DEEP NEURAL NETWORK FOR AUTOMATED LIP READING VISUAL SPEECH RECOGNITION

[1]Ms A.Sangeerani Devi, [2]Ramya.M, [3] Sri Ranjani.B, [4]Puranimaa.B

[1] Assistant Professor ,Department of Computer Science and Engineering , Sri Sai Ram Engineering College , Chennai- 600 044

[2,3,4] Department of Computer Science and Engineering , Sri Sai Ram Engineering College , Chennai- 600 044.

ABSTRACT

Lip Reading is a skill or act of recognizing a speaker's words by watching the lip movements. It is especially taught to deaf, so that they can communicate effectively. The art of lip reading can be used to help people by enhancing speech recognition in noisy areas, or by security forces in situations where it is necessary to identify a person's speech when the audio record is not available. Deep Neural Network is a good candidate for recognizing lip-articulated words. The goal of this project is to extract text words from a video input using visual lips tracking and Deep learning techniques.

I. INTRODUCTION

Mutism is the condition of total loss of speech. There are number of conditions that lead to muteness. It can also result from peripheral disorders such as laryngitis or vocal cord paralysis, endotracheal intubation, tracheostomy, or damage to the vocal cords or trachea from disease or trauma. Total laryngectomy is the treatment performed for advanced cancers of the larynx and hypopharynx, which is the surgical procedure in which the larynx, an organ essential for natural sound production is removed and includes separation of the airway from the mouth, nose and esophagus. Consequence of this surgery is associated with the loss of voice or normal verbal pronunciation affecting basic communication. Thus several diseases and various sites of pathology can produce mutism.

While considering hearing loss, Conductive hearing loss occurs when something blocks sound waves from reaching the inner ear. Ear infections, fluid build up behind the eardrum, holes in the eardrum, and problems with the middle ear bones can also cause deafness from conductive hearing loss. Sensorineural hearing loss is a condition caused by damage to inner ear or to the nerves that send sound to the brain. In rare cases, tumours are also a reason to cause conductive hearing loss – they block sound from getting into the inner ear. Resbycusis, or age-associated hearing loss, also has a genetic component. It is a condition which can make someone deaf over time as they age due to the slow decay of sensitive hair cells lining the inner ear. Even exposure to loud noise in certain areas of work such as construction machinery, amplified music, or heavy equipment can cause sensorineural hearing loss in people of all ages and forms the most common cause of hearing loss. Mutism and hearing loss affects the process of communication.

As communication facilitates the process of sharing information and knowledge, also helps people to develop relationships with others playing a vital role in human life. Therefore, the importance of communication cannot be underestimated. Thus this paper focuses on creating an aiding tool for deaf and dumb people.

Speech is the important mode of communication, for a person to express his/her feelings or to convey information. Speech is the natural way for communication. Human speech perception is well known to be a multimodal process.

Visual speech recognition is a process of understanding speech by observing only the lip movements without having access to the acoustic signal. Application of such system can be used in noisy environments as the visual signal is not affected by noise, it can also increase the performance of speech recognition systems and can be used whenever privacy is required in a public place.

The organisation of the paper is as follows. Section II explains the literature review used lip reading. Section III elaborates the proposed work and describes the results and implementation. Section IV concludes the paper.

II. LITERATURE REVIEW

This section presents the research work of some prominent authors in the same field and explains a short description of various techniques used.

Warunee Nittaya , Kaskaew Wetchasit and *Kittasil Silanon* (2018**) [1]** proposed a lip-reading computer assisted instruction (CAI) for hearing impairment students which consists of two units: the first unit is composed of the learning lesson. The student can learn the words in this module. The second unit is a multiple choice game. The student has to predict the word uttered by the speaker in lip-reading video and choose the correct answer choice.

Li Lu et al. (2018) [2] proposed a lip reading-based user authentication system, LipPass, which extracts unique behavioural characteristics of user's lip movements leveraging build-in audio devices on smartphones for user authentication using a deep learning-based method to extract efficient features from Doppler profiles, and employ Support Vector Machine and Support Vector Domain Description to construct binary classifiers and spoofer detectors for user identification and spoofer detection.

Cygert et al.(2018) [3] performed a analysis to evaluate to what extent visual data (i.e. lip reading) can enhance recognition accuracy in the multi-modal approach. In this analysis motion capture markers were placed on speaker's faces to obtain visual data by lips tracking during speaking. It includes testing of different parameterizations strategies and analysis of the accuracy of phonemes recognition in different experiments.

Joon Son Chung et al.(2017) [4] proposed a Watch, Listen, Attend and Spell(WLAS) network which learns to predict characters in sentences that were spoken, obtained from a video of a talking face, with or without audio. Discussed how the problems of audiovisual speech recognition (AVSR) and lip reading are closely linked. Mroueh *et al.* employed a feed-forward Deep Neural Networks (DNNs) that performed phoneme classification using a large non-public audio-visual dataset.

Yiting Li et al (2016) [5] proposed a system that processed dynamic feature is using convolution neural networks(CNN),the system is able to reduce the negative influence at the feature-extraction level. Dynamic feature represents difference image. A difference image is often used in object recognition, tracking and detection. In this way, the moving object is extracted clearly because the shade and colour of the moving object is distinct from the background.

III. PROPOSED WORK

This paper is concerned with automatic recognition of words which involves the conversion of the visual signals to text. Visual signals consists of words or phrases uttered by the speaker.

A. PROCESSING STEPS

Following steps are used to train a deep learning model for lip reading system.

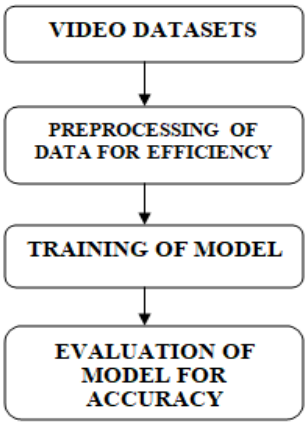

Fig 1: Deep Learning Model

B. DATA PREPROCESSING

The input video is converted into sequence of frames. It involves both face detection and lip detection. Pre-processing stage involves processing the data from video to obtain only the region of interest that speed up the training. The frames are saved as a grayscale image, thus reducing the size 3 times.

DETECTION AND TRACKING

An effective and robust object detection method using Haar feature-based classifiers proposed by Paul Viola and Michael Jones published in the paper *"Rapid Object Detection using a Boosted Cascade of Simple Features"* is used for face and mouth detection. Initially face is detected using frontal haar classifier. If face is not detected, then profile haar classifier is used for detection of profile face. If face is not detected during data preprocessing, video is skipped. After face detection, lip is detected using mouth haar classifier. Face and object detection is performed only for first frame. The remaining frames use object tracker.

MedianFlow Tracker offered by OpenCV library is used for tracking a specified object across multiple subsequent frames. Usually tracking algorithms are faster than detection algorithms and also tracking can help when detection fails.

FEATURE EXTRACTION

When extracting the mouth regions via the tracking method described above, grayscale images are converted and resized each frame to dimensions 24×32 pixels, which in total of 28 frames produces a tensor of size 28×24×32, with depth, height and width respectively. The values are populated in a NumPy 3D array and stored compressed on a hard disk which are the features used for model training.

C. MODEL TRAINING

The deep learning model refers to the model artefact that is created by the training process which can be used to get predictions on new data for which target is unknown. Deep Learning model is trained using Keras and Tensorflow.

CONVOLUTION

Convolution Layers comprises of a set of independent filters which are used for sliding by *N* pixels known as stride across the input image involves multiplying the values in the filter with the original pixel values of the image, thus computing element wise multiplications. We add the multiplication outputs to get the final integer which forms a single element of the output matrix. The final output matrix is known as Convolved Image, Activation Map or Feature Map.

BATCH NORMALIZATION

Batch Normalization is used to minimize the problem of covariance shift, means the change in the distribution of inputs to a unit as the training progresses, and therefore improve the training. So these inputs are normalized to have zero mean and unit variance, so that training can be drastically sped up.

POOLING LAYER

Max-Pooling operation is used, in which a filter convolves around each stride and produces an output which is the maximum number of that specific stride to the output matrix.It is used to reduce overfitting by providing an abstracted form of the representation. It also reduces the computational cost by reducing number of parameters to learn.

DROPOUT

Dropout is a technique used to reduce overfitting. It includes the process of setting the output of each hidden neuron with a certain

probability to zero. Thus dropped out neurons do not contribute to the forward pass and the Neural Network is forced to learn more robust features by finding an other activation path with conjunction of other neurons.

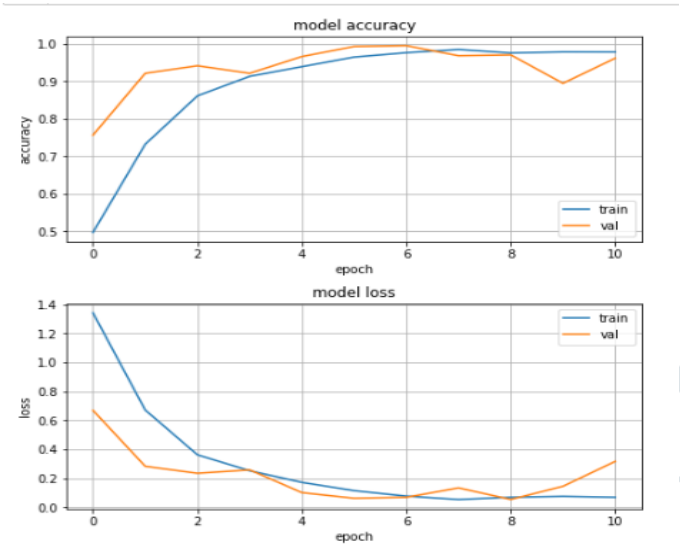

Fig 2 : Model accuracy and Model Loss

D. MODEL EVALUATION

Model Evaluation can be done using Confusion Matrix. It helps to find whether the model predicts the desired correct output. It provides a view of prediction for a large test dataset.

CONFUSION MATRIX

Confusion Matrix (CM) is tool for summarizing the performance of a classification algorithm. It can give a better idea of what classes the model predicts wrongly the most and for which words they are being confused. All values are normalized and the number in each individual cell represent show much is the given word being confused with another one (in percentage).

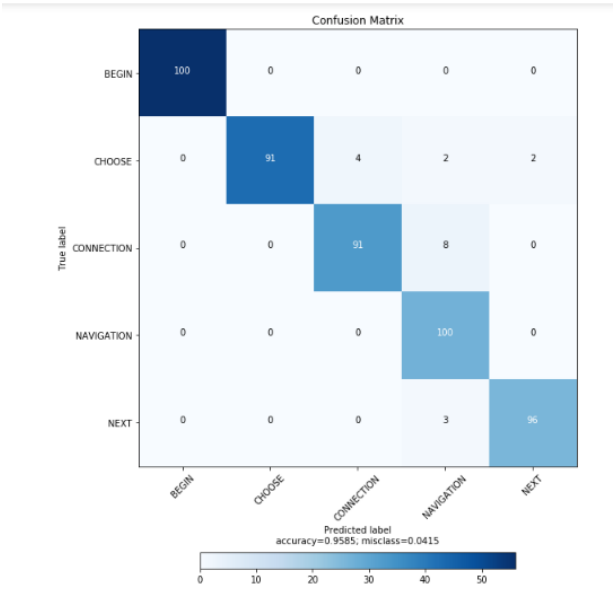

Fig 3 : Confusion Matrix

IV. CONCLUSION

Lip reading is a solution to understand or interpret speech visually especially by people with hearing difficulties and also an aiding tool for mute people to communicate. It helps, persons with a hearing impairment and people who cannot speak to remove obstacles to engage in social activities, without which communicating with others would be difficult.

Lip reading provides techniques for obtaining useful information in speech perception, efficient communication and language understanding, especially when the auditory speech is degraded. The benefits from visual speech cues have become the motivation for the significant interest in automatic lip-reading, which focus at improving automatic speech recognition by making use of informative visual features of a speaker's mouth region. Lip-reading technology usually includes the detection and localization of lip region, feature extraction and recognition method and fusion method, the classification of the mouth and the training the data sets.

REFERENCES

[1] Warunee Nittaya , Kaskaew Wetchasit and Kittasil Silanon “Thai Lip-Reading CAI for Hearing Impairment Student” in 2018 Seventh ICT International Student Project Conference (ICT-ISPC)

[2] Li Lu, Jiadi Yu, Yingying Chen, Hongbo Liu Yanmin Zhu, Yunfei Liu, Minglu Li "LipP ass: Lip Reading-based User Authentication on Smartphones Leveraging Acoustic Signals" in IEEE INFOCOM 2018 - IEEE Conference on Computer Communications

[3] S. Cygert, G. Szwoch, S. Zaporowski, A. Czyzewski "Vocalic segments classification assisted by mouth motion capture" in 2018 IEEE

[4] Joon Son Chung, Andrew Senior , Oriol Vinyals , Andrew Zisserman "Lip Reading Sentences in theWild " in 2017 IEEE Conference on Computer Vision and Pattern Recognition

[5] Yiting Li, Yuki Takashima, Tetsuya Takiguchi, Yasuo Ariki " Lip Reading Using a Dynamic Feature of Lip Images and Convolutional Neural Networks" in 2016 IEEE ICIS 2016, June 26-29, 2016, Okayama, Japan

[6] Chung, J. S.; Zisserman, "A. Lip Reading in the Wild " In Asian Conference on Computer Vision, 2016.

[7] Viola, P.; Jones, M. "Rapid object detection using a boosted cascade of simple features." In Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition.

[8] Hassan Akbari, Himani Arora, Liangliang Cao, Nima Mesgarani "LIP2AUDSPEC: SPEECH RECONSTRUCTION FROM SILENT LIP MOVEMENTS VIDEO" in 2018 IEEE

[9] Neeru Rathee "Investigating Back Propagation Neural Network for Lip Reading" in International Conference on Computing, Communication and Automation (ICCCA2016)

[10] Bor-Shing Lin, Yu-Hsien Yao , Ching-Feng Liu, Ching-Feng Lien And Bor-Shyh Lin "Development of Novel Lip-Reading Recognition Algorithm" in 2017 IEEE