ORIGINAL RESEARCH PAPER

Volume-8 | Issue-8 | August - 2019 | PRINT ISSN No. 2277 - 8179

INTERNATIONAL JOURNAL OF SCIENTIFIC RESEARCH

LIP READING THROUGH FOURIER BASED FACIAL LANDMARK

Computer Science

Nandini M S Dept. of IS & Engg., NIE Institute of Technology, Mysuru, Karnataka

Nagappa U. Bhajantri* Dept. of CS &Engg., Government, Engineering College, Chamarajanagara, Karnataka, India *Corresponding Author

Trisiladevi C Nagavi Dept.of CS &Engg., Sri Jayachamaraja College of Engg., JSS Science and Technology University, Mysuru, Karnataka, India

ABSTRACT

An algorithm has been presented to reduce noise distortion and to predict the movement of lips under noisy environment. However, such environment will be a challenging problem for a good hearing person. Further it is very difficult to listen and understand the words spoken by a person under noise distorted conditions. But, hearing impaired person reads the movement of lips by analyzing the shapes of lips provided language is known . The work is proposed to detect, track and recognize the shapes of lips and predict the movement of lips for Kannada language. The lip tracking, and prediction of sentences spoken by a person, as we have to keep track of every changes that are observed in lip movements recognized at a fraction of seconds measured in terms of time. Thus, time is a parameter that plays a vital role in analyzing and understanding the sentences spoken by a person. Further effort attempt to perform the task of estimating the shape of lips and then annotates the shapes of every lip movement with appropriate words. A minute change in movement of lips designates a specific prediction of words spoken. As the movement of tongue cannot be seen in videos properly, every change in shape of a lip designates a specific words of a sentences. Even while recognizing specific words like Amma and Appa are almost similar in nature even though they are different in meaning. Thus, we have came up with a model that analyzes and predicts the language spoken by a person with a good accuracy.

KEYWORDS

Fourier series, Facial Landmark, shape information, words classification, Kannada Words.

I. INTRODUCTION

In the recent era, Natural Language Processing is a subject of Machine learning that has a need for development of intelligent systems like supervised machine learning techniques(SML). The SML used in [12,19,21] works on the principle of training with sufficient data in the form of video. The Natural Languages related to lip reading is a challenging task that uses SML. Moreover, the human beings have a capability of visualizing certain objects and recognizing those objects into different classes based on the information learned during the learning stage. Similarly in case of artificial intelligence, the systems have to be trained with different methods of machine learning that helps system work like human beings. On the other hand, the artificially intelligent system must be trained with suitable algorithms as such the system works independently in situations based on the data trained to the system. Lip reading is similar to human readable format that reads and recognizes different languages based on certain training. If the system is trained to learn English language, the system recognizes it during testing phase. If the system is trained to learn local Language, the system recognizes Lip movements.

Fourier series and Facial Shape Landmark based feature learning plays a vital role in analyzing and understanding the features of frames of videos. Further the phase of training and testing as an important method of machine learning. The principle of mathematical operations required to make the system intelligent in identifying lip movement is very essential and while training the system to recognize Lip Movements. The term Fourier Series indicates series of expressions that may be suitable at different levels of features extracted while training the system at multiple levels, where each of different polynomial expression has a significant contribution towards recognizing the facial shape based Lip Movement. Expression at level 1 consisting of window of 128x128 kernel size, level 2 consists of 64x64 , level 3 consists of 32x32 features for recognizing lip movements. These multiple layers of extraction of features from different shapes of lip movement together constitutes the total number of features of a sequence of related frame in video. Multiple dimensional data or video data is transformed into one (1) dimensional data in the form of features vector, where the data is numerically represented for video data, which is separated from audio data.

The complete architecture is portrayed in the Fig 1. Recognizing of Lip Movement for English language has been carried out in some of the research paper [2,5,6,8,9], but the languages like Kannada and others are not more exploited with lip movement. Especially identification of lip movement for a Kannada sentences is very important and challenging task

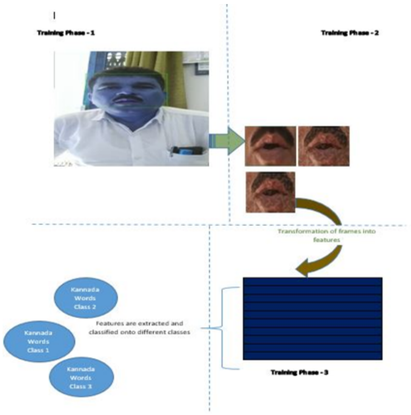

Fig.1. Architecture of the Learning and Top view of Pyramid representing and training for Recognition of Kannada Lip Movements.

Especially, when there is a small change in shape of an object, those changes must be notified and compared with ground truth shapes of objects to recognize different shapes of lips. The ground truth results are obtained from benchmark datasets of English Lip Reading, where the actual data and information is available with it.

Here the effort focuses on different aspects of lip-reading in related work in section II along with datasets and appropriate challenges in lip-reading. Section III presents proposed Fourier series and Facial Landmark based method for prediction of Kannada lip movements. Section IV describes the results and analyzes the performance of erected approach with respect to other contemporary attempts. Section V discusses the advantages with respect to existing methods and section VI concludes with the contributions.

II. RELATED WORK AND DATASET

The dataset consists of frames of videos of different directions of facial poses. Even then the system has been able to detect the facial features of a person and predicts the words spoken by a person by applying certain algorithms.

An active shape model (ASM) is a shape-constrained iterative fitting algorithm [28]. The shape features are considered from an ASM features that is extracted from an object as per [3,6,8] also known as a point distribution model (PDM),

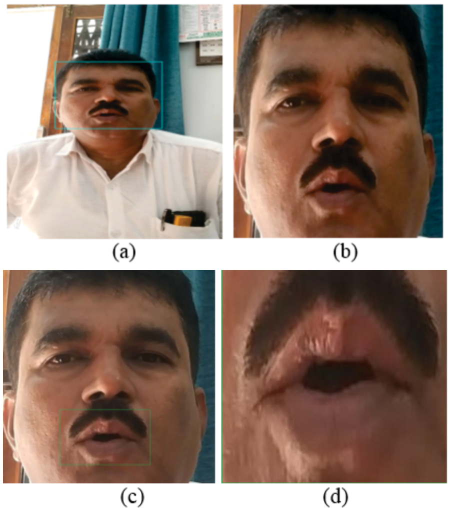
(a) (b)
(c) (d)

Fig. 2. Detected and cropped face is shown in (a) and (b) respectively. Detected mouth is shown in (c) and cropped mouth portion is shown in (d).

However, the mathematical operations like point distribution model is considered as a reference to extract features of lip in every instance of it. Even when a small changes are observed in facial features like lip, the shape features are noticed by annotating the facial information from a sequence of frames of a video as shown in Fig.2

To align the set of training models, the conventional iterative algorithm is used [27,28,29,30]. Given the set of aligned shape models, the mean shape, as it can be calculated along the axes describes most variance about the mean shape can be determined using a principal component analysis (PCA) as per[14,18].

III. FOURIER SERIES AND FACIAL LANDMARK BASED LEARNING FOR RECOGNITION OF LIP MOVEMENTS.

Let us consider a labelled features of training data trained with supervised learning problem that has access to the trained data.

A. Fourier Series based Representation Learning for Recognition of Lip movements

Here, SML has access to the training data of the form (x^i, y^i), where x^i is a feature vectors corresponding to feature labels y^iassigned during training phase. The training phase consists of extraction of features at different dimensions that minimizes the percentage of error that may arise while classifying the sequence of shapes of lip movement into different classes. Landmark based representation learning plays a vital role in solving complex, non-linear data into a form that fits feature vectors.

The proposed Landmark based feature representation learning, introduce the problem of solving the original eq. (1) into two different dimensions as shown in fig.3. The computational dimension that receives original input image with a boundary conditions like +1 and -1 as inputs and outputs the processed data in the form of Eq. (2) such that the input data is divided into different levels to process the information of image one by one and produces an output in the form of Eq. (3). Thus the method is defined by logistic regression, the original input image divided into different dimensions with boundary values like +1 and -1 that produces an output by mapping the function from input to output. These features are fed to the system along with boundary values +1 and -1 to the system called lower bound and upper bound of input interval function that maps the feature inputs to output.

Level 1 used 2- dimensional data two similar dimensions along with +1 and -1 are fed to the next level of feature learning, which receives input from previous level and combines the information of previous level data to the system along with these output are fed to the next layer. These kind of concatenating the data produces an output by one layer with next subsequent layers at different levels together forms that refines the data annotated with shapes into different layers. Further produces a prediction of appropriate shape corresponding to the input given to the system.

$$X_i = x_0 + x_1 \ldots x_{n-1} \qquad (1)$$

$$X_k = \sum_{k=0}^{n-1} x_k * e^{\frac{-2\pi ki}{N}} \qquad (2)$$

Where, $e^{i\theta} = cos\theta + i\ sin\theta$, the above eq. (2) shall be expanded to the form of eq. (3).

$$X_k = \sum_{m=0}^{n-1} x_m * \left[cos\left(\frac{2\pi ki}{m}\right) + i\ sin\left(\frac{2\pi ki}{m}\right)\right] \quad (3)$$

Where, $e^{i\theta} = cos\theta + i\ sin\theta$, the above eq. (2) shall be expanded to the form of eq. (3).

$$X_k = \sum_{m=0}^{n-1} x_m * \left[cos\left(\frac{2\pi ki}{m}\right) + i\ sin\left(\frac{2\pi ki}{m}\right)\right] \quad (3)$$

We know that cos $(-\theta) = cos\theta$ and sin $(-\theta) = -sin\theta$ and Eq. (3) represents the form of dimensionality that is divided into different dimensions. The Fourier Series in eq. (2) receives input from eq. (1) and performs the task of differential equations with respect to x between the intervals negative and positive waveforms of Fourier Series effort[17,22].

$$X_k = \sum_{m=0}^{n-1} x_m * \left[cos\left(\frac{2\pi ki}{m}\right) - i\ sin\left(\frac{2\pi ki}{m}\right)\right] \qquad (4)$$

$$X_k = \sum_{m=0}^{n-1} x_m * \left[cos\left(\frac{2\pi ki}{m}\right) - \sum_{m=0}^{n-1} x_m\ sin\left(\frac{2\pi ki}{m}\right)\right] \qquad (5)$$

$$P_k = \sum_{m=0}^{n-1} x_m * \cos\left(\frac{2\pi ki}{m}\right) \qquad (6)$$

$$Q_k = \sum_{m=0}^{n-1} x_m * \sin\left(\frac{2\pi ki}{m}\right) \qquad (7)$$

Eq. (5) receives input from Eq. (4) and performs certain computations like summating of Fourier series information extracted from an image.

$$\varepsilon(I) = \sum_{r=0}^{n-1} e^{i\theta} \qquad (8)$$

Further, eq. (6) and eq. (7) determines the Fourier information of features from different dimensions. Eq. (8) does the task of calculating the energy of the detected segment of the lip portion. Where the lip portion of reduced dimension is used as a part of processing an image and concatenating the result of processing into an appropriate results like feature vectors.

The output of these layers are fed to next subsequent layers together constitutes a very important information that concatenates multiple levels of Fourier information into one single system. This process of concatenating of data from one level of information into another level are as feed forward movement of data for processing the inputs given to the system.

In order to train system with multiple outputs, we need to mention the system with different classes of output. The output plays a very important role in predicting multiple classes of data. Some of the statistical features are attempts to exercise in proposed method, such as Mean, Standard deviation, Energy, Correlation, Entropy, Mean Standard deviation, Variance, Covariance, Difference Moments, and Correlation-2. These features together with facial feature localization points and its variations with respect to time in different frames are used to recognize the facial lip movements into different classes of Kannada words as narrated in the Fig. 1. These are used in every frames, where exactly the changes takes place in values of feature vectors according to the eq. (12), eq. (13), eq. (14), eq.(15) and eq.(16). The equations together with different channels of information is used to accurately recognize the facial lip movements.

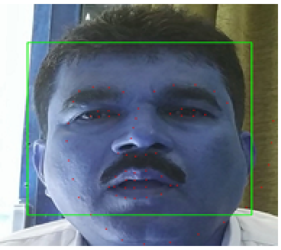

Fig.3. The lip localization from facial landmark.

Each level of architecture has different layers of operations for features extraction in addition to detection of lips using some of the algorithms like viola jones, which does the task of recognizing the facial features like lips for frontal faces, but the method is showing significant contribution towards recognizing shapes of lips for tilted faces like faces turned towards left and right are quite a challenging tasks of machine learning.

In the pretext of accumulating the features are as follows,
1. Grouped Consonants: The grouped consonants 25 are used in Kannada language very often. There are a total of 25 consonants x 10 statistical features x 3 different channels x 15 features spaces together constitutes to a total of 11,250 dimensional data + 250 dimensional data of a video along different frames

2. Miscellaneous Consonants: There are total of 10 miscellaneous consonants x 10 statistical features x along 3 different channels (R-G-B) x 15 feature vector data x 2 dimensional data reduction + 250 video data x along five folds 5 together constitutes 10,500 dimensional data.

Thus, finally we have obtained a total number of features dimensions of 11500 + 10500 features dimensions together forms a total of 22,000 features dimension of different frames of a video are extracted while the system is being trained with SML.

B. Proposed Algorithm

Algorithm: lip movement recognition with Fourier series

Algorithm: Lip Reading for Sentences		
Input: Audio with video data is input to the system		
Description: System determines the lip movement		
Output: Reduction of Noise Distortion and Lip movements are recognized for sentences		
Begin	**[Pre-processing]**	
Step 1:	Enforce eq. (1).	
Step 2:	**[Concatenation of Fourier Series with Facial Landmark Localization]**	
	Step 2.1	Compute with eq.(2)
	Step 2.2	Calculate with eq.(3)
	Step 2.3	Concatenate eq. (3) and eq.(4)
	Step 2.4	Compute with eq. (4) and assign it to eq.(5)
Step 3:	**[Fourier Series Feature Learning]**	
	Step 3.1	Assign labels to shapes of individual frames as per eq.(6)
	Step 3.2	Compute with eq. (7)
Step 4:	**[Recognitions]**	
	Step 4.1	Evaluate with Compute eq. (8)
	Step 4.2	If not end of frame Go to step 4.1
	Step.4.3	Recognize the Shapes
End		

concatenation with Facial Landmark based feature extraction at different dimensions for lip reading

The effort [1,4,7,10] has focused on reading lips by separating audio from visual data, then reduction of noise distortion, as we too focused towards annotating, recognizing as per [11,13,15] and understanding sentences, further need to perform certain tasks like preprocessing, psychophysical properties are analyzed to reduce the noise distortion, then pre-processing, annotations, features extraction, features classification are carried out to recognize sentences.

$$Precision = \frac{TP}{TP + FP} \quad (9)$$

$$Recall = \frac{TP}{TP + FN} \quad (10)$$

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (11)$$

$$\text{Mean}(\text{m}) = \frac{1}{N^2} \sum_{i,j=1}^{\infty} P(i,j) \quad (12)$$

$$\text{Standard deviation (sd)} = \frac{1}{\text{N}} \sqrt{\sum_{i,j=1}^{N} [p(i,j) - m]^2} \quad (13)$$

$$Entropy = \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} -P_{ij} * \log P_{ij} \quad (14)$$

$$Correlation\ 1$$

$$= \sum_{i=0}^{n-1} \sum_{j=0}^{n-1} \frac{(i,j)P_{ij} - \mu_x \mu_y}{\sigma_x \sigma_y} \quad (15)$$

$$Energy = \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} P_{ij}^2 \quad (16)$$

$$Correlation\ 2$$

$$= \sqrt{1 - exp\left[-2\left((i,j)P_{ij} - \mu_x \mu_y\right)\right]} \quad (17)$$

Subsequently, the parameters like precision and recall plays a significant role in assessing the accuracy of recognizing the movement of lips especially for Kannada language. There are various methods in deep learning, SML, Semi-SML algorithms In other words, measuring the efficacy of extended method with reference to the others.

IV. RESULTS

The effectiveness of the effort using Landmark based feature extraction for recognition of sentences shall be observed in terms of accuracy of existing methods where the exercised strategy has yielded with an accuracy 87.15% ,as narrated in the table 1.

Table 1. Analysis of erected method for Lip reading.

Sl.No	Methods	Accuracy
1	J.A. Bangham *et al.* [16]	68.46%
2	C. Bregler *et al.*[20]	52.50%
3	T.F. Cootes *et al.* 2017 [23]	84.50%
4	**Proposed Fourier Series Lip reading**	**87.15%**

Table 2. Analysis of method compared with different similarity indexes

Methods	Jaccard Similarity	Dice Similarity
J.A. Bangham *et al.* [16]	65.3	69.4
C. Bregler *et al.*[20]	71.2	69.15
T.F. Cootes *et al.* 2017 [23]	70.50	71.21
Fatemeh vakhshiteh et.al [24]	71.22	74.7
Towards Practical Lip Reading [25]	72.15	73.41
Gregorry J Wolff et.al [26]	69.51	68.53
B. Atal et.al [27]	70.33	79.83
J. Alegria et.al [28]	80.41	81.32
Eric petajan et.al [30]	84.56	88.92
Proposed Fourier Series and Facial Landmark Localization based Lip reading	**91.36**	**91.44**

In addition to some of the state of the art techniques like lip reading in the wild and other methods are enlisted in the table 2. The average precision is clearly demonstrating more accurate than contemporary method, the comparison observed in terms of graphical representation. Further, the efforts are more prone to detecting the shape of a mouth in different situations like tilted faces towards left or right. In addition to frontal faces of a person. It was a challenging task for developing an algorithm that does the task of recognizing words or sentences. The system determines the shape of mouth and does many tasks between annotations and prediction of sentences. Here, effort portrayed many advantages over existing work in terms of posterior time, accuracy, precision of accurately predicting sentences spoken by a person during test phase.

Table 3. Analysis of Noise Distortion and compared with existing methods

Sl.No	Folds	Fourier Method (dB)	Accuracy (%)
1	Fold 1	73.25	83.15
2	Fold 2	74.42	84.42
3	Fold 3	78.21	88.41
4	Fold 4	79.33	89.31
5	Fold 5	81.36	91.32

A. Performance Evaluation

Further, the parameters considered for evaluation of performance of a trained systems are precision, recall, sensitivity, specificity, accuracy, as we are measuring the accuracy of lip reading system in terms of recognizing the sentences that is spoken by a person. On the other hand the erected approach throws light on measuring time complexity with reference to existing efforts, as we are more concerned towards recognizing the shapes of lips and its movements for recognition of sentences. On other hand, which is helpful to differently abled personalities in addition to normal persons.

V. DISCUSSION

The experimentation has commenced with challenges to analyze and train the system to recognize shapes of lips by facial landmark localization for sequences. The reading of lip task for predicting certain characters like ಕ ಕಾ ಕಿ ಕೀ ಕು ಕೂ ಕೃ ಕೆ ಕೇ ಕೈ ಕೊ ಕೋ ಕೌ ಕಂ ಕಃ and ಗ ಗಾ ಗಿ ಗೀ ಗುas we are more focused towards recognizing lip movements and prediction of sentences. Even though the words seems to be spelt similar in manner, meaning are seem to be different, therefore some expressions were very difficult to predict. Similarly other challenging tasks addressed through some facial lip expressions are supposed to be identified with inherent shape changes. These are tend to be different and helps to measure the lip shapes by facial landmark localization as displayed in Fig. 3.

In view of experiment simplification of quantum of video frames. The entire video has been divided into equal partitions. In other words, video is broken down into folds, which are able to accommodate the fixed number of frames. Here, we have formed five such folds to measure accuracy and capability of erected strategy. The same is portrayed in table 3. Pertaining to noise distortion removal. Further, table 2 is revealing the performance with existing approaches. Here, to determine the way in which the two spoken sentences are similar. It can measure the commonality between the sentences. In precise, extend the static

Further, it will obtain the ratio with the number of words that are in the sentence sets, and this will give the similarity index referred to as Jaccard similarity index. However, Jaccard index can be extended on strings, where each string contains the characters. Generally, it does not considers the order of the characters, merely considers the presence of the characters. However, if we consider the index as 0, the two vectors have no elements in common. On contrary, similarity is 1, if two vectors have similar characters in union. Similarly, Dice similarity is employable primarily for empirical rather than theoretical. Its distance retain sensitivity in more heterogenous sentences spoken and gives less weight to outliers.

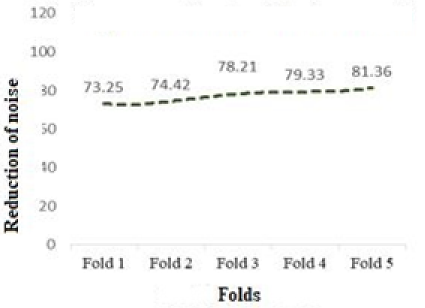

Fig. 4. Improvement in noise distortion removal

The dice coefficient and Jaccard index are monotonically related. So, we have extended both to evaluate the classifier performance. Since, they are completely monotonic in one another. However, Jaccard might be a little unintuitive. Because it is always less than or equal to minimum of precision and recall ,as shown in table 2 and eq. (18) and eq. (19). The research work has erected its good results over a dataset consisting of 25 sample videos. In order to obtain the results of identification of lip movements for 25 videos, the effort has been made to outperform existing methods in terms of accuracy.

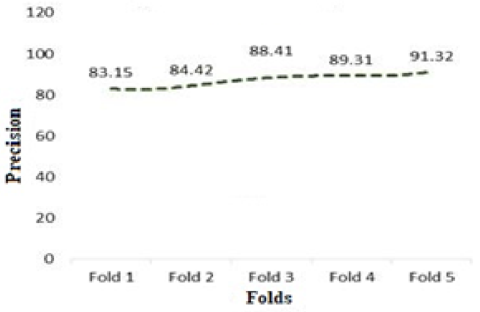

Fig.5. Comparison of precision along 5 folds.

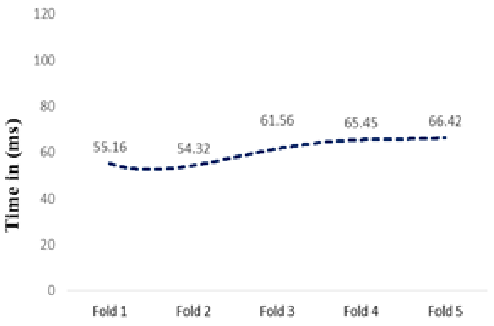

Fig.6. Measure of posterior time complexity for prediction of Sentences.

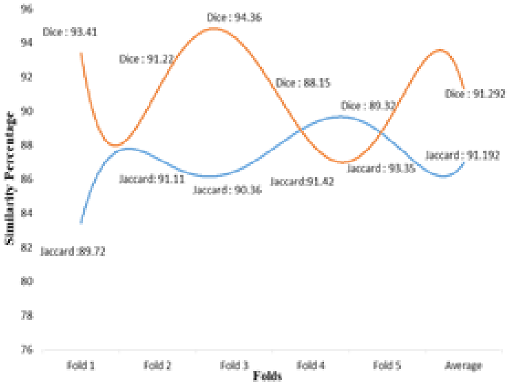

Fig.7. Measure of accuracy in terms of jaccard and dice similarity indexes for prediction of sentences that may be used as one of the important parameter.

Dice Similarity as

$$Dice = \frac{2*(P_k-Q_k)}{2*(P_k-Q_k) + 2*(P_k+Q_k)} \qquad (18)$$

Jaccard Similarity as

$$\backslash = \frac{\text{ь}^{\text{K}}+\text{ϭ}^{\text{K}}}{\text{ь}^{\text{K}}-\text{ϭ}^{\text{K}}} \qquad (\text{Iə})$$

tracking and lip movement is identified with a metric of accuracy is shown in fig.7.

However, the empirical experimentation revealing for the proposed Fourier feature based effort on certain metrics such as noise distortion removal ,precision, accuracy and posterior timing aspects have magnificently improve with respect to folds of video. Further, Jaccard and Dice similarity have shown similar evidences with respect to folds. Moreover, fold1 witnessing minimum accuracy, precision and time elapsed. On the other hand maximum metric value explored for fold5.Further,fold3 highlights maximum Dice coefficient similarity and minimum for fold5.Similarly,Jaccard index bears minimum and maximum for fold 1 and fold 5 respectively.

VI. CONCLUSION

The proposed method of recognition of words from lip movement has achieved few significant contributions such as detecting the facial features like lip in tilted faces of a person. These tilted faces of a person together with frontal faces forms a contribution for detecting the lip portion from every frames to be processed for feature extraction to recognize the language spoken by a person. The effort is towards recognizing lip movements. The desired objective has achieved at an accuracy of 87.15% of recognition of Kannada Sentences. The drawback of the algorithm is,it needs to perform the shape model operation on every change in shape of the lips. In other words, it may be involved in more computational aspects with cumbersome process.

REFERENCES

[1]. Campbell R (1998) Speech reading: advances in understanding its cortical bases and implications for deafness and speech rehabilitation. Scand Audio Suppl 49.

[2]. Bernstein LE, Demorest ME, Tucker PE (2000) Speech perception without hearing. Perception and Psychophysics. Perception and Psychophysics 62: 233–252.

[3]. Grant KW, Walden BE (1996) Evaluating the articulation index for auditory visual consonant recognition. J Acoust Soc Am 100: 2415–2424.

[4]. MacLeod A, Summerfield Q (1987) Quantifying the contribution of vision to speech perception in noise. Br J Audio l21: 131–141.

[5]. Massaro DW (1987) Speech perception by ear and eye: A paradigm for psychological inquiry. Hillsdale, NJ: Erlbaum.

[6]. Bernstein LE, Auer ET, Moore JK (2004) Audiovisual speech binding: convergence or association? In: Calvert GA, Spence C, Stein BE, eds. The handbook of multisensory processes. Cambridge, MA: MIT Press. pp. 203– 223.

[7]. Grant KW, Walden BE, Seitz PF (1998) Auditory-visual speech recognition by hearing-impaired subject: Consonant recognition, and auditory-visual integration. J Acoust Soc Am 103: 2677–2690.

[8]. Sumby WH, Pollack I (1954) Visual contribution to speech intelligibility in noise. J Acoust Soc Am 26: 212–215.

[9]. Erber NP (1969) Interaction of audition and vision in the recognition of oral speech stimuli. J Speech Hearing Res 12: 423–425.

[10]. Erber NP (1971) Auditory and audiovisual reception of words in low-frequency noise by children with normal hearing and by children with impaired hearing. J Speech Hearing Res 143: 496–512.

[11]. Erber NP (1975) Auditory-visual perception in speech. J Speech and Hearing Disord 40: 481–492.

[12]. Binnie CA, Montgomery A, Jackson PL (1974) Auditory and visual contributions to the perception of consonants. J Speech and Hearing Res 17: 619–630.

[13]. McCormick B (1979) Audio-visual discrimination of speech. Clin Otolaryngol Allied Sci 45: 355–361.

[14]. Meredith MA, Stein BE (1986) Spatial factors determine the activity of multisensory neurons in cat superior colliculus. Cogn Brain Res 369: 350–354.

[15]. Ross LA, Saint-Amour D, Leavitt VN, Javitt DC, Foxe JJ (2007) Do you see what i am saying? Exploring visual enhancement of speech comprehension in noisy environments. Cereb Cortex 17: 1147–1153.

[16] J.A. Bangham, P. Ling, and R. Young, "Mulitscale Recursive Medians, Scale-Space, and Transforms with Applications to Image Processing," IEEE Trans. Image Processing, vol. 5, no. 6, pp. 1043-1048, 1996.

[17]. Ernst MO, Banks MS (2002) Humans integrate visual and haptic information in a statistically optimal fashion. Nature 415: 429–433.

[18]. van Beers RJ, Sittig AC, Gon JJ (1999) Integration of proprioceptive and visual position-information: An experimentally supported model. J Neurophysiol 81: 1355–1364.

[19]. Kucera H, Francis WN (1967) Computational analysis of present-day American English. Providence, RI: Brown University Press.

[20] C. Bregler and S.M. Omohundro, "Learning Visual Models for Lipreading,"Computational Imaging and Vision, chapter 13, vol. 9, pp. 301-320, 1997.

[21]. Lidestam B, Lyxell B, Lundeberg M (2001) Speech-reading of synthetic and natural faces: effects of contextual cueing and mode of presentation. Scand Audiol 30: 89–94.

[22] Kai Xu, Dawei Li, Nick Cassimatis, Xiaolong Wang, "LCANet: End-to-End Lipreading with Cascaded Attention-CTC", IEEE Computer Vision, 2018.

[23] T.F. Cootes, G.J. Edwards, and C.J. Taylor, "Active Appearance Models," Proc. European Conf. Computer Vision, pp. 484-498, June 1998.

[24] Fatemeh vakhshiteh, farshad almasganj, ahmad nickabadi "Lip-Reading Via Deep Neural Networks Using Hybrid Visual Features," Image Anal Stereol, 159-171,36, 2018.

[25] Ziheng Zhou, Mattie Paitekainen, Guoying Zhao, "Towards Practical Lip Reading," IEEE CVPR, June 2011.

[26] Gregorry J Wolff, K Venkatesh Prasad "Lipreading by neural networks: Visual preprocessing, learning and sensory integration," Journal , vol. 28, pp. 1028-1034, 1980.

[27] B. Atal and L. Hanauer, "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave," J. Acoustical Soc. of America, vol. 50, pp. 637-655, 1971.

[28] J Alegria, J Lechat , "Phonological Processing in Deaf Children: When Lipreading and Cues Are Incongruent," Journal of Deaf Studies and Deaf Education vol. 10 no. 2, 2005.

[29] J.A. Bangham, R. Harvey, P. Ling, and R.V. Aldridge, "Morphological Scale-Space Preserving Transforms in Many Dimensions," J. Electronic Imaging, vol. 5, no. 3, pp. 283-299, July 1996.

[30] Eric petajan, Hans Peter Graf, "AUTOMATIC LIPREADING RESEARCH: HISTORIC OVERVIEW AND CURRENT WORK," Multimedia Communications and Video Coding, pp. 265-275, 1996.