International Journal of Applied Engineering Research ISSN 0973-4562 Volume 14, Number 13 (2019) pp. 3111-3118

Recognition of Kannada Lip Movements through Wavelets

Nandini M S
Department of IS & Engineering
NIE Institute of Technology,
Mysuru, Karnataka, India.

Nagappa U.Bhajantri
Department of CS & Engineering
Government Engineering College
Chamarajanaara, Karnataka, India.

Trisiladevi C Nagavi
Department of CS & Engineering
Sri Jayachamaraja College of Engg.,
JSS S&T U, Mysuru, Karnataka, India.

Abstract

The work proposed to detect, track and recognize the shapes of lips and predict the movement of lips. As we have to keep track of every changes that are observed in lip movements are to be 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 sentences. But a hearing impaired person reads the movement of lips by analyzing the shapes of lip, provided the language for known. The effort performs the task of wavelet feature based estimating the shape of lips and then annotates the shapes of every lip movement with appropriate words. As the movement of tongue cannot be seen in videos properly, every change in shape of a lip designates a specific words. Thus, we have designed a model that analyzes and predicts the language spoken. The method has yielded good accuracy. For hearing impaired person, language spoken by a person shall be understood by recognizing the shape of a lip.

Keywords: wavelet feature, annotation, shape information, words classification.

I. INTRODUCTION

In the latest decode Natural Language Processing is a subject of Machine learning that has a need for development of intelligent systems like supervised machine learning techniques. The supervised machine learning [2,4,5,6]works on the principle of training a system with sufficient data in the form of video. The Natural Languages related to lip reading is a challenging task that uses supervised machine learning techniques. 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. The artificially intelligent system [7,9,11] 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 done to it. 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 of such language.

Wavelet based feature representation learning is an important method of machine learning that trains the system by analyzing the principle of mathematical operations that is required to make the system intelligent in identifying lip movements of language. The term wavelets indicates different levels of features extracted while training the system at multiple levels, where each of different layers like level 1 consisting of window of 128x128 size, level 2 consists of 64x64 , level 3 consists of 32x32 size features for recognizing lip movements. These multiple levels of extraction of features from different shapes of lip movement together constitutes the total number of features of a frame in video. Multiple dimensional data or videos data are 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.

Recognition of Lip Movement for English language has been carried out in many of the research articles, but the local 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

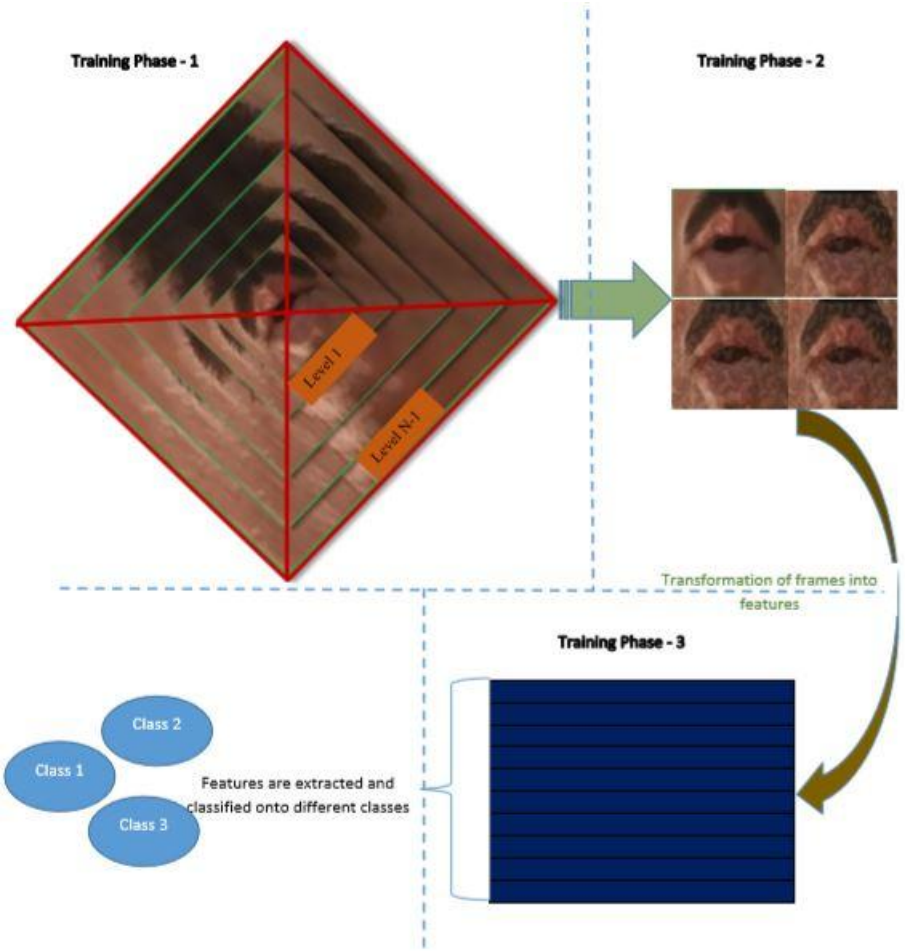


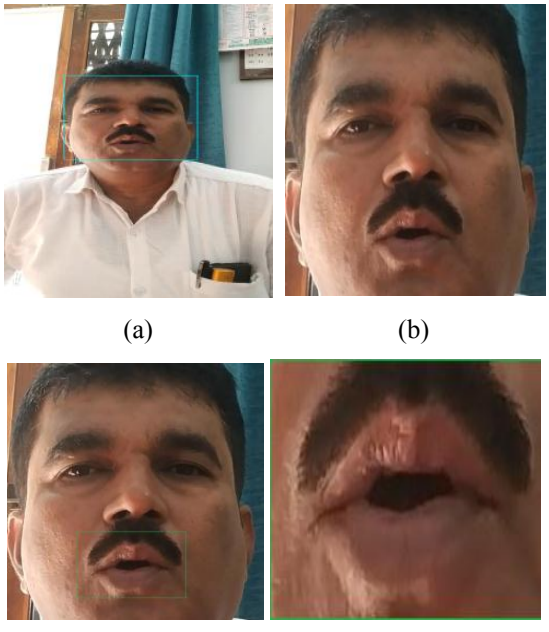
Fig.1. Architecture of the proposed Wavelet based Learning and Top view of Pyramid representing and training for Recognition of Kannada Lip Movements

The research work focuses on different aspects of lip-reading in related work in section 2, section 3 focuses on datasets and its challenges in lip-reading, section 4 presents our proposed wavelet based information for prediction of Kannada lip movements, section 5 describes the results and analyzes the performance of method with respect to other contemporary methods, section 6 discusses the advantages with respect to existing methods, section 7 concludes the research article .

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 through proposed algorithm.

Here, effort from an active shape model (ASM) is elaborated through a shape-constrained iterative fitting algorithm [19,21,28]. The shape features are considered from an ASM features that is extracted from an object as per [1,3,8,10,13] 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 face from frames are further processed to crop the mouth portion.

Statistical data computation have extended to analyze the extracted features. In this method the mathematical operations like point distribution model [12,17,18] is considered as a reference model as per [14, 15,16] 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 [20, 23, 25] the facial information from a sequence of frames of a video.

To align the set of training models, the conventional iterative algorithm is used [22, 24,26,27]. Given the set of aligned shape models, the mean shape, as it can be calculated along the axes that describes most variant data about the mean shape that can be determined using a principal component analysis (PCA).

III. PROPOSED WAVELET BASED REPRESENTATION LEARNING FOR RECOGNITION OF KANNADA LIP MOVEMENTS

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

A. Wavelet based Reperesentation Learning for Recognition of Lip movements in Kannada

Supervised Machine Learning 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^i assigned during training phase. The Wavelet based Representation learning plays a vital role in solving complex, non-linear data into a form that fits our feature vectors.

In order to describe the proposed wavelet based feature representation learning, we first introduce the problem of solving the original eq. (1) into two different dimensions as per fig.3. The dimension mentioned in fig.3 is a 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 are used to process the information of image one by one and produces an output in the form of eq. (3). The same model presented in fig.3 is extended to multiple levels of data input in the form of fig.4, where the input data represented by level 1, level 2, level 3 so on level N-1. 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.

In Level 1 we used 2 dimensional data two similar dimensions along with +1 and -1. These kind of concatenating the data produces an output by one level information with next subsequent levels together forms our proposed method and refines the data annotated with shapes into different levels and produces a data that helps in prediction of appropriate shape corresponding to the input given to the system. The above fig.2 and fig.3 indicates the number of levels and number of units used to represent an input into the system and to process the data input in levels with 4 similar dimensions and finally the output is produced with one (1) unit.

$$\psi_{s,\tau}(x) = \frac{1}{\sqrt{s}}\,\psi\left(\frac{x-\tau}{s}\right) \quad (1)$$

Eq. (2) represents the form of dimensionality that is divided into different dimensions. The wavelets W in eq. (2) receives input from eq. (1) and performs the task of differential equations with respect to x between the intervals -1 and +1.

$$W_{\varphi}(s,\tau) = \int_{-1}^{1} f(x) * \psi_{s,\tau}(x)\,dx \quad (2)$$

$$f(x) = \frac{1}{C_{\psi}} \int_{0}^{1}\int_{-1}^{1} W_{\psi}(s,\tau) * \frac{\psi_{s,\tau}(x)}{s^2}\,d\tau\,ds \quad (3)$$

Where, $C_{\psi} = \int_{-1}^{1}\left|\frac{\psi(\mu)}{\mu}\right| d\mu$ and ψ_{μ} is the Fourier transform of φ_{μ}. The eq. (3) indicates the function of different dimensional data that is fed to the system. The eq. (3) further divides the original image into different dimensions like LL, HH, LH, HL in two dimensional image, when it is split it more smaller dimensions.

$$\psi_{i,j}(x) = 2^{i/2}\,\psi(2^i x - j) \quad (4)$$

$$f(n) = \frac{1}{\sqrt{P}}\sum_{j} W_{\varphi}(i_0, j) * \varphi_{i_0,p}(n) + \frac{1}{\sqrt{P}}\sum_{i=i_0}^{1}\sum_{p} W_{\psi}(i,j) * \psi_{j,p}(n) \quad (5)$$

Eq. (5) receives input from output generated by eq. (4) and performs certain computations like summating of wavelets information extracted from an image.

$$W_{\phi}(i_0, j) = \frac{1}{\sqrt{P}}\sum_{x} f(x) * \phi_{i_0,j}(x) \quad (6)$$

$$W_{\psi}(i,j) = \frac{1}{\sqrt{P}}\sum_{x} f(x) * \psi_{i,j}(x) \quad (7)$$

$$\varepsilon(I) = \sum_{r=0}^{n-1}\sum_{c=0}^{n-1} I(r,c)^2 \quad (8)$$

Further, eq. (6) and eq. (7) determines the wavelet 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 a wavelet information into an appropriate results like feature vectors. The output of these levels are fed to next subsequent levels together constitutes a very important information that concatenates multiple levels of data into one single system. This process of concatenating the data from one level to another level of data for processing the inputs given to the system.

In order to train our system with multiple outputs, we need to mention the system with different classes of output. The output layer plays a very important role in predicting multiple types of data as output data.

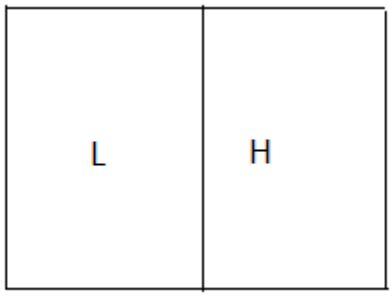


Fig.3. Splitting of original image into two dimensions

It's very much clear from the above pictorial representation, as to how the data at individual levels are used. Each level of architecture has different types 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 proposed method has done significant contribution towards recognizing shapes of lips for tilted faces like faces turned towards left and right are quite a challenging task

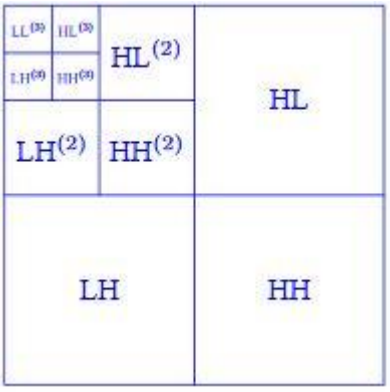


Fig.4. Different levels of splitting of original image into different dimensions

The fig.4 presents a method that has capability of extracting and training system to detect and recognize shapes of lips for sentences.

1. **Grouped Consonants:** The grouped consonants are of 25 types, these 25 consonants are used in Kannada language very often. There are a total of 25 consonants x 10 statistical features x 3 different channels x 10 features dimensions together constitutes to a total of 7500 dimensional data + 2100 data of a video along different frames = 9600 dimensional data.

2. **Miscellaneous Consonants:** Another set of consonants are miscellaneous consonants. There are 10 types of miscellaneous consonants. There are a total of 10 miscellaneous consonants x 10 statistical features x along 3 different channels (R-G-B) x 15 feature vector data + 100 features space data together constitutes 4600.

Thus, finally we obtained a number of features of size 9600+4600 a total of 14,200 of different frames of a video have trained.

B. Proposed Algorithm

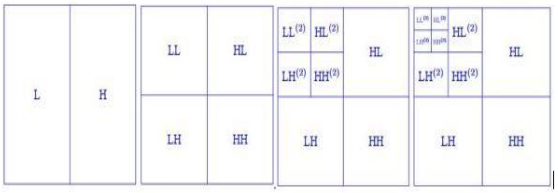


Fig.5. Division of images into different dimensions based on wavelet information.

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

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

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

The parameters like precision and recall plays a significant role in assessing the accuracy of recognizing the movement of lips especially for local language. Some of the research works have been carried out on English language for

measuring the accuracy of lip reading, but the work related to recognition of Kannada Lip Movement is not carried out to a large extent. Thus, we have focused our research method on recognizing Sentences especially on words that uses sophisticated sentences.

Algorithm: Lip Reading for Sentences

Input: Audio with video data is input to the system

Description: System determines the lip movement

Output: Lip movements are recognized for Sentences

Begin **[Pre-processing]**

Step 1:
- Compute eq. (1).

Step 2: **[Concatenation of different dimensions of wavelet features]**
- Step 2.1 Compute eq.(2)
- Step 2.2 Compute eq.(3)
- Step 2.3 Combine with next level of data
- Step 2.4 Compute eq. (4) as per And assign it to eq.(5)

Step 3: **[Wavelet Feature Learning]**
- Step 3.1 Assign labels to shapes of individual frames as per (6)
- Step 3.2 Solve eq. (7)

Step 4: **[Recognitions]**
- Step 4.1 Solve eq. (8)
- Step 4.2 If not an end of frame Go to step 4.1 else go to step 4.3
- Step.4.3 Compute and recognize Shapes

End

IV. RESULTS

In view of annotating, recognizing as per [16,30] and understanding Kannada Sentences, we need to perform certain tasks like preprocessing, annotations, features extraction, features classification are carried out to recognize Kannada sentences.

Table 1. Analysis of proposed method for Lip reading. Significant results are compared with existing efforts.

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 Multi-level Wavelet Lip reading**	**82.83%**

A. Performance Evaluation

Parameters considered for evaluation of performance of a trained systems are precision, recall, sensitivity, specificity, accuracy, as we are measuring the accuracy of a proposed lip reading system in terms of recognizing the sentences that is spoken by a person, it needs to incorporate few parameters mentioned above for evaluation of performance of proposed method with respect to the contemporary methods [29,30].

The effectiveness of the proposed research work using wavelet based feature extraction learning for recognition of sentences shall be observed in terms of accuracy of existing with respect to proposed method, where the proposed has yielded an accuracy of 82..83% accuracy for recognizing Kannada sentences. In addition to some of the state of the art techniques like lip reading in the wild. The average precision is clearly more precise and accurate than other contemporary, which shall be observed in terms of graphical representation of comparison. The system determines the shape of mouth and does many tasks between annotations and prediction of sentences.

The Posterior time complexity of proposed method have been shown in graphical representation as shown in Fig. 7. The proposed methods have many advantages over other existing methods in terms of time complexity, accuracy, precision of accurately predicting sentences. Further recognizing the shapes of lips and the time consumed in recognizing the movements of lips especially for Kannada Language and also helpful to differently abled personalities.

Table 2. Analysis of Significant results are compared in terms of precision

Sl.No	Folds	Precision
1	Fold 1	81.54%
2	Fold 2	84.26%
3	Fold 3	89.66%
4	Fold 4	96.45%
5	Fold 5	94.42%

Table 3 . Analysis of proposed method for Kannada Lip reading in terms of posterior computational aspect

Sl.No	Folds	Time
1	Fold 1	55.16
2	Fold 2	54.32
3	Fold 3	61.56
4	Fold 4	65.45
5	Fold 5	66.42

V. DISCUSSION

Every video of a dataset is divided into equal number of frames, where a total number of frames extracted from a video is divided into 5 folds consisting of equal number of frames in every folds. Hence, the performance has been assessed in terms of posterior computational aspects, accuracy, and precision. Further, the precision results obtained from the wavelet based method is shown in table 2, and the posterior computational aspect results are shown in Fig. 7

The challenging issues we observed in [22,28] while implementing an algorithm to recognize shapes of lips for sequences is quite a challenging task for predicting certain words like ಕ ಕಾ ಕಿ ಕೀ ಕು ಕೂ ಕೃ ಕೆ ಕೇ ಕೈ ಕೊ ಕೋ ಕೌ ಕಂ ಕಃ and ಗ ಗಾ ಗಿ ಗೀ ಗು ಗೂ ಗೃ ಗೆ ಗೇ ಗೈ ಗೊ ಗೋ ಗೌ ಗಂ,ಗಃ as we are more focused towards recognizing lips movement for prediction of Kannada sentences through multiple levels of feature extraction. Even though the words seems to be spelt similar in manner, meaning are seems to be different, therefore some expressions were very difficult to predict. Similarly other challenging tasks are addressed with some facial lip expressions are supposed to be identified with inherent shape changes and these changes are tend to be different and used to measure the lip shapes.

The facial changes that may cause the words to be spoken by a person is determined by analyzing the words annotated with shapes of lips. Thus, the proposed method has defined a way to recognize certain words to recognize the Kannada sentences. The drawback of the proposed method is that the algorithm needs to perform those shape model operation on every change in shape of the lips with respect to time

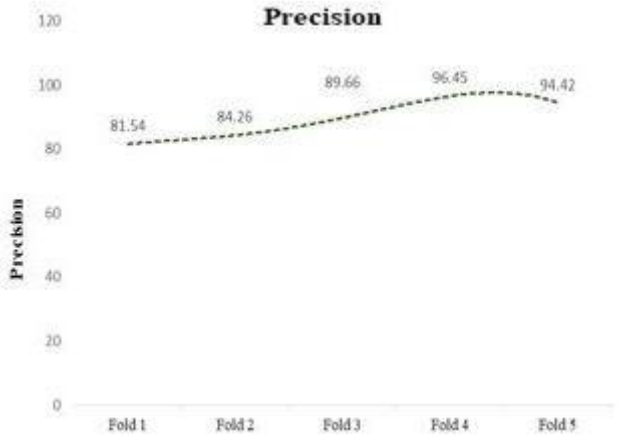


Fig. 6. Comparison of proposed method and its accuracies with respect to precision.

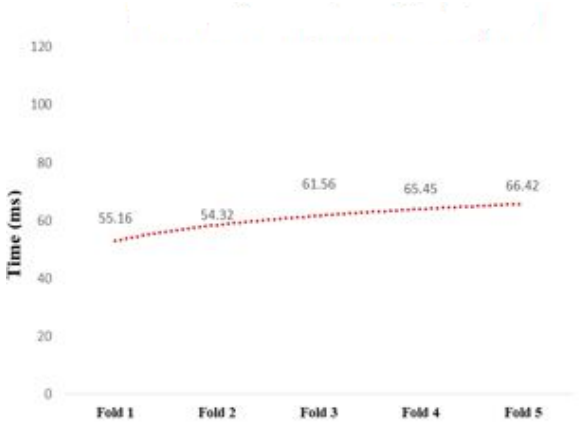


Fig.7. Measure of posterior time complexity for predict of lip movement.

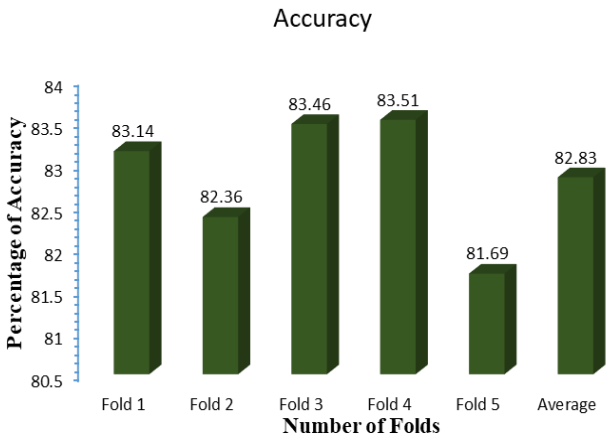


Fig.8. Classification accuracies of proposed method in different folds.

The accuracy of the proposed method with wavelet based multi-level features representation for prediction of words spoken based on lip tracking and lip movement is identified with a metric of accuracy as shown in fig.8. The accuracy of the proposed method is measured at different folds of measure like fold 1, fold 2, fold 3, fold 4 and fold 5.

VI. CONCLUSION

The wavelet based multi-level feature learning method has a few contribution towards recognition of words from lip movement. In addition to few other significant contributions like 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 in every frames of a video. The lip portion from every frames are to be processed for feature extraction for the purpose of recognizing the lip movements and to recognize the Kannada language. The next contribution is towards recognizing lip movements for Kannada language even though the words are similar in nature by expression, they are different by meaning. Thus, we developed a new method of recognizing the lip movement for Kannada Language known as Wavelet based feature learning. Thus our research contributions have thrown light on many new techniques that shall be incorporated into regional languages like Kannada. Lip movement recognition is done with good accuracy in comparison with other methods. Hence, the Wavelet based feature representation learning is a way of achieving the comparison with other methods. Hence, the Wavelet based feature representation learning is a way of achieving the desired objectives at an accuracy of 82.83% for recognition of Kannada Sentences.

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 nd Video Coding, pp. 265-275, 1996.