Lip Reading Recognition

[1]*Jyotsna Uday Swami*, [2]*Dr. Jayasimha S R*,
[1]*Student*, [2]*Assistant Professor*
[1,2]*Department of Master of Computer Applications*,
[1,2]*RV College of Engineering®, Bengaluru, India*

Abstract Machine Learning techniques give computers the capability to train and learn by using sample inputs and provide outputs which leads a model to test the test cases instead of being programmed. Lip reading recognition is a task of converting speech (lip movements) to readable format like text without audio. Lip reading recognition is a technology that helps for communication using devices. The uses of this concept are immense for the people who are who have hearing deficiency and vocally challenged. In situations where there is lot of noise, its very difficult to listen to each other. The noise around us can sometimes cause a huge impact on the conversation which may lead to miscommunication, the example can be a phone call in a theatre. In future, this problem can be solved by Lip Reading Recognition Technology. This technology helps one to transmit information without using vocal cords. This technology uses Lip Tracking which is a biometric system, it works based on a genuine system that can be developed using different levels of video processing, and hence it's possible to get lip contour and the exact location of key points in the subsequent frames which is usually referred as lip tracking. This technology works on the lip movement and generates frames where the other end person can easily get the result of it.

***Index Terms*- Lip Reading Recognition, Video frames.**

I. INTRODUCTION

Lip Reading Recognition technology is mainly used for speech tracking by means converting the speech into understanding format by externally read the developments of the lips, face, and tongue. In lively situations, where the acknowledgment of any discussion is troublesome, which can be visual discourse acknowledgment that gives a successful way to get discourse to get better results. Lip reading can be difficult sometimes because of the different types' accents, speed of talking, facial highlights, skin shading, and so forth. However, there are a large group of utilizations, because of which this issue expects noteworthiness. It is extremely useful to remove the communication gap especially when it is communicated in a noisy place, and so forth. Lipreading is a difficult task for anybody without the knowledge, especially without any context of the concept.

Most lip-reading actions, besides the lips may sometimes make use of tongue and teeth to pronounce a word or letter which are latent and difficult to disambiguate without context. Lip reading is a complex issue because of different kind of lip shapes & color. Lip-reading is similar to guesswork and will not be accurate because of pronunciation differences, bad position of faces, having hands over mouth, moustaches, too fast speech, low lighting, and beards etc.

The lip-reading recognition has been documented since long time, which includes the people with hearing deficiency use lip-reading method where it can be an adjustment to get and understand the fluent speech of the speaker. When it comes to converting these kinds of issues into automation there can be many difficult challenges comparing the ordinary recognition which can be audio. These audio speeches which has the pre-defined units are known as phonemes, mapping between the phonemes and the words are given by the pronunciation dictionaries. And its easy to map video than audio as a result because of the compressed video rates are higher than uncompressed audio. So, it's a easy and efficient way to get the better results using video than the audio formats.

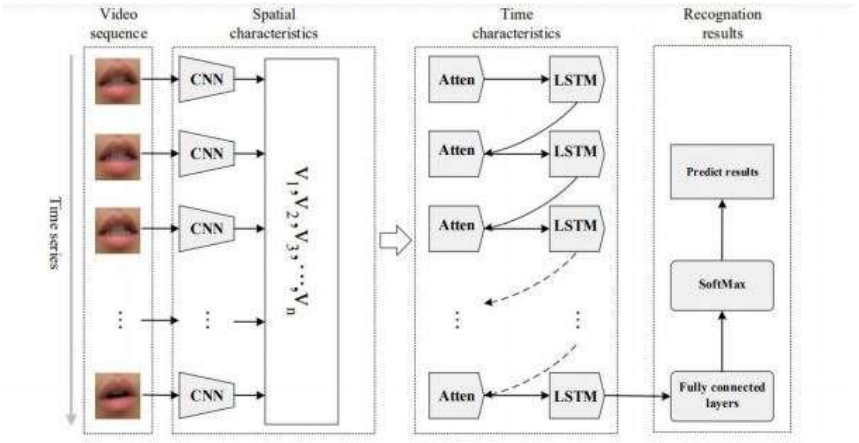

Fig. 1: Architecture Diagram

II. LITERATURE SURVEY

Speech recognition is used for recognizing the speech that relies on the lip movements speech without relying on the audio stream. This is especially useful in noisy surroundings where the audio signal is poor, and it can with acoustic speech recognizers in order to compensate for the poor performance due to noise [1].
The same two-level approach can be followed in deep learning concepts, where extraction step can be replaced by deep auto encoders and also Hidden Markov Models can be replaced by Long-Short Term Memory networks [2].
Recently, different end-to end research have been published. Such researches use either fully connected or convolutional layers [3].
The two-stream model is further extended with the state-of-the art approach for recognition of isolated words. This method consists of 3D convolutional layer that is followed by softmax layer, Bidirectional Gated Recurrent Unit and 18-layer Residual Network [4].
The former method is more relevant to tasks like isolated word detection, recognition and classification, also the latter to Sentence level differentiation and large vocabulary continuous speech recognition. however, recent publishes in speech recognition and natural language processing declare that direct modeling of words is efficient even for and large vocabulary continuous speech recognition [5].
A fully Long-Short Term Memory architecture is can be used for recognition, which gives superior results than traditional methods on the GRID audio visual, in an end-to-end sentence level lip reading recognition network is introduced, which combines spatiotemporal convolutional layers, Long-Short Term Memories and Connectionist Temporal Classification. It obtains 95.2% sentence level accuracy of the speech from GRID database, while trained on the all the left-over GRID speakers [6].
The new large-scale dataset has different generations and statistics, in which the LRS2-BBC feature is used to train the model and also evaluate it. The dataset has talking faces images and videos with a significant variety of facial poses. The network training gets the report a form of curriculum learning which is used to accelerate training of model. Finally, evaluation of the performance tastes with the models, which also includes for lips visual input only [7].
In most of the lip-reading models there will be two modules, the frontend and the backend module. Clip-level features, frame-Level features and Local motion patterns get attentions by the frontend module. The whole sequence level patterns and temporal dynamics of the sequence based on the output of front-end module are monitored by backend module [8].
Cosine learning rate scheduling to perform training with end-to-end procedure is proposed by Martinez [9].
There is a long history of research on lip reading. The most traditional methods are based on hand-crafted features along with shallow models, and so on and so on. With the developments of the deep learning techniques, researchers in sthis lip-reading model started introducing deep neural networks in recent years [10].

III. WORKING OF LIP-READING RECOGNITION WITH DIFFERENT TECHNOLOGIES

3.1 Dynamic time warping

The distance between two data and path wrapping values from sound are calculated by optimal wrapping path and the optimal wrapping path is calculated by Dynamic time warping algorithm. The distance between comparison is known as warping path. The two paths are said to be same, if the warping path produced are smaller. In different times two words from the same word by the same user can occur. For example, to can be pronounced with to or too. The time alignment problem is efficiently solved by Dynamic time warping method. Therefore, this algorithm is most often used in calculating pattern similarities. The processed data will always remain in time zone, so the sequence of data will vary at time. The following diagram illustrates between two sequence of numbers.

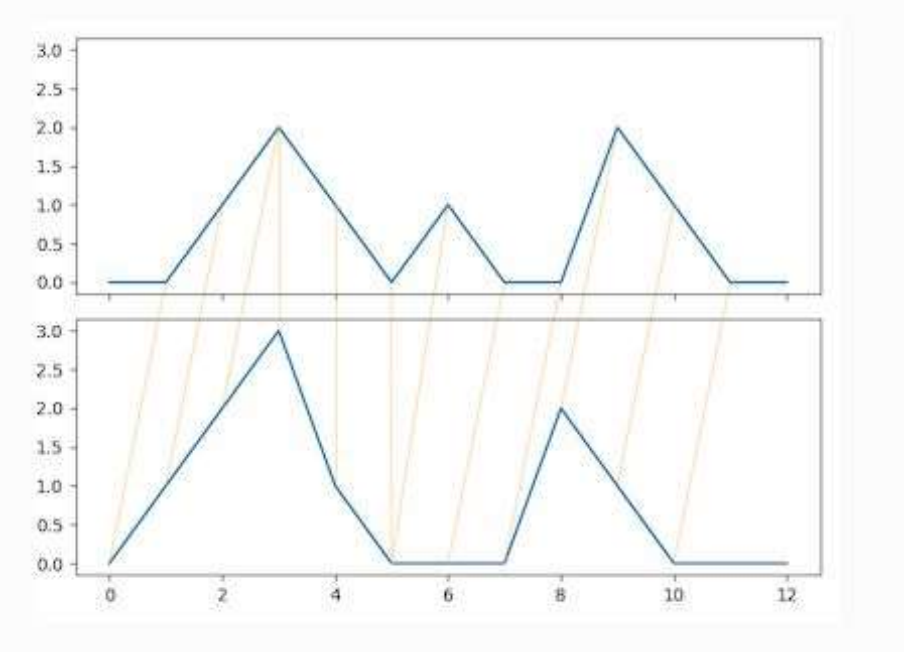

Fig. 2: Dynamic time warping

3.2 Shape template

shape template with snakes is used to extract lip contours from image. The parameterized shape template describes the lips. An object is modelled within this shape template. The model can be made to fit the object in the image by adjusting its parameters. The template shape is based on previous knowledge of shape of the lips. To describe the shape of upper and the lower lips shape template consist of parabolas. These shape templates will interact dynamically along with image through an energy function which will draw the shape template on to salient features by changing the parameters of the parabolas. To produce fields containing features of interest such as valleys, peaks or edges are done by preprocessing of image. These fields are interacted through energy function by shape template. As the energy equation is minimized the parabolas will tend towards already known lip shapes that is the advantage of internal potentials in shape templates. These reduces the probability of shape template getting on wrong medium which may be possible with snakes. The limited flexibility of using predefined lip shapes is the major drawback of shape template.
The following figure explains how templates are used to improve lip reading recognition.

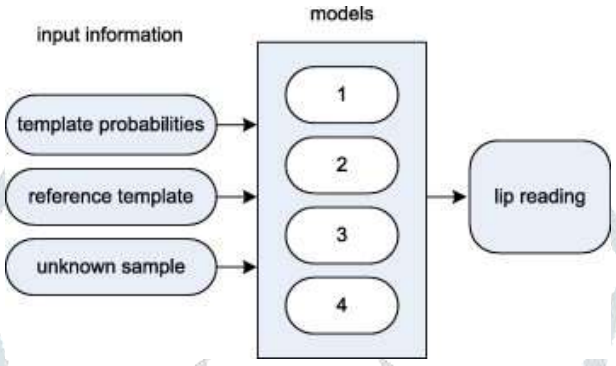

Fig. 3: Shape Templates used as inputs to Lip reading.

3.3 Snake's approach

The most effective lip-reading system is developed by Bregler and Omohundro [11] which was the combination of trained lip shapes and snake algorithm. Which also contains the technique of nonlinear manifolds. In the feature space the lips are represented as points by configuration and also the surface or manifold of the space contains all the possible lip configuration. The configuration space produced by trained dataset of points and the structure and dimensions points lie is induced. Manually controlled snakes are trained using collected data. Unknow lip shapes are tracked using the snakes which are controlled by using learned manifold. From lip manifold of the image is estimated back project using initial crude. The grey level gradient is result of one iteration around the snake points. After completion of each iteration the lip manifold gets updated by snake point projection, so throughout the iteration only legal lip shapes are considered. The snake is trained to converge only to legal lip shapes. The training data collected using manually controlled snake will determine the legal lip shapes. The system mentioned above uses modified snakes which controls through the use of learned lip shapes. It uses two methods that adapts to different lips without any previous training. The snakes can be controlled by two-dimensional pattern templates of lip edge contour instead of images gradient. The snakes should be able to provide the dynamics to move with the lips and adapt new shapes, while the pattern template should be able to provide stability.

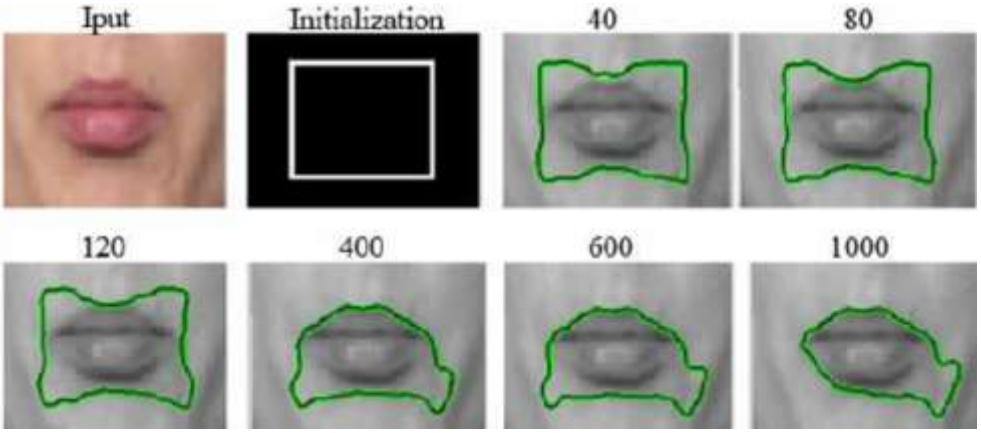

Fig. 4: Lip localization using Snake's approach on input image. The number above each image shows the number of iterations needed for snake parameters tuning.

3.4 Hidden Markov Model

Hidden Markov Model is trained with visual features which has three different states and diagonal covariance with Gaussian Mixture Model concerned with each task which all are sequence of embedded trained and which has tested with view angle dependence. Visual speech synthesis approaches and the generated visual parameters by HMM which uses constraints of dynamic ("delta") features. Visual parameter trajectories can predict the mouth motions rendered by the video. The drawback of Hidden Markov Model based in visual speech synthesis can be the resulted blurring because of the feature dimension which reduces statistical modeling. With this automatic lipreading which uses continuous density Hidden Markov Models as a pattern matching which is done by statistical components. The mixture of diagonal covariance matrix and the multi-dimensional gaussian mixtures are the result of the probabilities of the observed pattern matching of Hidden Markov Model. This method is specifically considered for lip-reading recognition where the independent context, the consideration of whole word, lip shape, facial expression, accent, etc. With the exception of some results which can be not that accurate considering different scenario all the Hidden Markov Model parameters are estimated by Viterbi training dataset, where a different training algorithm is used which is known as generalized probabilistic descent. To train this the available audio and the identical visual counterparts are synced together to obtain a string length which will be used for recognition.

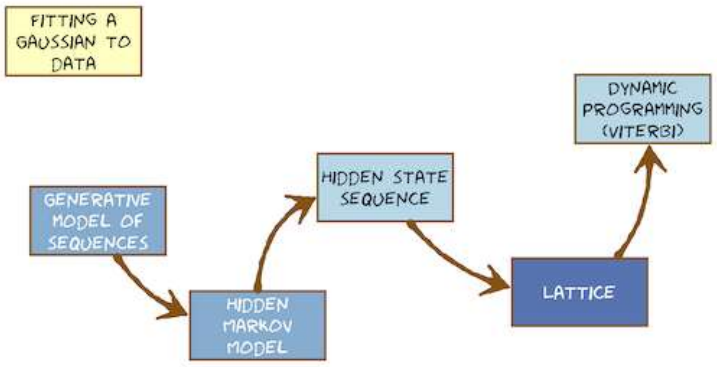

Fig. 5: Speech Recognition using Hidden Markov model

3.5 Convolutional Neural Networks Model

Lip reading issue can be resolved by CNN and LSTM methods, by using the excellent convolution neural networks the features of images are derived. CNN model contains 3 different convolution layers, these 3 layers are designed with regularization. AlexNet, GoogleNet supported CNN by structured are employed to differentiate and compare the performance with designed with CNN model. Classification model and dataset size can affect the performance of the module. Data augmentation techniques uses address data augmentation. Experimental studies that are based on CNN models which can be with or without augmented dataset it depends on various mini batch size values are to be performed.

- AvLetters dataset is used to transfer learning which is supported by CNN by using AlexNet
- CNN architecture is also compared and proposed by supporting transfer learning of CNN
- This CNN method is not only having better performance it is also easy to develop. It provides the whole information of described CNN and the model which are pre-trained used throughout the research.

Convolutional Neural Networks (CNN) comes under artificial neural network which is specialized to maintain multi-dimensional and large set of data. Least layers of convolution process is used by convolutional networks which are neural networks instead of using general matrix multiplication. The main and the basic components of CNNs are pooling layer, activation functions and convolution layer, fully connected layers, regularization, loss layer, optimization. The learnable filter set is included by convolutional layer. The ability to learn with a fully connected layer is the structure layer of convolutional network.

The important parameters for this layer are number of filters, stride and spatial extend. The pooling layer reduces the network model's cost. Pooling layer makes the system resistant toa very small position changes, the pooling method

constantly uses the operations such as Minimum, Maximum, Sum, Average, etc. There is also choice of activating the function as ELU, sigmoid, ReLU, etc., apparently it also affects the performance of CNN. ReLU, linear function which is piecewise, the return negative inputs make to zero activation function and, also without changing the output positive input. To achieve higher classification accuracy ELU which is an activation function is used, which also allows neural networks.

The frequently used function is sigmoid function which continuous and derivable function. After convolution and pooling layers fully, connected layer comes in CNN. This study has all the layers and neurons are fully connected to previous layers. The last layer of CNN is the Loss Layer, which results the comparison should be evaluated between the predicted labels and the final evaluated labels during the training of the model. To prevent loss function issue SoftMax is used and to resolve overfitting problem Regularization techniques is used, which is important the most important problem for deep neural networks. The following diagram illustrates how the CNN and LSTM methods are divided into two layers and by layering them how we get the results.

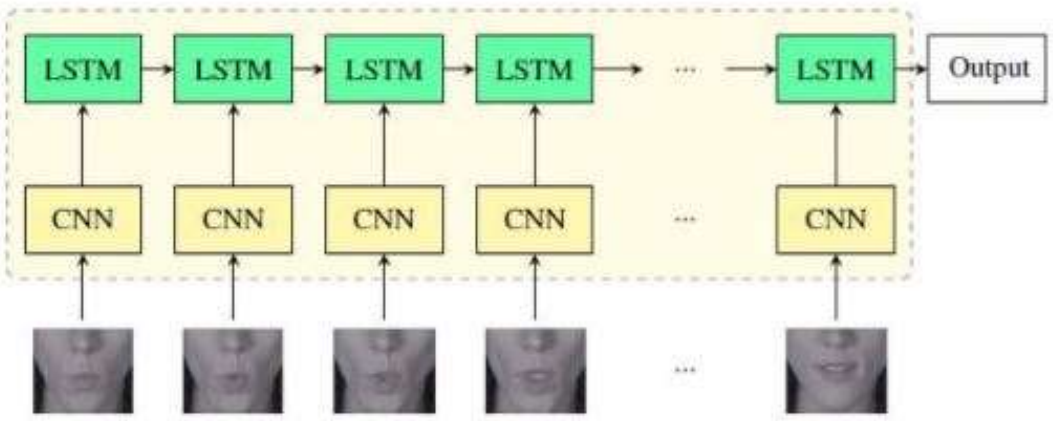

Fig.6: Lip-Reding Recognition using Convolutional Neural Networks (CNN) model.

IV. RESULTS

The research still on for the model for lip tracking is done for the most accurate and the strong processing model. In lip-reading many researchers have strive to get the perfect and the efficient processing pipeline with expect to overcome the inherent issues. This paper quest in the opinion may not be that efficient and relevant to the lip-reading recognition results as the general assumptions. To obtain recognition rates the entirely different set of methods can be tested with accuracy which is sufficient resulted by transformation, which is very important resultant for the community of lip-reading. The researchers may use the plugs to track the mouth this method is appeared as the most robust in this era, then the recognition modules which concentrates on assuming on the plug. The CNN method introduced in this paper is also used in several different kind of experiments till now which appears to be very competitive method till now with the others while competing with some strong and accurate points on in its favor. Along with CNN the different mentioned equivalent geometry methods and non-geometry-based methods suggests the results from associated processing models.

V. CONCLUSION

In this paper, the comparison of different modules of lip-reading which contains Dynamic Time Warping, Shape Template, CNN, Snake's approach and Hidden Markov model. A dataset subset from GRID is applied to generate the result of performance in the particular term of accuracy rate of word. To obtain the CNN features, a novel LSTM is utilized to get and track all the set of targeted shape points on the lips. Each model is mostly verified by different kinds on independent speakers who has different speaking manners, tone, accent, and lip shape for testing. From the general observation in CNN method with different types of appearance parameters to perform more efficiently with other types of feature, which concludes that the appearance gives more information than the shape to get the readings. In the final output shows that even the pixel-based methods and LSTM can be used for normalizing the images with removal of the shapes, with fine variation from the region of interest. The particular improvement in recognition work was observed with the snake's and the shape template method. This paper can be modified with further advance work with the following tasks. The recognition of words can be replaced with the classifiers which will be the trained visemes for recognition. These can be the smallest units of the visuals which can be distinguished in reading the lips and it can be similar to audio speech phonemes. On the other hand, as per the observation the shape normalized appearance also performs similar to CNN feature, it is equivalent to its inner appearance of the mouth area which contains most affective information, despite the fact that further researches will need to examine these features on the designed on the inner appearance of the mouth with helps to get the perfect readings for better results.

REFERENCES

[1] Brais Martinez, "lipreading using temporal convolutional networks", Samsung AI Research Center, Cambridge, UK, 23 Jan 2020

[2] Y. Yamaguchi, "Audio-visual speech recognition using deep learning," Applied Intelligence, vol. 42, no. 4, pp. 722– 737, 2015.

[3] T. Stafylakis and G. Tzimiropoulos, "Combining residual networks with LSTMs for lipreading," in Interspeech, 2017. "SAS Visual Analytics" ["Online"] Avaiable : https://www.sas.com/en_us/software/visual-analytics.html. Accessed: 23-Mar-2018.

[4] S. Petridis, T. Stafylakis, P. Ma, F. Cai, G. Tzimiropoulos, and M. Pantic, "End-to-end audiovisual speech recognition," in IEEE Int'l Conference on Acoustics, Speech and Signal Processing, 2018.Sankey Flow Show - Attractive flow diagrams made in minutes!" [Online]. Available: http://www.sankeyflowshow.com/. [Accessed: 28-Sep-2017].

[5] K. Audhkhasi, B. Ramabhadran, G. Saon, M. Picheny, and D. Nahamoo, "Direct acoustics-to-word models for english conversational speech recognition," arXiv preprint arXiv:1703.07754, 2017.Google, "Google Dashboard" [Online]. Available: h ttps://sites.google.com/a/pressatgoogle.com/googledashboard/.

[6] "Sisense Reviews | TechnologyAdvice." [Online]. Available: http://technologyadvice.com/products/sisense-reviews/. [Accessed: 09-Oct-2017].

[7] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman. Lip reading sentences in the wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017.

[8] Themos Stafylakis and Georgios Tzimiropoulos, "Combining residual networks with lstms for lipreading. in: Interspeech 2017, 20-24 august 2017, stockholm, sweden.,"

[9] Brais Martinez, Pingchuan Ma, Stavros Petridis, and Maja Pantic, "Lipreading using temporal convolutional networks," in ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020, pp. 6319–6323.

[10] Joon Son Chung and AP Zisserman, "Lip reading in profile," 2017.

[11] C. Bregler and S. M. Omohundro. Nonlinear manifold learning for visual speech recognition. In Proceedings of the Fifth International Conference on Computer Vision, pages 494– 499, Los Alamitos, CA, June 1995. IEEE Computer Society press.

[12] W. Feng, "Audio visual speech recognition with multimodal recurrent neural networks," in 2017 International Joint Conference on Neural Networks (IJCNN), pp. 681– 688, May 2017.

[13] M. Yuksel "Performance improvement of deep neural network classifiers by a simple training strategy," Engineering Applications of Artificial Intelligence, vol. 67,pp. 14 – 23, 2018.

[14] H. Badem "Classification of high resolution hyperspectral remote sensing data using deep neural networks," Journal of Intelligent & Fuzzy Systems, vol. 34, pp. 2273–2285, 04 2018.

[15] A. Ben-Hamadou "Human machine interaction via visual speech spotting," in Advanced Concepts for Intelligent Vision Systems (S. Battiato, J. Blanc-Talon, G. Gallo, W. Philips, D. Popescu, and P. Scheunders, eds.), (Cham), pp. 566–574, Springer International Publishing, 2015.

[16] S. Agrawal, Ranvijay, "Lip reading techniques: A survey," in 2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT), pp. 753–757, July 2016.

[17] J. Noyola, "Lip reading using CNN and LSTM," in Technical Report, 2016.

[18] Y. Takashima, Y. Kakihara, R. Aihara, T. Takiguchi, Y. Ariki, N. Mitani, K. Omori, K. Nakazono, "Audio-visual speech recognition using convolutive bottleneck networks for a person with severe hearing loss," IPSJ Transactions on Computer Vision and Applications, vol. 7, pp. 64– 68, 2015.

[19] S. Das, "CNNs architectures: LeNet, AlexNet, VGG, GoogLeNet, ResNet and more . . . ." https://medium.com/@siddharthdas-32104, 2017

[20] K. He, X. Zhang, S. Ren, J. Sun, "Deep residual learning for image recognition," CoRR, vol. abs/1512.03385, 2015

[21] Chung J.S., Zisserman A. (2017) Out of Time: Automated Lip Sync in the Wild. In: Chen CS., Lu J., Ma KK. (eds) Computer Vision ACCV 2016 Workshops. ACCV 2016. Lecture Notes in Computer Science, vol 10117. Springer, Cham

[22] Anina I., Zhou Z., Zhao G., and Pietikinen M. (2015) "OuluVS2: A multi-view audiovisual dataset for non-rigid mouth motion analysis", In Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition (FG15), Ljubljana, Slovenia, 1-5.

[23] A. Mesbah, A. Berrahou, M. El Mallahi, H. Qjidaa. Fast and efficient computation of three-dimensional Hahn moments. Journal of Electronic Imaging, 25 (6), doi: 10.1117/1.JEI.25.6.061621, (2016)

[24] T. Stafylakis, G. Tzimiropoulos, Combining Residual Networks with LSTMs for Lipreading, In Interspeech, (2017)

[25] Stavros Petridis, Maja Pantic, "Deep Complementary Bottleneck Features for Visual Speech" IEEE, pp. 2304-2308, 2016.