian Conference on Computer Vision (pp. 303-316). Springer, Cham.
48. Petajan, E., Bischoff, B., Bodoff, D., & Brooke, N. M. (1988, May). An improved automatic lipreading system to enhance speech recognition. In Proceedings of the SIGCHI conference on Human factors in computing systems (pp. 19-25).
49. Bennani, Y., & Gallinari, P. (1991, April). On the use of TDNN-extracted features information in talker identification. In Acoustics, Speech, and Signal Processing, IEEE International Conference on (pp. 385-388). IEEE Computer Society.
50. Bregler, C., & Konig, Y. (1994, April). " Eigenlips" for robust speech recognition. In Proceedings of ICASSP'94. IEEE International Conference on Acoustics, Speech and Signal Processing (Vol. 2, pp. II-669). IEEE.
51. Sujatha, P., & Krishnan, M. R. (2012, February). Lip feature extraction for visual speech recognition using Hidden Markov Model. In 2012 International Conference on Computing, Communication and Applications (pp. 1-5). IEEE.
52. Thangthai, K., Bear, H. L., & Harvey, R. (2018). Comparing phonemes and visemes with DNN-based lipreading. arXiv preprint arXiv:1805.02924.
53. Viola, P., & Jones, M. J. (2004). Robust real-time face detection. International journal of computer vision, 57(2), 137-154.
54. Zhang, K., Zhang, Z., Li, Z., & Qiao, Y. (2016). Joint face detection and alignment using multitask cascaded convolutional networks. IEEE signal processing letters, 23(10), 1499-1503.
55. King, D. E. (2009). Dlib-ml: A machine learning toolkit. The Journal of Machine Learning Research, 10, 1755-1758.
56. Wand, M., & Schmidhuber, J. (2017). Improving speaker-independent lipreading with domain-adversarial training. arXiv preprint arXiv:1708.01565.
57. Wand, M., Schmidhuber, J., & Vu, N. T. (2018, April). Investigations on end-to-end audiovisual fusion. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 3041-3045). IEEE.
58. Petridis, S., Li, Z., & Pantic, M. (2017, March). End-to-end visual speech recognition with LSTMs. In 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP) (pp. 2592-2596). IEEE.
59. Sui, C., Bennamoun, M., & Togneri, R. (2015). Listening with your eyes: Towards a practical visual speech recognition system using deep boltzmann machines. In Proceedings of the IEEE International Conference on Computer Vision (pp. 154-162).
60. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.
61. Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078.
62. LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), 2278-2324.
63. Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009, June). Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition (pp. 248-255). Ieee.
64. Alom, M. Z., Taha, T. M., Yakopcic, C., Westberg, S., Sidike, P., Nasrin, M. S., ... & Asari, V. K. (2018). The history began from alexnet: A comprehensive survey on deep learning approaches. arXiv preprint arXiv:1803.01164.
65. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research, 15(1), 1929-1958.
66. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... & Rabinovich, A. (2015). Going deeper with convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1-9).
67. Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4700-4708).
68. Bear, H. L., Harvey, R. W., Theobald, B. J., & Lan, Y. (2014). Which phoneme-to-viseme maps best improve visual-only computer lip-reading?. In Advances in Visual Computing: 10th International Symposium, ISVC 2014, Las Vegas, NV, USA, December 8-10, 2014, Proceedings, Part II 10 (pp. 230-239). Springer International Publishing.

69. Garg, A., Noyola, J., & Bagadia, S. (2016). Lip reading using CNN and LSTM. Technical report, Stanford University, CS231 n project report.
70. Parkhi, O. M., Vedaldi, A., & Zisserman, A. (2015). Deep face recognition.
71. Li, Y., Takashima, Y., Takiguchi, T., & Ariki, Y. (2016, June). Lip reading using a dynamic feature of lip images and convolutional neural networks. In 2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS) (pp. 1-6). IEEE.
72. Chung, J. S., & Zisserman, A. (2017). Out of time: automated lip sync in the wild. In Asian conference on computer vision (pp. 251-263). Springer, Cham.
73. Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). Imagenet classification with deep convolutional neural networks. Communications of the ACM, 60(6), 84-90.
74. Saitoh, T., Zhou, Z., Zhao, G., & Pietikäinen, M. (2017). Concatenated frame image based CNN for visual speech recognition. In Asian Conference on Computer Vision (pp. 277-289). Springer, Cham.
75. Lin, M., Chen, Q., & Yan, S. (2013). Network in network. arXiv preprint arXiv:1312.4400.
76. Zhang, X., Gong, H., Dai, X., Yang, F., Liu, N., & Liu, M. (2019, July). Understanding pictograph with facial features: end-to-end sentence-level lip reading of Chinese. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 33, No. 01, pp. 9211-9218).
77. Ji, S., Xu, W., Yang, M., & Yu, K. (2012). 3D convolutional neural networks for human action recognition. IEEE transactions on pattern analysis and machine intelligence, 35(1), 221-231.
78. Tran, D., Bourdev, L., Fergus, R., Torresani, L., & Paluri, M. (2015). Learning spatiotemporal features with 3d convolutional networks. In Proceedings of the IEEE international conference on computer vision (pp. 4489-4497).
79. Tran, D., Ray, J., Shou, Z., Chang, S. F., & Paluri, M. (2017). Convnet architecture search for spatiotemporal feature learning. arXiv preprint arXiv:1708.05038.
80. Qiu, Z., Yao, T., & Mei, T. (2017). Learning spatio-temporal representation with pseudo-3d residual networks. In proceedings of the IEEE International Conference on Computer Vision (pp. 5533-5541).
81. Graves, A., & Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural networks, 18(5-6), 602-610.
82. Fung, I., & Mak, B. (2018, April). End-to-end low-resource lip-reading with maxout CNN and LSTM. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 2511-2515). IEEE.
83. Torfi, A., Iranmanesh, S. M., Nasrabadi, N., & Dawson, J. (2017). 3d convolutional neural networks for cross audio-visual matching recognition. IEEE Access, 5, 22081-22091.
84. Xu, K., Li, D., Cassimatis, N., & Wang, X. (2018, May). LCANet: End-to-end lipreading with cascaded attention-CTC. In 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018) (pp. 548-555). IEEE.
85. Srivastava, R. K., Greff, K., & Schmidhuber, J. (2015). Training very deep networks. Advances in neural information processing systems, 28.
86. Yang, S., Zhang, Y., Feng, D., Yang, M., Wang, C., Xiao, J., ... & Chen, X. (2019, May). LRW-1000: A naturally-distributed large-scale benchmark for lip reading in the wild. In 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019) (pp. 1-8). IEEE.
87. Stafylakis, T., & Tzimiropoulos, G. (2017). Combining residual networks with LSTMs for lipreading. arXiv preprint arXiv:1703.04105.
88. Margam, D. K., Aralikatti, R., Sharma, T., Thanda, A., Roy, S., & Venkatesan, S. M. (2019). LipReading with 3D-2D-CNN BLSTM-HMM and word-CTC models. arXiv preprint arXiv:1906.12170.
89. Xiao, J., Yang, S., Zhang, Y., Shan, S., & Chen, X. (2020, November). Deformation flow based two-stream network for lip reading. In 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020) (pp. 364-370). IEEE.
90. Luo, M., Yang, S., Shan, S., & Chen, X. (2020, November). Pseudo-convolutional policy gradient for sequence-to-sequence lip-reading. In 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020) (pp. 273-280). IEEE.
91. Zhang, Y., Yang, S., Xiao, J., Shan, S., & Chen, X. (2020, November). Can we read speech beyond the lips? rethinking roi selection for deep visual speech recognition. In 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020) (pp. 356-363). IEEE.
92. Zhao, X., Yang, S., Shan, S., & Chen, X. (2020, November). Mutual information maximization for effective lip reading. In 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020) (pp. 420-427). IEEE.
93. Petridis, S., Stafylakis, T., Ma, P., Cai, F., Tzimiropoulos, G., & Pantic, M. (2018, April). End-to-end audiovisual speech recognition. In 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP) (pp. 6548-6552). IEEE.
94. Petridis, S., Wang, Y., Li, Z., & Pantic, M. (2017). End-to-end audiovisual fusion with LSTMs. arXiv preprint arXiv:1709.04343.
95. Petridis, S., Wang, Y., Li, Z., & Pantic, M. (2017). End-to-end multi-view lipreading. arXiv preprint arXiv:1709.00443.
96. Saitoh, T., Morishita, K., & Konishi, R. (2008, December). Analysis of efficient lip reading method for various languages. In 2008 19th International Conference on Pattern Recognition (pp. 1-4). IEEE.
97. 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.
98. Lee, B., Hasegawa-Johnson, M., Goudeseune, C., Kamdar, S., Borys, S., Liu, M., & Huang, T. (2004). AVICAR: Audio-visual speech corpus in a car environment. In Eighth International Conference on Spoken Language Processing.
99. Zue, V., Seneff, S., & Glass, J. (1990). Speech database development at MIT: TIMIT and beyond. Speech communication, 9(4), 351-356.
100. Cox, S. J., Harvey, R. W., Lan, Y., Newman, J. L., & Theobald, B. J. (2008, September). The challenge of multispeaker lip-reading. In AVSP (pp. 179-184).
101. Ortega, A., Sukno, F., Lleida, E., Frangi, A. F., Miguel, A., Buera, L., & Zacur, E. (2004, May). AV@ CAR: A Spanish Multichannel Multimodal Corpus for In-Vehicle Automatic Audio-Visual Speech Recognition. In LREC.
102. Movellan, J. (1994). Visual speech recognition with stochastic networks. Advances in neural information processing systems, 7.
103. Vanegas, O., Tokuda, K., & Kitamura, T. (1999, October). Location normalization of HMM-based lip-reading: experiments for the M2VTS database. In Proceedings 1999 International Conference on Image Processing (Cat. 99CH36348) (Vol. 2, pp. 343-347). IEEE.
104. Messer, K., Matas, J., Kittler, J., Luettin, J., & Maitre, G. (1999, March). XM2VTSDB: The extended M2VTS database. In Second international conference on audio and video-based biometric person authentication (Vol. 964, pp. 965-966).

105. Patterson, E. K., Gurbuz, S., Tufekci, Z., & Gowdy, J. N. (2002, May). CUAVE: A new audio-visual database for multimodal human-computer interface research. In 2002 IEEE International conference on acoustics, speech, and signal processing (Vol. 2, pp. II-2017). IEEE.
106. Anina, I., Zhou, Z., Zhao, G., & Pietikäinen, M. (2015, May). Ouluvs2: A multi-view audiovisual database for non-rigid mouth motion analysis. In 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG) (Vol. 1, pp. 1-5). IEEE.
107. Petridis, S., Shen, J., Cetin, D., & Pantic, M. (2018, April). 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.
108. Chitu, A. G., Driel, K., & Rothkrantz, L. J. (2010). Automatic lip reading in the Dutch language using active appearance models on high speed recordings. In Text, Speech and Dialogue: 13th International Conference, TSD 2010, Brno, Czech Republic, September 6-10, 2010. Proceedings 13 (pp. 259-266). Springer Berlin Heidelberg.
109. Tamura, S., Miyajima, C., Kitaoka, N., Yamada, T., Tsuge, S., Takiguchi, T., ... & Nakamura, S. (2010). CENSREC-1-AV: An audio-visual corpus for noisy bimodal speech recognition. In Auditory-Visual Speech Processing 2010.
110. Igras, M., Ziółko, B., & Jadczyk, T. (2012). Audiovisual database of Polish speech recordings. Studia Informatica, 33(2B), 163-172.
111. Xu, Y., Du, L., Li, G., Wu, P., & Zhang, X. (2000). Chinese audiovisual bimodal speech database CAVSR1. 0. In Proc. Int. Symp. Chin. Spoken Lang. Process. (pp. 98-101).
112. Rekik, A., Ben-Hamadou, A., & Mahdi, W. (2015). Human machine interaction via visual speech spotting. In Advanced Concepts for Intelligent Vision Systems: 16th International Conference, ACIVS 2015, Catania, Italy, October 26-29, 2015. Proceedings 16 (pp. 566-574). Springer International Publishing.
113. Chung, J. S., & Zisserman, A. (2017). Lip reading in the wild. In Computer Vision–ACCV 2016: 13th Asian Conference on Computer Vision, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part II 13 (pp. 87-103). Springer International Publishing.
114. Sanderson, C. (2002). The vidtimit database (No. REP_WORK). IDIAP.
115. Hazen, T. J., Saenko, K., La, C. H., & Glass, J. R. (2004, October). A segment-based audio-visual speech recognizer: Data collection, development, and initial experiments. In Proceedings of the 6th international conference on Multimodal interfaces (pp. 235-242).
116. McCool, C., Marcel, S., Hadid, A., Pietikäinen, M., Matejka, P., Cernocký, J., ... & Cootes, T. (2012, July). Bi-modal person recognition on a mobile phone: using mobile phone data. In 2012 IEEE international conference on multimedia and expo workshops (pp. 635-640). IEEE.
117. Wang, J., Wang, L., Zhang, J., Wei, J., Yu, M., & Yu, R. (2018, June). A large-scale depth-based multimodal audio-visual corpus in mandarin. In 2018 IEEE 20th International Conference on High Performance Computing and Communications; IEEE 16th International Conference on Smart City; IEEE 4th International Conference on Data Science and Systems (HPCC/SmartCity/DSS) (pp. 881-885). IEEE.
118. Lin, X., Yao, H., Hong, X., & Wang, Q. (2008, December). HIT-AVDB-II: A new multi-view and extreme feature cases contained audio-visual database for biometrics. In 11th Joint International Conference on Information Sciences (pp. 357-363). Atlantis Press.
119. Petrovska-Delacrétaz, D., Lelandais, S., Colineau, J., Chen, L., Dorizzi, B., Ardabilian, M., ... & Amor, B. B. (2008, September). The iv 2 multimodal biometric database (including iris, 2d, 3d, stereoscopic, and talking face data), and the iv 2-2007 evaluation campaign. In 2008 IEEE Second International Conference on Biometrics: Theory, Applications and Systems (pp. 1-7). IEEE.
120. Petr, C., Miloš, Ž., Zdeněk, K., Jakub, K., Jan, Z., & Luděk, M. (2005). DESIGN AND RECORDING OF CZECH SPEECH CORPUS FOR AUDIO-VISUAL CONTINUOUS SPEECH RECOGNITION.
121. Trojanová, J., Hrúz, M., Campr, P., & Železný, M. (2008). Design and recording of czech audio-visual database with impaired conditions for continuous speech recognition.
122. Verkhodanova, V., Ronzhin, A., Kipyatkova, I., Ivanko, D., Karpov, A., & Železný, M. (2016). HAVRUS corpus: high-speed recordings of audio-visual Russian speech. In Speech and Computer: 18th International Conference, SPECOM 2016, Budapest, Hungary, August 23-27, 2016, Proceedings 18 (pp. 338-345). Springer International Publishing.
123. Bachman, G., Le-Jan, G., Souviraà-Labastie, N., & Bimbot, F. (2011). BL-Database: A French audiovisual database for speech driven lip animation systems (Doctoral dissertation, INRIA).
Le, Y. B. G. B. G., & Bimbot, J. N. S. L. F. (2011). BL-Database: A French audiovisual database for speech driven lip animation systems.
124. Fernandez-Lopez, A., Martinez, O., & Sukno, F. M. (2017, May). 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.
125. 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.
126. Afouras, T., Chung, J. S., Senior, A., Vinyals, O., & Zisserman, A. (2018). Deep audio-visual speech recognition. IEEE transactions on pattern analysis and machine intelligence, 44(12), 8717-8727.
127. Chung, J. S., & Zisserman, A. P. (2017). Lip reading in profile.
128. Afouras, T., Chung, J. S., & Zisserman, A. (2018). LRS3-TED: a large-scale dataset for visual speech recognition. arXiv preprint arXiv:1809.00496.
129. Shillingford, B., Assael, Y., Hoffman, M. W., Paine, T., Hughes, C., Prabhu, U., ... & de Freitas, N. (2018). Large-scale visual speech recognition. arXiv preprint arXiv:1807.05162.
130. Lan, Y., Theobald, B. J., & Harvey, R. (2012, July). View independent computer lip-reading. In 2012 IEEE International Conference on Multimedia and Expo (pp. 432-437). IEEE.
131. Estival, D., Cassidy, S., Cox, F., & Burnham, D. (2014). AusTalk: an audio-visual corpus of Australian English.
132. Bailly-Bailliére, E., Bengio, S., Bimbot, F., Hamouz, M., Kittler, J., Mariéthoz, J., ... & Thiran, J. P. (2003). The BANCA database and evaluation protocol. In Audio-and Video-Based Biometric Person Authentication: 4th International Conference, AVBPA 2003 Guildford, UK, June 9–11, 2003 Proceedings 4 (pp. 625-638). Springer Berlin Heidelberg.
133. Fox, N. A., O'Mullane, B. A., & Reilly, R. B. (2005). VALID: A new practical audio-visual database, and comparative results. In Audio-and Video-Based Biometric Person Authentication: 5th International Conference, AVBPA 2005, Hilton Rye Town, NY, USA, July 20-22, 2005. Proceedings 5 (pp. 777-786). Springer Berlin Heidelberg.
134. Lucey, P., Potamianos, G., & Sridharan, S. (2008). Patch-based analysis of visual speech from multiple views. In AVSP (pp. 69-74).

135. Morade, S. S., & Patnaik, S. (2014). Lip reading by using 3-D discrete wavelet transform with dmey wavelet. International Journal of Image Processing (IJIP), 8(5), 384.
136. Pei, Y., Kim, T., & Zha, H. (2013). Unsupervised Random Forest Manifold Alignment for Lipreading. 2013 IEEE International Conference on Computer Vision, 129-136.
137. Petridis, S., & Pantic, M. (2016, March). Deep complementary bottleneck features for visual speech recognition. In 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 2304-2308). IEEE.
138. Hu, D., & Li, X. (2016). Temporal multimodal learning in audiovisual speech recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 3574-3582).
139. Papandreou, G., Katsamanis, A., Pitsikalis, V., & Maragos, P. (2009). Adaptive multimodal fusion by uncertainty compensation with application to audiovisual speech recognition. IEEE Transactions on Audio, Speech, and Language Processing, 17(3), 423-435.
140. Rahmani, M. H., & Almasganj, F. (2017, April). Lip-reading via a DNN-HMM hybrid system using combination of the image-based and model-based features. In 2017 3rd International Conference on Pattern Recognition and Image Analysis (IPRIA) (pp. 195-199). IEEE.
141. Kamper, H., Wang, W., & Livescu, K. (2016, March). Deep convolutional acoustic word embeddings using word-pair side information. In 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 4950-4954). IEEE.
142. Tailor, J. H., & Shah, D. B. (2015). Review on Speech Recognition System for Indian Languages. International Journal of Computer Applications, 119(2).
143. Nandini, M. S., Nagavi, T. C., & Bhajantri, N. U. (2019, March). Deep Weighted Feature Descriptors for Lip Reading of Kannada Language. In 2019 6th International Conference on Signal Processing and Integrated Networks (SPIN) (pp. 978-982). IEEE.
144. Patil, M. S., Chickerur, S., Meti, A., Nabapure, P. M., Mahindrakar, S., Naik, S., & Kanyal, S. (2019). LSTM Based Lip Reading Approach for Devanagiri Script.
145. Parikh, R. B., & Joshi, H. (2020). Gujarati Speech Recognition–A Review. no, 549, 6.
146. Patel, J., & Nandurbarkar, A. (2015). Development and implementation of algorithm for speaker recognition for gujarati language. International Research Journal of Engineering and Technology, 2(2), 444-448
147. Vijayendra, A. D., & Thakar, V. K. (2016). Neural network based Gujarati speech recognition for dataset collected by in-ear microphone. Procedia computer science, 93, 668-675.
148. Valaki, S., & Jethva, H. (2017, March). A hybrid HMM/ANN approach for automatic Gujarati speech recognition. In 2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS) (pp. 1-5). IEEE.
149. Tailor, J. H., & Shah, D. B. (2018). HMM-based lightweight speech recognition system for gujarati language. In Information and Communication Technology for Sustainable Development: Proceedings of ICT4SD 2016, Volume 2 (pp. 451-461). Springer Singapore.
150. Raval, D., Pathak, V., Patel, M., & Bhatt, B. (2020, December). End-to-End Automatic Speech Recognition for Gujarati. In Proceedings of the 17th International Conference on Natural Language Processing (ICON) (pp. 409-419).
151. Pandit, P., Bhatt, S., & Makwana, P. Automatic speech recognition of Gujarati digits using artificial neural network. In Proceedings of 19th Annual Cum 4th International Conference of GAMS On Advances in Mathematical Modelling to Real World Problems (pp. 141-146).
152. Tailor, J. H., Rakholia, R., Saini, J. R., & Kotecha, K. (2022). Deep Learning Approach for Spoken Digit Recognition in Gujarati Language. International Journal of Advanced Computer Science and Applications, 13(4).
153. Fernandez-Lopez, A., & Sukno, F. M. (2017). Automatic viseme vocabulary construction to enhance continuous lip-reading. arXiv preprint arXiv:1704.08035.
154. Bhaskar, S., & Thasleema, T. M. (2023). LSTM model for visual speech recognition through facial expressions. Multimedia Tools and Applications, 82(4), 5455-5472.
155. Rudregowda, S., Patil Kulkarni, S., HL, G., Ravi, V., & Krichen, M. (2023, March). Visual Speech Recognition for Kannada Language Using VGG16 Convolutional Neural Network. In *Acoustics* (Vol. 5, No. 1, pp. 343-353). MDPI.

7. PUBLICATIONS

1. “A Brief study on Lip-Reading Methods” in 13^{th} International conference on Science and Innovative Engineering 2023

2. “An insight into Lip-Reading Dataset and Languages” in 13^{th} International conference on Science and Innovative Engineering 2023

3. “ViLiDEx- A Lip Extraction Algorithm for Lip Reading” in International Journal on Recent and Innovation Trends in Computing and Communication, 11(9), 3672-3675.