International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

ViLiDEx- A Lip Extraction Algorithm for Lip Reading

Ms. Vibhavari Patel
Research Scholar, CSE Department, Faculty of Technology & Engineering,
The M S University of Baroda, Vadodara, Gujarat, India
vibhavarihpatel@gmail.com

Dr. Vishwas Raval
Associate Professor, Department for Studies in Strategic Technologies (Cyber/Space),
School of National Security Studies, Central University of Gujarat, Gandhinagar, Gujarat, India
vishwas.raval@cug.ac.in

Abstract— *Technology is evolving at an immense speed every day. In the lap of technology, computer vision and machine learning are also growing fast. Many real time applications are running without human interaction just because of Computer vision and machine learning. In this paper, we are using computer vision and machine learning for lip feature extraction for Gujarati language. For this task we have created dataset GVLetters for Gujarati alphabets. We have taken videos of 24 speakers for 33 alphabets of Guajarati language. Face landmark algorithm from dlib is used for deriving ViLiDEx (Vibhavari's algorithm for Lip Detection and Extraction). ViLiDEx is applied for 24 speakers and 5 alphabets from each class (Guttural, Palatal, Retroflex, Dental and Labial). This algorithm calculates total number of frames for each speaker, keep 20/25 frames as a dataset and removes extra frames. Depending on number of frames, frame numbers divisible by prime numbers are chosen for removal.*

Keywords- *Lip extraction and detection, Speech recognition, dataset, Indian language, Dlib library*

I. INTRODUCTION

A great deal of research was carried out in the field of Speech recognition system. Speech recognition is the ability of a machine or a program to interpret audio signals or sequence of images and convert them into specific language. Apple's Siri and Google's Alexa are applications of AI based speech recognition systems which interprets audio signals. When Image processing algorithms were developed, Video based speech recognition applications were designed for different languages. Different Face detection and lip extraction methods are used for video-based speech recognition. Shape based or face-structure based methods and model-based methods were used for lip detection and extraction. After rising AI and machine learning techniques, many pre-trained models are used for this task. Video based speech recognition methods are useful to learn different languages for hearing-impaired people. These methods can be used in the noisy environment where audio recognition is difficult to implement. These methods can also be used in video surveillance to track the speech of a particular person.

Here in this paper we are using pre-trained model Dlib[11] for lip detection and extraction. Here we have used our own dataset GVLetters for Gujarati language. GVLetters consists of 24 speakers, each speaking 34 alphabets, and three shots for each alphabet. Gujarati alphabets are classified in five different classes: guttural, palatal, retroflex, dental and labial. We have modified face landmark algorithm for lip detection and extraction. Our modified algorithm ViLiDEX counts the total number of frames for each alphabet video, removes the extra frames having sequence number match with prime numbers listed by the algorithm, extract lip area of remaining 20 frames and store them. Frame numbers divisible by prime numbers (2,3,5,7,11 …) will be discarded.

II. RELATED WORK

A. History of lip detection and extraction

Lip detection and extraction has a history of seven decades. Different factors like complexity of videos, different types of background like static and rotating, different face structures of speakers etc. make this task challenging. Some speakers have short lip movements as compared to others. Speakers from different regions have different accents, different style and different angle of speech. Hence there is a need to create robust models of automated speech recognition.

Lip detection and extraction is mainly divided into three steps: Lip detection extraction, feature extraction and classification. The first step, Lip detection and extraction involves locating and extracting area of interest (ROI) from raw data. After detecting the ROI, in the second step, effective features are extracted which will be given as input for further transformation. Transformation will reduce the dimensions of features which will be used in the last step for final classification. There are mainly two approaches for lip detection and extraction. In first approach, various image processing methods are used for ROI extraction and feature extraction algorithm was designed based on image processing algorithms.[14][18] In second approach, different pre-trained models are used for ROI detection and iterative learning methods are designed to automatically extract the features. Here we are using Dlib Library [11] for lip detection and extraction.

Different datasets of different languages are available for

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

this task. Lot of research work carried out for different languages like English, Chinese, Japanese, Russian, Persian etc. we have focused on Gujarati language which is one of the most widely spoken Indo-Aryan languages of India. Most Indian languages and hence Gujarati language is derived from Devanagari scripts.

B. Indian Languages and Devanagari scripts

Most Indian languages are derived from Devanagari scripts. Languages derived from Devanagari scripts have some special characteristics compared to The English language and other non-Indian languages. They have a scientific way of speaking wherein the alphabets are categorized based on how they are spoken. The arrangements of letters in the Gujarati language are called ***"Mulakshar"*** which means basic letters. In English alphabets, the arrangement of letters is not logical. There is no reason why vowels are scattered around in the alphabet set or why the letter G comes before the letter H. in Devanagari script and hence all languages derived from it, consonants, and vowels are categorized separately. The alphabets (vowels and consonants) are arranged based on where and how the sound of that letter is produced inside the mouth. For easiness and as we are working on it, we discuss for Gujarati language only.

C. Characteristics of Gujarati Language

There are 36 consonants and 12 vowels in the Gujarati language. Unlike the English language, vowels are not scattered in between consonants. They are separated and arranged based on where and how the sound is produced while speaking.

Classification of consonants based on spoken style for the Gujarati language is shown in Figure 1. The first five consonants as shown in Figure 1 are called guttural as the sound of these consonants comes from the throat. Similarly, palatal group consonants are articulated when the tongue touches the hard palate. Retroflex group consonants are articulated when the tongue curls back a bit and touches the roof of the palate. A dental group of consonants is produced when the tongue touches the upper teeth and labial is produced using lips.

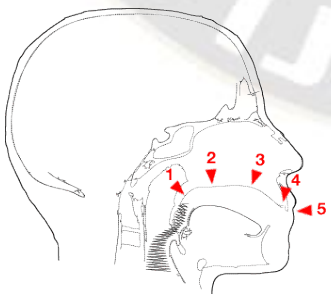

Figure 1. Spoken style of alphabets (1. Guttural, 2. Palatal, 3. Retroflex, 4. Dental, 5. Labial)

Name of the class	Alphabets of the class	Spoken by
Guttural	'ક', 'ખ', 'ગ', 'ઘ', 'ઙ', 'હ'	back of the tongue touches the velum
Palatal	'ચ', 'છ', 'જ', 'ઝ', 'ઞ', 'ય', 'શ'	the tongue touches the hard palate
Retroflex	'ટ', 'ઠ', 'ડ', 'ઢ', 'ણ', 'ર', 'ષ'	the tongue curls back a bit and touches the alveolar ridge
Dental	'ત', 'થ', 'દ', 'ધ', 'ન', 'લ', 'સ'	the tongue touches the back of the teeth
Labial	'પ', 'ફ', 'બ', 'ભ', 'મ', 'વ'	rounded lips

Table 1. Classification of Devanagari Alphabets

Gujarati language has special alphabets like 'ણ', 'ળ', 'ઙ', 'ઞ', 'છ', 'ક્ષ', 'જ્ઞ', 'ત્ર', 'શ્ર' and these alphabets have no specific equivalent unique alphabet in The English language. People, who have a native language as English, cannot pronounce these alphabets easily. Combinations of alphabets in these languages have the same spellings in English. If we want to differentiate them, we need to check their phonic as shown in Table 2.

Gujarati Alphabet	English spelling	Phonics
ટ, ત	Ta	ṭa, ta
થ, ઠ	Tha	tha, ṭha
ડ, દ	Da	ḍa , da
ઢ, ધ	Dha	ḍhe, dhe
ષ, સ	Se	ṣe, se
ળ, લ	La	ḷe, la
ન, ણ	Na	ne, ṇe

Table 2. Alphabets and their corresponding phonics

III. DLIB TOOLKIT

Dlib toolkit is a cross platform, an open source library written in C++ language, provides environment for developing machine learning software. Dlib design is based on contract and component-based software engineering. It is a collection of independent software components, each accompanied by documentation and debugging modes. This library is useful in both research and real world projects. Dlib library is general purpose library which contains graphical applications to create Bayesian networks and various tools for handling threads, network I/O, and other tasks.

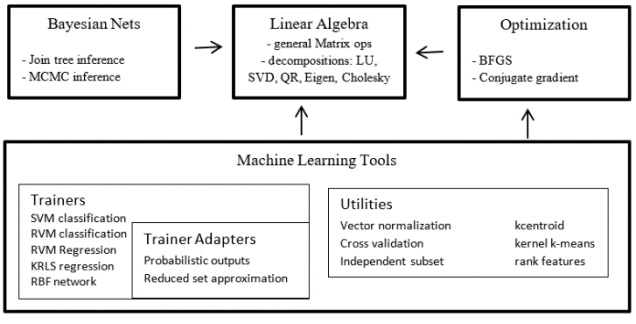

Figure 2. Dlib-ML toolkit with dependency

The four components of Dlib toolkit (Figure 2) are linear algebra, machine learning tools, Bayesian nets and optimization. The linear algebra component provides core functionality while remaining three provides various tools.

IV. PROPOSED WORK – VILIDEX ALGORITHM

Here we are going to show about lip detection, extraction and database creation for Gujarati Alphabets. We have taken

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 September 2023*

videos using Nikon D 5300 camera with 1920 X 1080 full HD resolutions and 30frames/second. 3 shots of 24 speakers are taken for 33 alphabets; total 2376 samples are collected for 5 classes of Gujarati Alphabet. 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. Among 25 speakers, 4 are school going children. Here we have modified face landmark algorithm to detect and extract lips for 5 alphabets from each class and extracted 20 and 25 frames for each.

Facial landmarks 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 3).

Figure 3. Landmarks of face

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. Prime numbered (2, 3, 5, 7, 11, 17…) frames will be removed depending on total number of frames. A Flow chart of ViLiDEx algorithm is given below.

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. 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.

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 3. ViLiDEX algorithm

ViLiDEX algorithm (Table 3) is designed to remove extra frames from an input video and extract lip area from the remaining 20 frames and store. This algorithm successfully removes extra frames. If total frames are too large, then only this algorithm extracts 20 key frames correctly (Table 4).

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

Table 5. Working of ViLiDEX algorithm

For lip extraction face landmark algorithm uses 49-68 points, which extracts lip area accurately. This algorithm is not able to extract the lip area when the speaker has an abrupt change in face movement. For 24 speakers, accuracy of this algorithm is 95.83%. Figure 4 (a, b) shows 20 frames for alphabet 'Ka', where all frames are extracted correctly. In figure 4 (b) frame number 9 and 10 are not extracted correctly due to abrupt change in face movement of speaker.

Figure 4(a). Frames correctly extracted for alphabet 'Ka'

Figure 4(b). Frames extracted erroneously for alphabet 'Ka'

V. FUTURE WORK AND LIMITATIONS

This algorithm removes extra frames and extracts lip area from the remaining 20 frames and store. Frame removal task is based on prime numbers that works well with any number of total frames. If the number of frames increases, it may lose key frames needed for feature extraction. This algorithm can be improved to remove duplicate and repeating frames at start and end position. In figure 5, the initial 8 frames are similar and not needed for feature extraction. Such frames should be replaced with one frame only, so other key frames could be included.

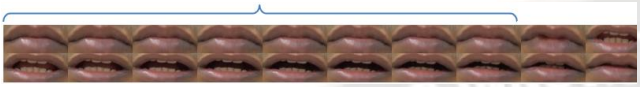

Figure 5. Similar frames those must be replaced with one frame

VI. CONCLUSION

ViLiDEx algorithm removes extra frames using prime number-based frame removal method and keeps 20 frames needed for feature extraction. Total 24 speakers uttered 34 alphabets in 3 shots to make GVLetters dataset for Gujarati language. This algorithm gives 95.83% accuracy for lip detection and extraction.

ACKNOWLEDGMENTS

We are thankful to the Omnipotent God for making us able to do something for the society. We are thankful to our parents for bringing us to this beautiful planet. We are grateful to our department and University for providing support and resources for this work. Finally, we acknowledge the authors and researchers whose papers helped us to move ahead with this work.

REFERENCES

[1] 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.

[2] Datta, A. K., & Ganguli, N. R. (1980). Automatic Speech Recognition in Intelligence Communication. IETE Journal of Research, 26(1), 82-84.

[3] Pearson, D. (1981). Visual communication systems for the deaf. IEEE Transactions on Communications, 29(12), 1986-1992.

[4] Paliwal, K. K., Sinha, S. S., & Agarwal, A. (1983). An isolated word recognition system for Hindi digits using linear time normalization. IETE Journal of Research, 29(1), 18-22.

[5] Viola, P., & Jones, M. J. (2004). Robust real-time face detection. *International journal of computer vision*, *57*(2), 137-154.

[6] Furui, S. (2005). 50 years of progress in speech and speaker recognition research. ECTI Transactions on Computer and Information Technology (ECTI-CIT), 1(2), 64-74.

[7] Hong, X., Yao, H., Wan, Y., & Chen, R. (2006, December). A PCA based visual DCT feature extraction method for lip-reading. In 2006 International Conference on Intelligent Information Hiding and Multimedia (pp. 321-326). IEEE.

[8] 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

[9] 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.

[10] Gunes, H., & Piccardi, M. (2008). Automatic temporal segment detection and affect recognition from face and body display. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 39(1), 64-84.

[11] King, D. E. (2009). Dlib-ml: A machine learning toolkit. The Journal of Machine Learning Research, 10, 1755-1758.

[12] Anusuya, M. A., & Katti, S. K. (2010). Speech recognition by machine, a review. arXiv preprint arXiv:1001.2267.

[13] Barkhan, M., Alizadeh, F., & Maihami, V. (2019). Designing and implementing a system for Automatic recognition of Persian letters by Lip-reading using image processing methods. Journal of Advances in Computer Engineering and Technology, 5(2), 71-80.

[14] Mestri, R., Limaye, P., Khuteta, S., & Bansode, M. (2019, April). Analysis of Feature Extraction and Classification Models for Lip-Reading. In 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI) (pp. 911-915). IEEE.

[15] Huang, Y., Chen, F., Lv, S., & Wang, X. (2019). Facial expression recognition: A survey. Symmetry, 11(10), 1189.

[16] 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.

[17] Mesbah, A., Berrahou, A., Hammouchi, H., Berbia, H., Qjidaa, H., & Daoudi, M. (2019). Lip reading with Hahn convolutional neural networks. Image and Vision Computing, 88, 76-83.

[18] Hao, M., Mamut, M., Yadikar, N., Aysa, A., & Ubul, K. (2020). A Survey of Research on Lipreading Technology. IEEE Access.

[19] Parikh, R. B., & Joshi, H. (2020). Gujarati Speech Recognition–A Review. no, 549, 6.