Visual Speech and Speaker Recognition for Marathi and Gujarati language using Deep learning

[1] Ramesh R. Naik, [2]Jaiprakash Verma, [3]Mohd Zuhair, [4]Priya Mehta, [5]Tisha Kotadia
[1]Assistant Professor, [2]Associate Professor, [3]Assistant Professor,[4]Student, [5]Student,
Institute of Technology, Nirma University, Ahmedabad, Gujarat, India.

Abstract

People use coordinated multisensory inputs to perceive their surroundings in a multimodal fashion. Numerous studies in psychology and cognitive science show that the processing and interpretation of human speech is multimodal complicated. These discoveries served as the impetus for the emerging and quickly expanding study field of audio-visual speech recognition. Up to this point, extensive research has shown that systems that combine visual and auditory information perform better than their conventional blind versions. The goal of this work is Design and development of a working tool of Visual Speech and Speaker Recognition. To examine visual elements and assess their importance in voice recognition from an automated point of view. The region around the lips contains speech-related information, and obtaining its representation in parameterized form is a vital necessity. This difficult topic entails tough vision tasks including face detection and facial tracking. Next face detection, it's crucial to reliably detect the mouth area across the following video sequences. A method for lip boundary and lip area tracking will be developed in this study, and approaches for visual speech and speaker recognition will receive the majority of our attention. Support Vector Machine and Deep learning.

Keyword: *Speech, Speaker Recognition, Deep learning, visual speech.*

1. INTRODUCTION

Speech is the means by which people communicate with one another. Speech is used by humans to connect or communicate with one another. Visual speaking human face with an eye-catching expression is referred to as visual speech. Visual speech recognition, often known as lip reading, is the process of understanding speech only from visual input. A speaker is captured on camera with video data. The speaker's face movements is minimized in the enormous data stream being reported. The movement of the tongue, chin, teeth, and lips is taken into account [1, 2]. Speaker Recognition is just a computer programme that recognises the speaker or the person speaking. The speaker might be identified using audiovisual information or even only through visual cues. Based on visual cues

derived from the speaker's face region, the system detects tracks and recognises speakers automatically. Despite the wide range of existing strategies for visual speech recognition, numerous methods have been put forth for solving the problem in the literature. However, research is still being done in this field to find the best features and classification techniques to discriminate as well as possible between different mouth shapes while keeping the mouth shapes produced by different people in the same class. And demand as little processing of the mouth picture [2]. For the purpose of reading human lips, the visibility and relationships of the tongue, lower teeth, and upper teeth are crucial [3]. According to study, the distance between the lips, both horizontally and vertically, fluctuates depending on the proximity of similar words. The lip-reading task is even difficult when there is no frontal view of the face. To handle these situations, a pose normalization block is introduced in a standard system and generates virtual frontal views from non-frontal images [5].
When using visual speech recognition, it's also important to consider factors like phrasing, stress, intonation, emotions, and the lifting and shape of the eyebrows [6].

2. LITERATURE REVIEW

In the literature, various approaches have been put out to address the issue of visual speech recognition. Regarding the feature kinds, the classifier employed, and the class dentitions, the various sorts of solutions adopted vary greatly. While Support Vector Machine with Viterbi Algorithm achieved good word recognition rates in comparison to the state-of-the-art results from the literature, Automatic segmentation of Visual Speech Corpora Based found the time boundaries between subsequent phonemes in the corpus using Hidden Markov Model with Viterbi Algorithm [1, 2]. The majority of academics have attempted to compare the two photos using conventional image comparison methods. Although one attempted to investigate the usage of neural networks in lip reading approaches, the results obtained were of lesser efficiency taking into account the efficiency reducing limits of the image comparison technique utilised in the approach [4]. For interpolating between frames in a picture sequence and for generating features for recognition, Bergeler and Stephan utilised manifold for tracking and extracting lip movement [6]. Luettin creates whole-word hidden Markov models (HMM) for visual speech recognition after first using active shape models to represent various mouth forms and grey level distribution profiles (GLDPs) around the outer and/or inner lip contours as feature vectors [7]. Movellan utilises HMMs for generating visual word models as well, but after some straightforward pre-processing to take advantage of the vertical symmetry of the mouth, he uses the grey levels of mouth images directly as features [8].Past research on lip reading has mostly focused on frontal face lip reading, completely ignoring the fact that image-based cues from the profile view (PV)

can also be useful for lip reading [9]. In order to condense the enormous amount of data into a reasonable set of low-level picture statistics representing the region of interest around the mouth, video preprocessing employs a number of straightforward algorithms [10]. It is possible to think of the parameterized face as a fully compatible addition to the current KTH text-to-speech system [11]. Mostly used data corpora for visual speech (lip reading) are TULIPS1, AVletters, CUAVE, DUTAVSC, AVICAR, XM2VTSDB, M2VTS, and Vid-TIMIT. A comparison of the characteristics and stated purpose is presented in [23]. Among these data corpora, M2VTS is in French, XM2VTS is in 4 languages and DUTAVSC is in Dutch language, rest all are in English only. Also, many databases are not publically available. So, the researcher has to build their own dataset. The set of guidelines for building data corpus for visual speech recognition are given in [24]. Language quality is also an important parameter for lip reading and the database should have good coverage of the phonemes and visems of the targeted language. It should contain phonetically rich words and sentences. Most of existing data corpora are built for small recognition task like digit recognition or letter recognition. In our proposed research work we will mainly focus on Marathi and Gujarati and Language. So, we need to build our own dataset in Marathi and Gujarati Language.

The following Table1 Shows the literature review of Audio-Visual Databases

Database Name	Speakers	Database
Audio-Visual speech In a Car (AVICAR).[12]	100 (50 Male, 50 Female)	20 phone numbers with 10 digits each, 20 sentences randomly chosen out of 450 TIMIT sentences for each speaker
AVLetters [13]	10 (5 Male, 5 Female)	A-Z Letters
AVLetter2[14]	5	26 isolated letter seven times
AV-TIMIT. [15]	233 (117 Male, 106 Female)	Utterances from phonetically balanced TIMIT sentences
Clemson University Audio-Visual Experiments (CUAVE) [16]	36 (17 Male, 19 female)	Isolated and Continuous digits with head movement in side-to-side, back-forth and tilting.
IBMSmart-Room[17]	38	Continuous digit string

Language Independent Lip-Reading (LILiR) [18]	20	Resource management Corpus
MOBIO[19]	152 (100 Male, 52 Female)	short-response questions, free-speech questions
OuluVS[20]	20 (17 Male, 3 Female)	10 daily use English phrases per speaker with nine repetition each
Grid AV[21]	34 (18 Male, 16 Female)	1000 Sentences per speaker
XM2VTSDB[22]	295	Continuous Digits

3. PROPOSED SYSTEM

In our proposed model, there are five steps. In First step we have collected the database. Second step pre-processing Third step contains Feature Extraction After Feature Extraction Classification has been done. All the Steps of our model are explained in detail, in the Following Sections. Following figure 1 shows the detail outline of the complete proposed process.

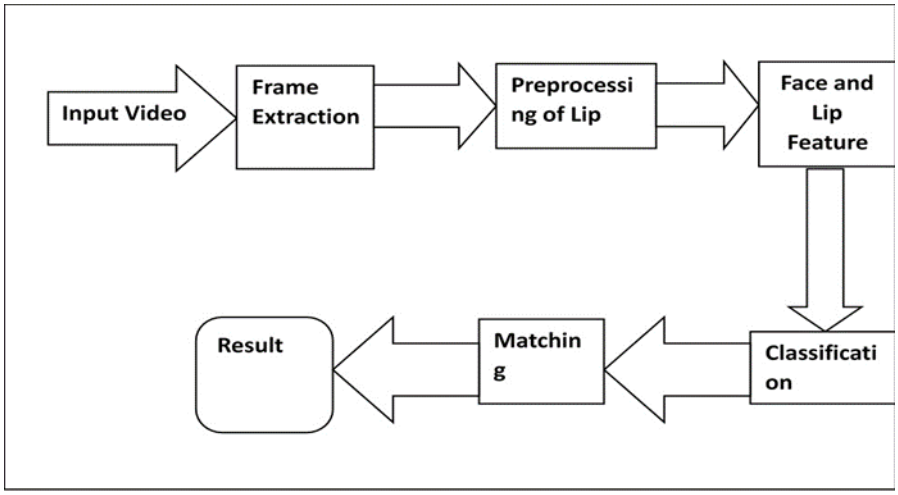


Fig. (1). system architecture

Steps

1.Collection of Database - The first step in the proposed system implementation is capturing frames from a video source.

2.Pre-processing - The second step in the proposed system is Pre-processing has been carried out for face images. After the image captured by the camera the lips are detected. During our lip identification process, we explored two distinct methods. The first approach, based on OpenCV, exhibited commendable accuracy when applied to front-facing images. However, its performance suffered notably when dealing with side views.

3.Feature Extraction - In feature extraction, we first extracted basic features of the lip to detect its view. If the lip view is not frontal, we apply normalization techniques for further calculation. Next, speaking lips undergo further processing for lip feature extraction. During this process, the extracted lip features determine whether the lip is open, closed, rounded, or showing labial dental features. Additionally, processing the inner contour of the lip helps identify if lips and teeth are visible, and provides measurements for the height and width of both the inner and outer lip contours.

4.Classification Techniques – The next step in the proposed model is to applied classification techniques there are different classification techniques available based on the extracted features of the lip images lip frames are classified into speaking lips or non-speaking lips. Classification has been done using the technique Support Vector Machine and Deep learning.

4. DATABASE:

First part consists of creating a dataset in Gujarati and Marathi, two Indian languages. In this we have recorded videos of 30 speakers which include male and female with different ages. Each speaker has spoken 10 words. Each word is repeated 5 times. All the words are spoken in Gujarati and Marathi language. Therefore, currently we have total words recorded in a database are 1500 words, i.e., 30*10*5. The format of the videos recorded is mp4 and all these videos are recorded by Sony FE 90mm F2.8 Macro G OSS (SEL90M28G) E-Mount Full-Frame, Mid-telephoto Macro Lens SEL90M28G, using this camera. The distance of the video recording is 4 feet with wall at background.

Table2: Sample 10 words in Gujarati and Marathi Language.

Sr.No.	Gujarati words	Marathi Words
1	ઘર	घर
2	આકાશ	आकाश
3	ખેતર	शेत
4	ફૂલ	फुल
5	સફેદ	पांढरा
6	કાડો	काळा
7	માતા	आई
8	રાત	रात्र
9	સવાર	सकाळ
10	મન	मन

Figure2: Samples frame extraction

Figure3: Sample cropped frames for Ghar word for female

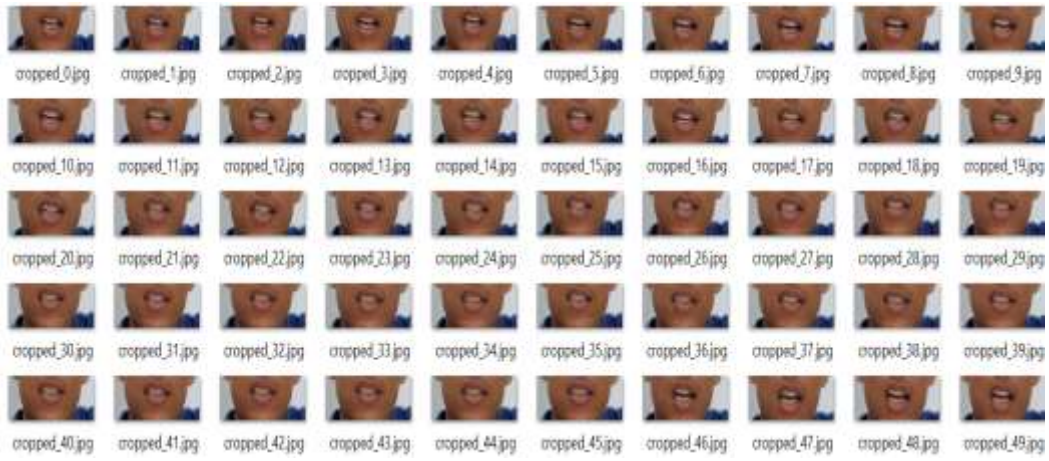


Figure4: Sample cropped frames for Ghar word for male

Figure5: Sample Final output for Ghar word for female

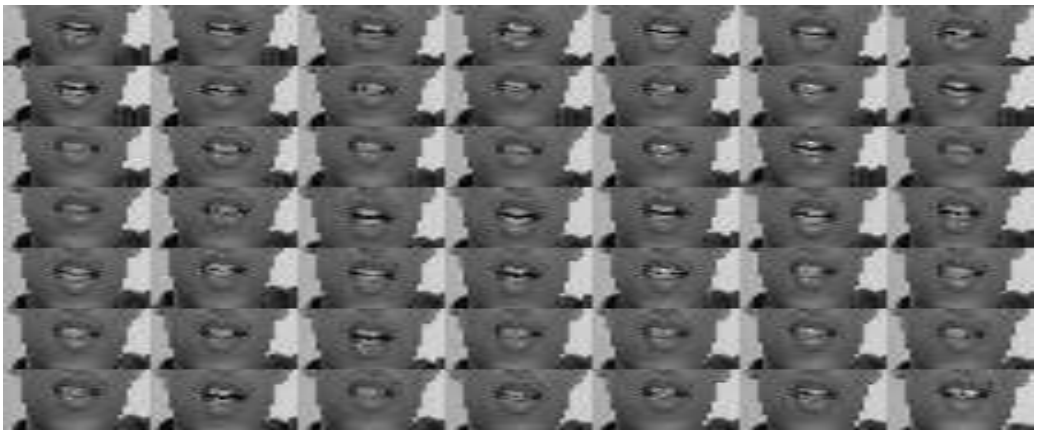


Figure6: Sample Sample Final output for Ghar word for male

5. RESULT

The goal of pattern recognition is to categorize objects of interest into various classes. These objects, referred to as patterns, are sequences of feature vectors extracted from input images in our scenario. For speech recognition, models are independently trained on features extracted from face frames and lip frames. Specifically, a Cubic Support Vector Machine (SVM) model is trained for speech recognition. SVM Classification Support Vector Machine a supervised learning technique is used for classification. In SVM, each feature is transformed as a point in n-dimensional space. Here n is the number of feature vectors used and feature value is used as value of a particular coordinate. Classification using SVM involves separating data into training and testing sets. Each instance in the training set contains one target value and several features. SVM's aim is to develop a model (based on the training data) which predicts target values of the

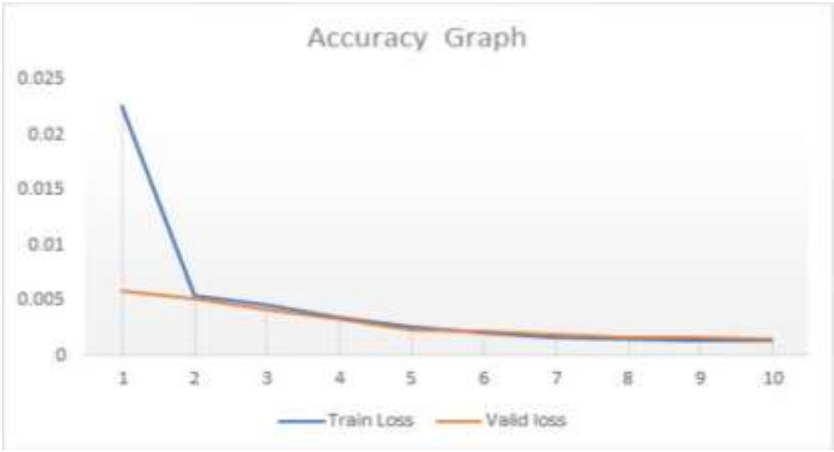
test data given only attributes of the test data [25]. SVM is structured to differentiate from a training set Images divided into two groups (x1 , y1), (x2,Y2), ... (xn, yn) where d-dimensional xi in Rd Function space, and yi in class label {-1,+1},In i=1.. n. SVM establishes maximal separation Hyper planes with kernel (K) function. All images, the vector of which lies on one side In the hyper plane, class -1 and class -1 belong Others are from Class +1 [25]. Output of the classifier is evaluated based on the following formulas, Sensitivity: Measure of correct predictions of presence of abnormality in the image out of total number of images with abnormality. It is also called as True Positive Rate. $Sensitivity=TP$ / $(TP+FN)$ X 1 00 -------- (1) Specificity: Measure of correct predictions of absence of abnormality in the image out of total number of images without abnormality. It is also called as True Negative Rate. $Specificity=TN/(FP+TN)$ X 100 ------- (2) Accuracy: Measure of correct predictions of presence or absence of the abnormality in the image out of total number of images. $Accuracy=TP+TN/$ $(TP+FN+FP+TN)$ X 100 ---- (3) Classification for Speech Recognition To assess the accuracy of speech recognition, we utilized the entirety of the samples for training. Employing a Cubic Support Vector Machine model with 5-fold cross-validation, we trained a model. The average accuracy achieved by the Cubic Support Vector Machine on the dataset was 90 %.

In the initial phase of our Research, we began by capturing frames from a video source. Subsequently, we aimed to detect and isolate faces within these frames while also identifying the specific regions corresponding to the lips. During our lip identi-fication process, we explored two distinct methods. The first approach, based on OpenCV, exhibited commendable accuracy when applied to front-facing images. However, its performance suffered notably when dealing with side views. In order to mitigate these limitations,

Table5: Results.

Epoch	Train loss	Valid loss
1	0.0226	0.0059
2	0.0054	0.0052
3	0.0046	0.0042
4	0.0035	0.0033
5	0.0026	0.0023
6	0.0020	0.0022
7	0.0017	0.0019
8	0.0015	0.0017
9	0.0014	0.0016
10	0.0013	0.0015

we adopted a deep learning model. We ran this model for a total of 10 epochs, and with each passing epoch, we observed a pro-gressive improvement in accuracy. The table below presents the accuracy achieved for each epoch, highlighting the model's steady enhancement over time.



Graph1: Accuracy

6. CONCLUSION

In this research we mainly focused on methods for visual speech recognition has been developed. First part consists of creating a dataset in Gujarati and Marathi, two Indian languages. We have expanded our video data by recording 30 people from different age groups and gender. Each person was supposed to speak 10 words, 5 times each. Therefore, currently we have dataset of size 1500 words, i.e., 30*10*5. The format of the videos recorded is mp4 and all these videos are recorded by Sony FE 90mm F2.8 Macro G OSS (SEL90M28G) E-Mount Full-Frame, Mid-telephoto Macro Lens SEL90M28G, using this camera. We present a new approach to multimodal speech recognition that focuses on the visual cues provided by the lips. The lips are a rich source of in-formation about speech, and they can be used to track the movements of the mouth and to identify the phonemes being spoken. We proposed a new method for tracking the lip boundary and lip area, and we evaluated the performance of our approach on a standard dataset of lip videos. Our results show that our approach to lip tracking is effective, and that multimod-al-al speech recognition can achieve significant improvements over traditional audio-only speech recognition. The experimental results show that using SVM classifier accuracy is 90% for classifying speech.

Acknowledgements

We would like to acknowledge and thanks to Dept of Computer Science and Engineering, Nirma University, Ahmedabad, Gujarat. supporting to this work.

REFERENCES

1. *Zdenek Krnoul, &et.al. The Automatic Segmentation of the Visual Speech Department of cybernetics, University of West Bohemia Univerzitnı 8, 306 14 Plzen, Czech Republic.*
2. *Gordan, &et.al. (2002, July). Visual speech recognition using support vector machines. In 2002 14th International Conference on Digital Signal Processing Proceedings. DSP 2002 (Cat. No. 02TH8628) (Vol. 2, pp. 1093-1096). IEEE.*
3. *Trojan ova, &et.al. (2007, January). Development and testing of new combined visual speech parameterization. In Proc. Int. Conf. Auditory-Vis. Speech Process.(AVSP) (pp. 1-4).*
4. *Abhay Bagai, &et.al. (April 2009) Lip Reading using Neural Networks IJCSNS International Journal of Computer Science and Network Security VOL.9 No.4*
5. *Virginia Estellers, &et.al.(2012).Multi-pose lip-reading and audio-visual speech recognition Estellers and Thiran EURASIP Journal on Advances in Signal Processing*
6. *Bregler, &et.al. (1995, June). Nonlinear manifold learning for visual speech recognition. In Proceedings of IEEE International Conference on Computer Vision (pp. 494-499). IEEE.*
7. *Luettin, &et.al. (1997). Speech reading using probabilistic models. Computer vision and image understanding, 65(2), 163-178.*
8. *Wysoski, &et.al. (2010). Evolving spiking neural networks for audiovisual information processing. Neural Networks, 23(7), 819-835.*
9. *Kumar, &et.al. (2007, April). Profile view lip reading. In 2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07 (Vol. 4, pp. IV-429). IEEE.*
10. *Michelsanti, &et.al. (2021). an overview of deep-learning-based audio-visual speech enhancement and separation. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29, 1368-1396.*
11. *Agelfors, &et.al. (1999). Synthetic visual speech driven from auditory speech. In AVSP'99-International Conference on Auditory-Visual Speech Processing.*
12. *B. Lee, M. Hasegawa-Johnson, C. Goudeseune, S.Kamdar, S. Borys,M. Liu, T. Huang, AVICAR: audiovisual speech corpus in a car environment, Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH), 2004, pp. 380–383.*
13. *I. Matthews, T. Cootes, J. Bangham, S. Cox, R. Harvey, Extraction of visual features for lip reading, IEEE Trans. Pattern Anal. Mach. Intell. 24 (2) (2002) 198–213.*

14. *S. Cox, R. Harvey, Y. Lan, J. Newman, B. Theobald,The challenge of multi speaker lip-reading, Proc. Int.Conf. Auditory-Visual Speech Process. (AVSP), 2008,pp. 179–184.*
15. *T. Hazen, K. Saenko, C. La, J. Glass, A segment-based audio-visual speech recognizer: data collection, development, and initial experiments, Proc. Int. Conf.Multimodal, Interfaces, 2004, pp. 235–242.*
16. *E. Patterson, S. Gurbuz, Z. Tufekci, J. Gowdy, CUAVE:a new audio-visual database for multimodal humancomputer interface research, Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), vol. 2, 2002, pp. 2017–2020.*
17. *P. Lucey, G. Potaminanos, S. Sridharan, Patch-based analysis of visual speech from multiple views, Proc. Int.Conf. Auditory–Visual Speech Process. (AVSP), 2008,pp. 69–74.*
18. *http://www.ee.surrey.ac.uk/Projects/LILiR/index.html.*
19. *McCool, Chris, Sebastien Marcel, Abdenour Hadid, Matti Pietikainen, Pavel Matejka, Jan Cernocky, Norman Poh et al. "Bi-modal person recognition on a mobile phone: using mobile phone data." In Multimedia and Expo Workshops (ICMEW), 2012 IEEE International Conference on, pp. 635-640. IEEE, 2012.*
20. *G. Zhao, M. Barnard, M. Pietikäinen, Lipreading with local spatiotemporal descriptors, IEEE Trans.Multimedia 11 (7) (2009) 1254–1265.*
21. *M. Cooke, J. Barker, S. Cunningham, X. Shao, An audio-visual corpus for speech perception and automatic speech recognition, J. Acoust. Soc. Am. 120 (5) (2008) 2421–2424.*
22. *K. Messer, J.Matas, J. Kittler, J. Luettin, G. Maitre, XM2VTSDB: the extended M2VTS database, Proc. Int. Conf. Audio, Video-Based Biometrics Person Authentication (AVBPA), 1999.*
23. *Alin G. Chitu and Leon J.M. Rothkrantz,"Visual Speech Recognition Automatic System for Lip Reading of Dutch," Journal of Information Technologies and Control,2009,3 pp. 2-9.*
24. *Alin G. Chitu and Leon J.M. Rothkrantz,"Building a Data Corpus For Audio-Visual Speech Recognition," In Euromedia 2007.Apr. pp.88–92. Available at:http://mmi.tudelft.nl/pub/alin/APTEC-04.pdf.*
25. *Mihaela Gordan, Constantine Kotropoulos Ioannis Pitas "VISUAL SPEECH RECOGNITION USING SUPPORT VECTOR MACHINES".*

******