International Scientific Journal of Engineering and Management
Volume: 02 Issue: 05 | May – 2023 DOI: 10.55041/ISJEM00975 ISSN: 2583-6129
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata www.isjem.com

TEXTIFY – LIP READING TO TEXT

Deeksha Shetty K, Kshama D, Rakshitha, Swathi

A J Institute of Engineering and Technology, Mangalore

Mrs. Sharanya P S

(Project Guide)

Assistant Professor, Department of Information Science and Engineering,

A J Institute of Engineering and Technology, Mangalore

Email: sharanya@ajiet.edu.in

Abstract-- Textify is the process of identifying spoken words by watching lip movement. Visual listening techniques include lip reading. The ability of Textify to extract text from streaming or recorded video may present opportunities for solving connectivity and practical problems. This entails attempting to decipher the speaker's meaning from the expression on their face. For the conversion, the deep learning approach can be used to analyze lip movement. The system that is being proposed in this project transforms video data into text data by lip-reading. Text conversion is the most effective way to speak to people who are deaf or dumb. The application will receive frames from a live or recorded video as input. Additionally, the application must track down and understand the feature patterns.

Keywords- lip reading, deep learning, convolutional neural network.

I. INTRODUCTION

Recently, there has been a lot of research interest in automated lip reading, and this discipline has made a lot of progress thanks to machine learning-based methods. Automated lip reading is sometimes referred to as "visual speech recognition" when it is done without audio support. Recent deep learning-based automated lip-reading techniques mainly concentrate on decoding protracted speech segments in the form of words and sentences by using words or characters as recognition classes. As it is still difficult to automatically lip-read people speaking sentences with a wide variety of vocabulary and words not present in the training phase, word-based approaches have generally been more successful in achieving correctness than lip-reading sentences. This is so that lip-reading systems that employ words or characters as classes can only predict words that the systems have been trained to predict. Words must be encoded as classes during the training phase.

Recent technical advancements have generated a lot of interest in lip reading systems, which have shown to be useful in a variety of applications include converting audio into legible text for meetings and spy cameras in government operations. Focusing on facial clues and emphasizing the lips from any aspect of the face are the first steps in a lip-reading method. By tracking lips of various sizes and shapes, this technique has the advantage of enabling the system to obtain an accurate set of points for the lips. The most difficult part of tracking lips, however, is identifying their movements because everyone's lips move differently because they are all different sizes and shapes. The spacing between the lips is predetermined for each unique phoneme, which is the sound made when a syllable is spoken in any language, which connects these movements. As a result, the lip movements will be categorized according to this distance, which is unique to a syllable and is known as the Euclidean distance.

II. PROBLEM STATEMENT

The majority of the time, stupid people use sign language to communicate, yet they often struggle to interact with non-sign language users. Since there is a barrier between these two communities, written information is the most effective way to communicate with D/deaf people. People who are born deaf or hard of hearing may find reading more challenging because it requires them to make the connection between how a word looks and sounds. For kids who are born deaf or hard of hearing, this is especially true. Communication between two communities won't be a problem. People with hearing loss, D/deafness, or other hearing impairments will find the lip reading to text app to be more convenient.

Whether audio is available or not, a technique

ISSN: 2583-6129
DOI: 10.55041/ISJEM00975
www.isjem.com

called Textify, uses lip movement to decipher spoken words. The Lip-Reading device will function using data from both live video and movies or videos. These systems are highly beneficial for deaf individuals since they convert audio into text that is simple to read.

The input for the application will be frames from a complete video. The application must be able to tell the face (lips) apart from the rest of the edge and understand, follow, and interpret the minute-by-minute examples of the lip over the time limit. This can be done using Deep Convolutional Neural Network (CNN) Model and computer vision (feature extraction). Lip reading framework is tough to implement because to complex picture handling, challenging classifier preparation, and lengthy recognition procedures. Automated lip-reading technology is a vital part of human-computer interface technology.

III. LITERATURE SURVEY

Saakshi Bhosale, Rohan Lure et. ed [1] in "A utility to change over Lip movement into Discernible text" have characterized that programmed lip-concentrating on age is an absolutely significant issue of human-PC transaction period. For visual perception and human communication, lipreading is essential. To translate spoken language, the application extracts lip movements from videos. Feature extraction and categorization are involved in conventional lip-reading systems. In computer vision, encompassing image processing, object detection, human behavior recognition, and video analysis, deep learning has made considerable strides.

Feng hour souheil et. A lip-reading system based on neural networks is suggested in ed. [2]'s article titled "Lip Reading Sentences Using Deep Learning with Only Visual Cues." When evaluated on the BBC Lip Reading Sentences 2 benchmark dataset, a viseme-based lip-reading system that just uses visual cues and without a lexicon demonstrated dramatically improved performance with a 15% reduced word error rate compared to recent efforts. Visemes as classes enable classification of words not seen during training and decrease computational bottlenecks. They can also be applied to other languages. Visemes, on the other hand, are shorter than words, increasing visual ambiguity and decreasing the amount of temporal information available for class detection, necessitating special considerations while designing such a system.

Mandar Gogate, Ahsan Adeel, and in "Lip-Reading Driven Deep Learning Approach for Speech Enhancement," ed. [3] offers a novel framework for speech enhancement based on lip-reading. This research suggests an audio-visual speech enhancement architecture that makes use of both deep learning and analytical acoustic modelling. It uses an improved Wiener filter and a novel lip-reading regression model to estimate the clean audio spectrum. The framework's effectiveness in enhancing speech quality and understandability is demonstrated in dynamic real-world circumstances. The deep learning-driven lip-reading model for context-aware, autonomous AV speech augmentation is currently being improved.

Nadeem Hashmi, Saquib, and A lip-reading model for audio-free video data with variable-length sequence frames was presented in ed. [4] in "A Lip-Reading Model Using CNN with Batch Normalization." They start by isolating the lip region from each face image in the video sequence and concatenating them into a single image. In order to train the model and extract the visual features from start to finish, the next step is to build a twelve-layer convolutional neural network with batch normalization on two levels. Batch normalization can be used to lessen both internal and external variations in a variety of characteristics, including the speaker's accent, the quality of the lighting and picture frame, the speaking rate, and posture.

Myo Kalyar San et al. A work on lip movement recognition for Myanmar consonants was proposed in ed. [5] in "Lip Movements Recognition Towards an Automatic Lip-Reading System for Myanmar Consonants". The paper suggests a color transformation, algorithm, and classifier-based lip-reading method for Myanmar consonant recognition. Additionally, it offers a visual training technique for hearing-impaired people that makes use of the AVASR system and facial feature extraction. The system is flexible in an unrestricted context and successfully distinguishes facial features.

Warunee Nittaya and other the lip-reading computer assisted instruction (CAI) for students with hearing impairment was proposed by ed [6] in "Thai Lip-Reading CAI for Hearing Impairment Student." The CAI system consists of a multiple-choice game and a lesson. Students with hearing impairment in elementary school benefit from better pronunciation. Ten students were used to test the method, and the results revealed that after using it repeatedly, the rate of accurately identifying mouth shapes increased.

Apurva H. Kulkarni and others [7] in the book "Artificial Intelligence: The book "A Survey on Lip-Reading Techniques" examines various lip-reading methods and language datasets in the age of deep learning. This essay describes the key elements of lip-reading technology and offers an overview of automatic lip-reading techniques. Additionally, they talk about lip-

ISSN: 2583-6129
DOI: 10.55041/ISJEM00975
www.isjem.com

reading difficulties like phoneme confusion, irregular lighting, and awkward head postures. The objective is to help deaf people communicate more easily in noisy environments. DNNs need a large amount of training data, and a generalization dataset must be created. Patients with injuries to the vocal cords or throat can also benefit from this technology, which can also be used to examine CCTV footage.

Fan Yang, Jianguo Wei, et al. In "Three-dimensional Joint Geometric-Physiologic Feature for Lip-Reading," ed. [8] In order to capture similarities in speech patterns, a 3D lip physiology feature was derived from color photos using a new deep learning technique called Dense Nets. In the final Dense Nets layer, the color and lip features were combined. It has proven beneficial to incorporate depth information into lip reading utilizing affordable RGB-D cameras, however due to the large difference in pronunciation between speakers, typical feature extraction techniques are insufficient for describing the information of lip movements.

Kurniawan, Adriana, and According to ed. [9] in "Syllable-Based Indonesian Lip-Reading Model," Lip reading is a technique for turning videos into text that uses visual speech recognition. To deal with input variances, a 3D architecture and deep learning are used. A syllable-based model is suggested because predicting words outside of the dictionary presents an OOV issue. In order to increase the data, augmentation is done 40 times, producing testing accuracy of 100%. Ten OOV words, however, display a lower accuracy of 80%.

Jiadi Yu, Li Lu, and A lip reading-based user authentication system was proposed by ed [10] in "Lip Reading-Based User Authentication Through Acoustic Sensing on Smartphones". The paper proposes Lip Pass, a user identification system based on lip reading that uses acoustic sensors on smartphones to extract specific behavioral traits from users' speaking mouths. The system uses a deep learning-based approach to characterize mouth movements and build multi-class identifiers, binary classifiers, and spoofer detectors for mouth state identification, user identification, and spoofer detection. Lip Pass achieved 90.2% accuracy in user identification and 93.1% accuracy in spoofer detection through rigorous tests involving 48 participants in four genuine scenarios. The system aims to provide a more secure and reliable method of user authentication using lip reading.

IV. SYSTEM DESIGN

Proposed system for lip-reading model:

- Predict word/phrase from lip movement, with/without audio input.
- Stored/live videos/images.
- Concatenate frames to create a sequence. Adjust for variations.
- Train CNN with batch normalization on concatenated images to classify words/phrases and extract relevant info.
- Extract visual info from upper/lower lips, including width, height, area, perimeter. Depict mouth shape.
- Uses face detection model to extract faces and Haar Cascade Facial Landmark detector to isolate lips.
- Trained on LRW dataset of 1000 words spoken by hundreds of speakers.

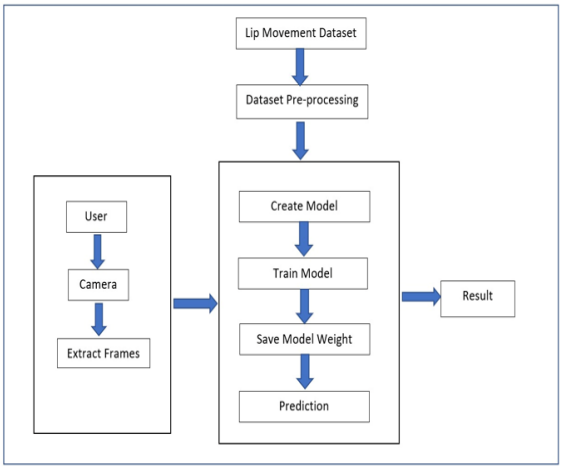
The proposed lip-reading model aims to predict words/phrases from lip movement, with or without audio input, using a three-step process that involves concatenating frames, training a CNN with batch normalization, and extracting visual info from lips. The model utilizes face detection and Haar Cascade Facial Landmark detection to isolate lips, and is trained on a large dataset of words spoken by various speakers.

Architectural Design

Architecture of a lip reading to text system:

- User speaks in front of camera.
- Camera captures video, sent to system.
- System extracts frames, preprocesses them (remove noise, normalize, crop).
- Frames used to create dataset, undergo preprocessing (e.g. data augmentation).
- CNN created to extract features from preprocessed frames.
- Model trained on dataset to map visual features to words.
- Model weights saved for future use.
- To transcribe new speech, system extracts feature and uses language model to produce word probability distribution.
- Output is sequence of transcribed words, may undergo post-processing.

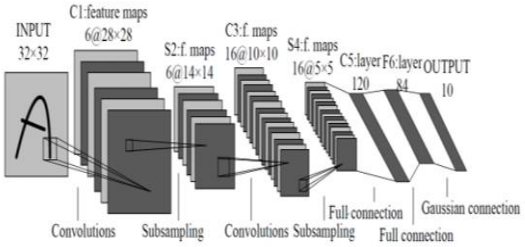
Lip reading to text system architecture involves the user speaking in front of a camera, with the captured video being preprocessed and used to train a CNN to map visual features to words. The system can transcribe new speech using the trained model to produce a word sequence as output.



Architectural Design

Working of CNN

In lip reading to text systems, convolutional neural networks (CNNs) are used because they have been shown to be successful in extracting pertinent characteristics from lip movements. Lip reading to text is a multi-step process that begins with the video input of a speaker. This input has been preprocessed to isolate the mouth area and remove any background noise that might have an impact on the system's accuracy.



Workflow of CNN Diagram

In lip reading to text systems, convolutional neural networks (CNNs) are frequently employed. The steps in the procedure are as follows:

- A person speaking is captured on camera and fed into the lip-reading system.
- Background noise is removed from the video and the mouth area is isolated.
- A CNN is then fed the preprocessed footage, and it extracts pertinent elements from the lip movement.
- A dataset of labelled lip movements and the related text transcripts is used to train the CNN.
- By using the mastered mapping between fresh lip movements and text, the CNN can predict the text transcription of new lip movements once it has been trained.
- To fix mistakes and increase accuracy, extra postprocessing procedures may be applied to the anticipated text.

In recent years, the application of CNNs in lip reading to text systems has demonstrated encouraging results, with high accuracy rates attained on a variety of datasets. The architecture and CNN hyperparameters, as well as the quantity and diversity of the training dataset, all have a significant impact on the system's performance.

Working of Haar Cascade Algorithm

The Haar Cascade algorithm is a popular object detection algorithm used in computer vision. It works by detecting patterns of intensity changes in an image. In the context of lip reading to text, the Haar Cascade algorithm can be used to detect the position and shape of the lips in an image or video frame.

- Data collection: Collect a dataset of images or video frames of people speaking. These could be recordings of people speaking specific words or phrases.
- Preprocessing: Preprocess the images or video frames to extract the region of interest (ROI), which is the area containing the lips. This can be done using techniques such as face detection or facial landmark detection.
- Training: Train a Haar Cascade classifier using the preprocessed images or video frames. This involves selecting positive samples (images containing the lips) and negative samples (images not containing the lips), and using a machine learning algorithm to learn the features that distinguish the positive and negative samples.
- Testing: Apply the trained Haar Cascade classifier to new images or video frames to detect the position and shape of the lips. This can be done by sliding a window over the image or video frame and classifying each window as containing the lips or not.
- Lip reading: Once the position and shape of the lips are detected, use techniques such as optical flow or lip shape analysis to convert the lip movements into text.

The Haar Cascade algorithm can be a useful tool in the process of lip reading to text by providing accurate detection of the lips in an image or video frame. However, it is important to note that lip reading is a challenging task and can be affected by various factors such as lighting conditions, occlusion, and individual differences in lip movement.

V. METHADOLOGY

Lip reading to text has the potential to be a valuable tool in improving communication accessibility and inclusivity for people with hearing impairments.

- Data collection: Video data of people speaking needs to be collected. This data can be obtained by recording

ISSN: 2583-6129

people speaking in different environments, lighting conditions, and camera angles. The more diverse the data, the better.

- Data preprocessing: The collected video data needs to be preprocessed before it can be used for lip reading to text. Preprocessing involves removing noise, adjusting the contrast and brightness of the video, and aligning the video frames.
- Feature extraction: The preprocessed video data is then used to extract features. This involves identifying the key visual features of the mouth and surrounding area, such as the shape and position of the lips, and mapping them to a numerical representation. Popular feature extraction techniques include deep learning-based methods such as Convolutional Neural Networks (CNNs).
- Model training: Once the features have been extracted, the next step is to train a lip reading to text model. This involves using the extracted features as input to a machine learning algorithm, such as a neural network, to learn how to predict the corresponding words or phonemes from the video data.
- Evaluation: The final step is to evaluate the performance of the lip reading to text model. This involves testing the model on a separate set of data to see how accurately it can transcribe spoken words.

VI. IMPLEMENTATION

- **Dataset Creation**

The system detects the mouth region in real-time video frames using OpenCV's Haar cascade classifier. It captures frames from the default camera and resizes them using a scaling factor. The classifier is applied to the grayscale image of each frame to detect the mouth region. If the classifier detects a mouth region, then crops the mouth image and saves it to a specified folder. It also draws a green rectangle around the detected mouth region on the video frame. This process is repeated for a specified number of frames, and then the system stops capturing video. The system useful for collecting a dataset of mouth images for lip reading to text applications. The dataset can be used to train a machine learning model for lip reading.

- **Building the CNN Model**

The system trains a Convolutional Neural Network (CNN) model for lip reading to text using the Keras library. It reads in images from a specified directory, preprocesses them using data augmentation techniques, and splits them into training and validation sets. It then defines the architecture of the CNN model, with convolutional and classification layers. The model is compiled with categorical cross-entropy loss, and metrics such as precision, recall, specificity at sensitivity, sensitivity at specificity, and accuracy are used for evaluation. The model is then trained on the training data, with validation performed on the validation data, and the best model is saved. Finally, the training history is saved to a file and plotted for visualization.

- **Accuracy for Created CNN Model**

Textify involves several steps, including data collection, preprocessing, feature extraction, model training, and testing. In the data collection phase, a large dataset of multiple frames is required. These are then preprocessed to isolate the lip region using various computer vision techniques. Next, features such as lip shape, motion, and appearance are extracted from the preprocessed images using machine learning-based techniques.

Once the features are extracted, a deep learning model is trained to map the extracted features to the corresponding text. The model can be trained to predict either words or phrase using various techniques such as convolutional neural networks (CNNs). Once the model is trained, it can be tested on a separate test dataset to evaluate its performance.

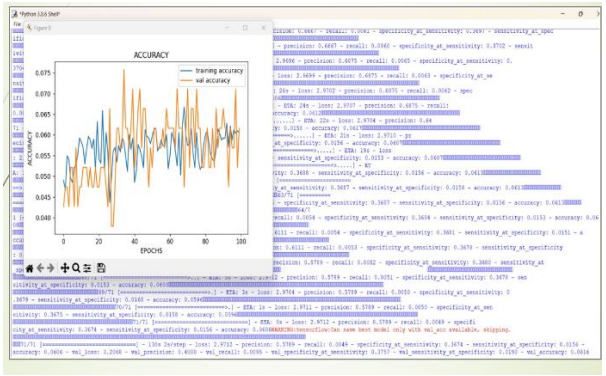
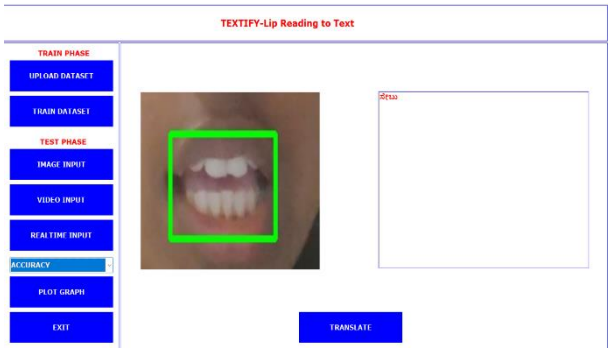


Diagram of Accuracy for Training Model

To improve the performance of the lip-reading system, various techniques such as data augmentation, transfer learning, and ensemble learning can be used. Data augmentation involves creating new training samples by applying various transformations such as rotations, scaling, and flipping to the existing training dataset. Transfer learning involves using pre-trained models on large datasets to extract features from lip images, which can then be fine-tuned on a smaller lip-reading dataset. Ensemble learning involves combining multiple models to improve the overall performance of the lip-reading system.

International Scientific Journal of Engineering and Management
Volume: 02 Issue: 05 | May – 2023 DOI: 10.55041/ISJEM00975 ISSN: 2583-6129
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata www.isjem.com



User Interface Diagram

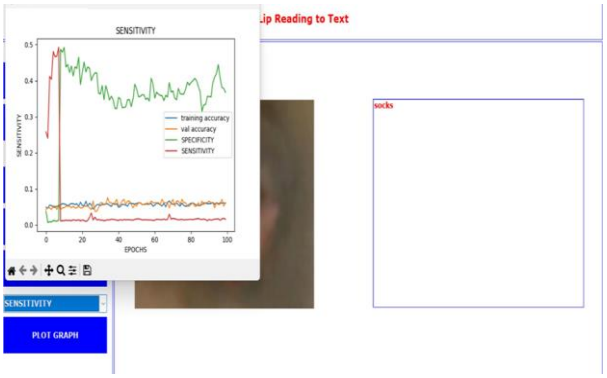
This user interface consists of three main components: a training section, a testing section, and a graph section. In the training phase, which consists of two phases such as load dataset and train dataset, the model is created. We will extract the dataset from load dataset and feed it into the system to be trained.

Graph for Specificity

The second phase is more crucial since it will test the model in three different ways, starting with images from the dataset, moving on to stored video, and concluding with actual video. If it worked in every scenario, the model is sound. The third section has four graphs, including ones for accuracy, specificity, sensitivity, and F1 score.

Graph for Accuracy

The lip-reading system's performance over time is shown by the accuracy graph in the text to lip reading conversion. As more data is fed into the system, it demonstrates how well the system can faithfully translate spoken words into text.



Graph for Sensitivity

A statistical metric called specificity quantifies how well a system can recognize unfavorable occurrences. Negative instances in the context of lip reading to text relate to situations where the system mistakenly interprets non-speaking visual cues as speech (such as facial expressions, mouth movements caused by eating, or yawning), yielding misleading positive findings. A lip-reading system's sensitivity might be tricky because lip motions can be quite unpredictable and influenced by things like illumination, camera angle, and speaker variability.

Graph for F1 Score

The F1 score reflects improved performance in accurately interpreting lip movements as spoken words. The translator was created since, although our system can recognize English, using it can be challenging for those who don't speak the language. Translator aids in translating into different languages for speakers who are comfortable.

- **Working of Live Video**

The system uses OpenCV and Keras libraries to detect the mouth from a video captured by a camera and classify it into two categories using a trained model. The system loads a pre-trained mouth cascade classifier xml file to detect the mouth in a grayscale image captured by the

camera. If the mouth is detected, it extracts the region of interest around the mouth and saves it to a temporary directory on the disk.

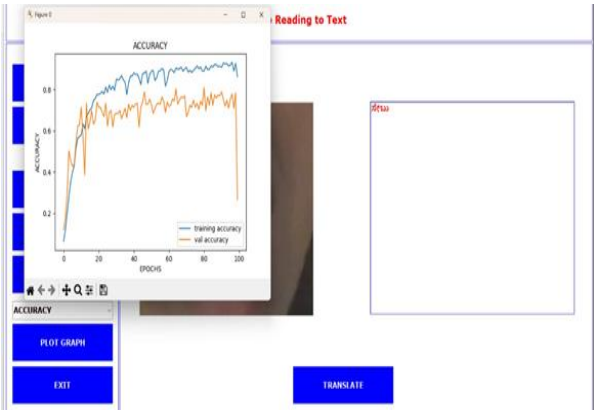
Live Video Capturing

The system then loads the pre-trained model using Keras and predicts the class of the image by mapping the predicted label to the actual label using a dictionary. Then displays a rectangle around the detected mouth region in the original color image and shows the resulting image in a window. The code continues to capture frames and detect the mouth until a predefined number of frames is reached or the user interrupts the program.

- **Translator**

A translator in lip reading to text is a software program that can recognize lip movements and convert them into text. The software uses computer vision and machine learning algorithms to analyze the movements of the lips and convert them into words or sentences. The lip movements are captured using a camera or webcam and the software then processes the video frames to identify the movements of the lips.

Once the lip movements are identified, the software uses machine learning algorithms to recognize the corresponding words or sentences. The recognized words or sentences are then converted into text and displayed on the screen or outputted to a text file. This type of software is commonly used by people who are deaf or hard of hearing to understand spoken language by reading the lips of the speaker. It can also be used in noisy environments where it may be difficult to hear the spoken words.



Translator in GUI

VII. CONCLUSION

We decided to create a system to assist with lip-reading and predicting speech. We reviewed different deep learning algorithms and datasets to improve automatic lip-reading. Our system uses a face detection model to extract faces and the Haar Cascade Facial Landmark detector to extract lips from the face image. We trained a CNN with batch normalization to classify words or phrases and extract information from image sequences.

VIII. REFERENCES

[1.] A Lip-Reading Model Using CNN with Batch Normalization published by Harexpe the Gupta, Dhruv Mittal Proceedings of 2018 Eleventh International Conference on Contemporary Computing (IC3), 2-4 August, 2018, Noida, India.

[2.] Saakshi Bhosale, Rohan Bait, Shivangi Jotshi, Rohan Bangera, Prof. Jinesh Melvin “An Application to Convert Lip Movement into Readable Text” in International Journal of Engineering Research & Technology (IJERT). ISSN: 2278-0181.

[3.] Souheil Fenghour, Daqing Chen in “Lip Reading Sentences Using Deep Learning with Only Visual Cues” published in November 9, 2020, accepted.November 18, 2020, date of publication November 26, 2020, date of current version December 11, 2020.

[4.] Ahsan Adeel, Mandar Gogate, Amir Hussain, and William M. Whitmer in “Lip-Reading Driven Deep Learning Approach for Speech Enhancement” IEEE TRANSACTIONS ON EMERGING TOPICS IN COMPUTATIONAL INTELLIGENCE.

[5.] H. Kulkarni and D. Kirange, "Artificial Intelligence: A Survey on Lip-Reading Techniques,"

2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT),2019,pp.1-5, Doi:10.1109/ICCCNT45670.2019.8944628.

[6.] T. Thein and K. M. San, "Lip movements recognition towards an automatic lip-reading system for Myanmar consonants," 2018 12th International Conference on Research Challenges in Information Science(RCIS),2018,pp.1-6, Doi:10.1109/RCIS.2018.8406660.

[7.] W. Nittaya, K. Wetchasit and K. Silanon, "Thai Lip-Reading CAI for Hearing Impairment Student," 2018 Seventh ICT International Student Project Conference (ICTISPC), 2018, pp. 1-4, Doi: 10.1109/ICT-ISPC.2018.8523956.

[8.] J. Wei, F. Yang, J. Zhang, R. Yu, M. Yu and J. Wang, "Three-Dimensional Joint Geometric-Physiologic Feature for Lip-Reading," 2018 IEEE 30th International Conference on Tools with Artificial Intelligence (ICTAI), 2018, pp. 1007-1012, Doi: 10.1109/ICTAI.2018.00155.

[9.] AKurniawan and S. Suyanto, "Syllable-Based Indonesian Lip-Reading Model," 2020 8th International Conference on Information and Communication Technology (ICoICT), 2020, pp. 1-6, Doi: 10.1109/ICoICT49345.2020.9166217.

[10.] Lu et al., "Lip Reading-Based User Authentication Through Acoustic Sensing on Smartphones," in IEEE/ACM Transactions on Networking, vol. 27, no. 1, pp. 447-460, Feb. 2019, Doi: 10.

ISSN: 2583-6129
www.isjem.com