WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Lip Reading using Computer Vision Techniques and Deep Learning Algorithms for Deaf and Dumb People

Lokesh M R [a], Arpitha C L[b], Hemalatha [b], Krupa K G [b]

[a] Professor, Department of computer science and Engineering, Vivekananda College of Engineering and Technology, Puttur, Dakshina Kannda Distict , Karnataka , India,

[b] UG Students, Department of computer science and Engineering, Vivekananda College of Engineering and Technology, Puttur, Dakshina Kannda Distict , Karnataka , India,

DOI: https://doi.org/10.55248/gengpi.5.0524.1130

A B S T R A C T

Lip reading to text for deaf and dumb people introduces a transformative solution aimed at enhancing communication for the deaf and dumb community through the development of a Lip Reading to Text Conversion system. Leveraging advanced computer vision techniques and deep learning algorithms, the system interprets and translates visual lip movements into real-time, contextually relevant text. Key functionalities include language support for diverse communication needs, adaptability to various lip shapes and accents, and a user-friendly interface accessible through wearable devices. The system prioritizes real-time processing, ensuring seamless and instantaneous conversion for natural and fluid communication. Extensive testing with diverse scenarios involving deaf and dumb individuals validates the system's accuracy and effectiveness. To addressing the unique challenges faced by the community, the Lip Reading to Text Conversion system stands as a promising assistive technology, breaking down communication barriers and fostering inclusivity for individuals with hearing and speech impairments.

Keywords: Lip-reading; Adaptability; Deep Learning; speech impairment; Text Conversion System; Expanding Language Support.

1. Introduction

Lip Reading, also known as speech reading, is a technique of understanding speech by visually interpreting the movements of the lips, face and tongue when normal sound is not available. The communication barriers faced by deaf and dumb individuals have been a longstanding challenge, hindering their ability to interact seamlessly with the hearing world. One promising technological solution involves the development of lip-reading systems that can convert visual information from lip movements into text. This paper presents an overview of the state-of-the-art lip-reading technologies and their application in converting lip movements into textual information for the benefit of the deaf and mute community.

In recent years the existing state of art proposes a secure and effective lip-reading system that can accurately detect lips movements, even when face masks are worn. The system utilizes radio frequency (RF) sensing and ultra-wideband (UWB) radar technology, which overcomes the challenges posed by traditional vision-based systems (Saeed, Umer, et al). This survey also provides comparisons of all the different components that make up automated lip-reading systems including the audio-visual databases, feature extraction, classification networks and classification schema(Fenghour, Souheil et al) .The proposed method consists of imaging from event data, face and facial feature points detection, and recognition using a Temporal Convolutional Network(Kanamaru, T en. al.). The improved Efficient-Ghost Net is used to perform lip spatial feature extraction, and then the extracted features are inputted to the GRU network to obtain the temporal features of the lip sequences, and finally for prediction (Zhang en. al.). To integrate facial expression features; Expression based feature and action unit-based feature into the lip-reading method (Shirakata, T en. al.). Deaf and hearing-impaired people use sign language as the main way of communication in everyday life. Sign language is a structured form of hand gestures and lips movements involving visual motions and signs, which is used as a communication system Ivanko en. al.).

The proposed lip-reading system employs advanced computer vision techniques, including deep learning algorithms, to accurately interpret lip movements and translate them into meaningful text. The system is trained on large datasets of diverse lip movements, encompassing various languages and speech patterns. The integration of neural networks allows the model to learn and adapt to different individuals' unique lip shapes and movements, enhancing the system's overall accuracy and usability. Furthermore, the system incorporates real-time processing capabilities, enabling instantaneous conversion of lip movements into text, thereby facilitating natural and fluid communication.

The development of user-friendly interfaces, such as mobile applications or wearable devices, ensures accessibility and ease of use for individuals with varying levels of technological proficiency. To evaluate the effectiveness of the lip-reading system, extensive testing has been conducted with deaf and dumb individuals in diverse communication scenarios. The results demonstrate significant improvements in communication efficiency and accuracy, highlighting the potential of lip-reading technology as a valuable tool for enhancing the quality of life for the deaf and dumb community. This project

introduces a lip-reading system designed to empower deaf and mute individuals by converting visual lip movements into text. The system employs sophisticated algorithms, beginning with the collection of diverse lip movement datasets.

The rest of the paper is organized as follows: First in Section II, the different real video databases used to train and test lip-reading systems for decoding the word are described; then in Section III, an overview of the different pre-processing aspects that make up lip-reading systems is given. This is followed by a comparison of the different CN network architectures used for feature extraction in Section IV, a system design is proposed in Section V, and a comparison of the different modeling is implemented g in Section VI. In Section VII, a result is given with accuracy graph and EF3-Architecture. Finally in Section VIII, concluding remarks are given along with suggestions for further research and a summary of current challenges faced in the domain of automated lip-reading.

2. Literature Survey

The work provides the realm of computer vision, the integration of event-based cameras with a local cross-channel interaction strategy has emerged as a promising approach. This strategy operates without dimensionality reduction, allowing for the preservation of intricate details within the visual data. Moreover, by incorporating facial expression features, the system gains a deeper understanding of human emotions and intentions. In parallel, advancements in deep learning techniques have revolutionized speech recognition systems, particularly in noisy environments.

The lip-reading as a new application by the event-based camera Kanamaru et.al.. This paper proposes an event camera-based lip-reading for isolated single sound recognition. The proposed method consists of imaging from event data, face and facial feature points detection, and recognition using a Temporal Convolutional Network. Furthermore, this paper proposes a method that combines the two modalities of the frame-based camera and an event-based camera. In order to evaluate the proposed method, the utterance scenes of 15 Japanese consonants from 20 speakers were collected using an event-based camera and a video camera and constructed an original dataset. Several experiments were conducted by generating images at multiple frame rates from an event-based camera. As a result, the highest recognition accuracy was obtained in the image of the event-based camera at 60 fps. Moreover, it was confirmed that combining two modalities yields higher recognition accuracy than a single modality.

The optimizes and improves GhostNet(Gaoyan Zhang and Yuanyao Lu.) , a lightweight network, and improves on it by proposing a more efficient Efficient-GhostNet, which achieves performance improvement while reducing the number of parameters through a local cross-channel interaction strategy, without dimensionality reduction. The improved Efficient-GhostNet is used to perform lip spatial feature extraction, and then the extracted features are inputted to the GRU network to obtain the temporal features of the lip sequences, and finally for prediction.

In integrate facial expression features (Shirakata and Saitoh), Expression based feature and action unit-based feature into the lip-reading method. Evaluation experiments are conducted with three public databases of OuluVS, CUAVE, and CENSREC-1-AV. As a result, it is confirmed that the recognition accuracy is improved by integrating the facial expression feature for all databases.

Deaf and hearing-impaired people use sign language as the main way of communication in everyday life (D. Ivanko et.al.[4].) Sign language is a structured form of hand gestures and lips movements involving visual motions and signs, which is used as a communication system. Since sign language includes not only hand gestures, but also lip movements that mimic vocalized pronunciation, it is of interest to investigate how accurately such a visual speech can be recognized by a lip-reading system, especially considering the fact that the visual speech of hearing-impaired people is often characterized with hyper-articulation, which should potentially facilitate its recognition.

Lip-reading technology play an important role to improve the recognition rate of the speech recognition systems on the noise environment and to help the disabled in hearing communicate with others (Pingxian et.al..) Because the shape of lip has obvious changes when somebody is pronouncing, it is difficult to directly detect the lip region; on the contrary, the nose shape cannot be able to have the obvious changes, and the face detection in opencv software has achieved considerable effect on detection face, so the paper has presented a method of detection lip; this method adopts the relative position of lip against to face and nose to detect the lip region. The experiments show that this method can quickly and efficiently extract the lip region, and improve the lip segmentation accurately.

The automated lip-reading approaches is presented in this paper with the main focus being on deep learning related methodologies which have proven to be more fruitful for both feature extraction and classification (Souheil Fenghour.) This survey also provides comparisons of all the different components that make up automated lip-reading systems including the audio-visual databases, feature extraction, classification networks and classification schemas. The main contributions and unique insights of this survey are: 1) A comparison of Convolutional Neural Networks with other neural network architectures for feature extraction; 2) A critical review on the advantages of Attention-Transformers and Temporal Convolutional Networks to Recurrent Neural Networks for classification; 3) A comparison of different classification schemas used for lip-reading including ASCII characters, phonemes and visemes, and 4) A review of the most up-to-date lip-reading systems up until early 2021.

3. Real-time System Lip Reading from Images Processing Modeling Architecture

The propose a system that will take in an image from dataset as input from the user. The work carried non inclined values and to help recognize the face in a better manner. The next step is to detect the region of interest that is the mouth and crop it out. This cropped ROI is to be passed to the convolution neural network (CNN) for further processing. Here the visual features are extracted, and the model is trained, based on which the spoken word with different styles are decoded to text.

- Design and implement a real-time system capable of lip reading from images.
- Train the networks on a lip-reading dataset, so that they can operate as lip reading systems.
- Evaluate the accuracy on simplified letter classification task.
- Explore the possibility to recognize words in trained image.
- Build a prototype that present capabilities of deep learning algorithm.

Convolutional Neural Network

The system architecture is designed based on working of a Convolutional Neural Network (CNN). A Convolutional Neural Network is a subset of machine learning algorithm which can take an input image, assign importance to various aspects/objects in the image and be able to differentiate one from the other. To recognize spoken words from lip movements, a CNN can be trained on a large dataset of video clips of people speaking different words and phrases. The CNN can be trained to learn the features of the lip movements and map them to corresponding words. The input to the CNN is typically a sequence of frames from a video, which are fed through a series of convolutional layers.

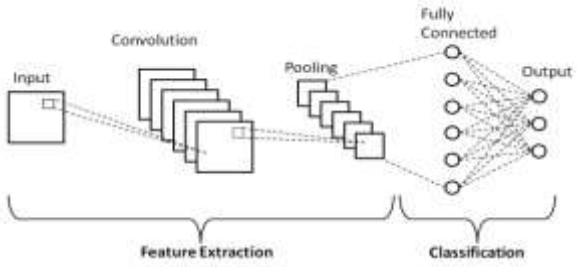


Fig. 1 - Architecture of CNN

The convolutional layers learn the spatial features of the lip movements in each frame, while the pooling layers down sample the feature maps, reducing the spatial dimensionality of the data. Finally, fully connected layers are used to classify the sequence of frames as a spoken word or phrase. One of the key challenges in lip reading and recognition is dealing with variations in lighting, camera angles, and speaker characteristics.

- Real Video

Here we take real-video as input for dataset creation. It is used to capture a video of the person and speaking their lip movements. The video is then processed using computer vision and machine learning techniques to recognize the spoken words.

- Face Detection and Cropping

Takes input as real time video, the system will detect the face in the video if it exists and for the simplicity of our project, we are assuming that our system will be able to detect faces with full frontal view only discarding the possibility of having partial or side views of a human. We make use of Haar features to be able to detect a face in the frame. After the detection of the face, the frames with no face will be discarded. The next step will be to be able to identify our Region of Interest (ROI) which is the lips and the mouth region in this case. It is to be identified with help of the Haar cascade classifier itself. Once the mouth region has been identified, cropped, identify the lip moment and for further processing and training of our system.

- Feature Extraction and Normalization

After the images are stored as an array, the features from the ROI need to be extracted. The spatial temporal features need to be extracted and fed into the CNN as an input for training of the model. Normalization of the image frames is necessary to avoid any irregularities in the dataset. For example, a person might take one second to pronounce a word, while another individual may take two seconds to pronounce the same word. Leaving such irregularities unattended may cause discrepancies in training and the results. So, we make use of normalization to be able to have an even training data.

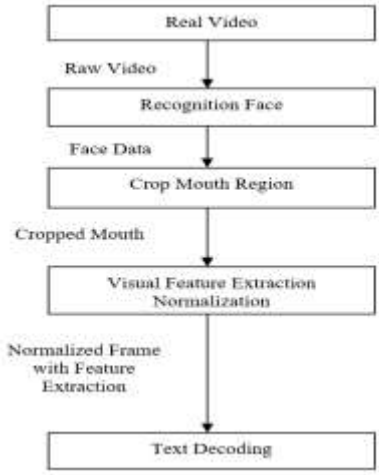


Fig. 2 - Real-time system lip reading from images Processing

- Text Classification and Decoding

Once the normalization is done, the data will be fed into the CNN for training and text decoding. The CNN learns on its own by having many epochs and passing the information learnt among the multiple hidden layers. The decoding will be done by matching the lip movement with the image data and the given dataset used for training, the words spoken will be predicted. The words which were spoken by the individual in the dataset.

4. Real-time System Lip Reading from Images Processing Modeling

The primary purpose of Convolution in case of a CNN is to extract features from the input image. Convolution preserves the spatial relationship between pixels by learning image features using small squares of input data. Every image is considered by a computer as a matrix of numbers (where numbers specify pixel color intensity). We create a Convolution Layers by convolving or" sliding" a filter (sometimes referred to as a kernel) by N pixels (also called stride) across the input image, where the current region below the filter is called receptive field, and 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 called Convolved Image, Activation Map or Feature Map. [13] As an example, considering an I as an input image, matrix K as a filter/kernel of size h × w, we can compute the Convolved Image $I * K$ as

$$(I * K)_{xy} = \sum_{i=1}^{h} \sum_{j=1}^{w} K_{ij} * I_{x+i-1,y+j-1}$$

The below figure shows diagrammatical overview of the above Formula and the result of applying convolution over an image.

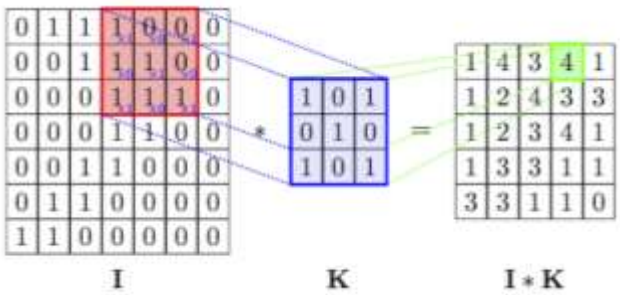


Fig. 2 - Diagrammatical overview

When applying convolutions filter the input shape and filter shapes are important. In case of images, we generally have either a 2D matrix of pixel values with dimensions height × width, representing a grayscale image, or a 3D matrix of shape height × width × 3, in case of colour image with red, green and blue channels. When handling a grayscale image with input shape of W × H, a kernel of shape k × k is used convolving in two directions (x, y) across the image, and 2D output matrix is obtained - Figure 2. With colour image, the input is W × H × L, (L = 3 channels), a kernel of shape k × k × L is used and the convolving is still performed in two directions (x, y), thus we again get a 2D matrix as an output.

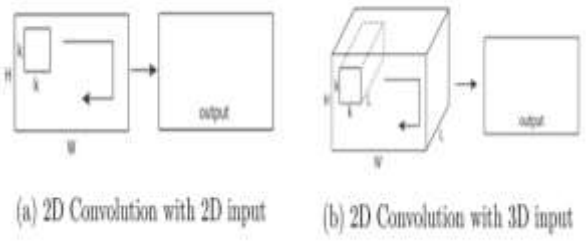


Fig. 3 - Image convolving

When building a ConvNet as shown in Figure 3, for classifying 3D objects, represented as a point cloud or 3D mesh, or when classifying a video, which can be imagined as an image sequence that can be stacked, we can represent the input layer by a 3D matrix. We can then use 3D convolutions (input of shape W × H × L, and a kernel of shape k × k × d, where d < L) were discussed in figure 4. Since kernel depth is smaller than the depth of the input volume, we convolve in three directions (x, y, z) and therefore the output will also be a 3D volume.

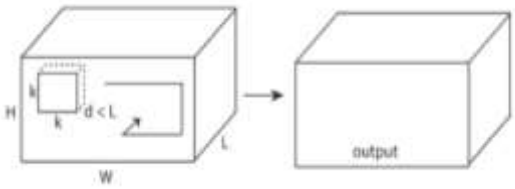


Fig. 4 - Video, 3D objects convolving

5. Real-time System Lip Reading from Images Processing Modeling Implementation

5.1 Dataset

One dataset commonly used for lip reading to text for deaf and dumb people is the LRW (Lip Reading in the Wild) dataset. LRW contains videos of people speaking words from a large vocabulary, recorded under various lighting conditions and camera angles.

Fig. 4 - Image Frame Sequence from a sample input image of the dataset

The dataset includes both the audio recordings of the spoken words and corresponding visual data of the speaker's lips as they articulate the words. This data is annotated with the corresponding text, allowing machine learning models to learn to transcribe spoken words into text based solely on visual lip movements. This technology can greatly benefit individuals who are deaf or hard of hearing, as it provides a way for them to understand spoken language through visual cues.

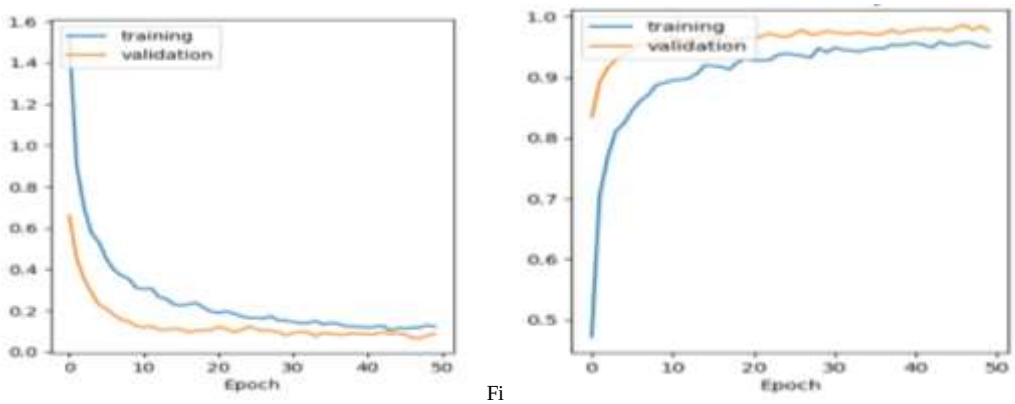
5.2 Result Discussion

When training the Neural Network we have used various architectures. First, we have built a model as is suggested in LRW, the EF-3 architecture in the below table, shows which is based on VGG-M model.

Table 1 – EF-3 Architecture

Layer Type	Output Shape	Parameter
Conv3D	(None,28,24,32,48)	1344
Batch Normalization	(None,28,24,32,48)	192
MaxPooling3D	(None,9,8,10,48)	0
Conv3D	(None,9,8,10,256)	332032
MaxPooling3D	(None,28,24,32,48)	0
Batch Normalization	(None,28,24,32,48)	1024
Conv3D	(None,3,2,3,512)	353956
Conv3D	(None,3,2,3,512)	7078400
Batch Normalization	(None,3,2,3,512)	2048
Flatten	(None,9216)	0
Dense	(None,20)	184340
Total params: 11,138,836		
Trainable Params: 11,137,204		
Non-trainable params: 1,632		

The accuracy graph of lip reading to text typically represents the performance of a lip reading system in converting lip movements into text. The x-axis of the graph usually represents different test samples or time frames, while the y-axis represents the accuracy of the system in correctly transcribing the lip movements into text.



Fi

Fig. 5 - (a) Model Loss; (b) Model Accuracy.

In Fig. 5 - (a) shows graph may show fluctuations in accuracy across different samples or time frames, which can be influenced by various factors such as the complexity of the spoken words, the clarity of the speaker's lip movements, background noise, and the quality of the lip reading algorithm.

In Fig. 5 – (b) shows overall trend of the graph can indicate the general performance of the lip-reading system. A rising trend may indicate improvements in accuracy over time, possibly due to algorithm refinements or increased training data. Conversely, a fluctuating or stagnant trend may suggest limitations or challenges faced by the system in accurately transcribing lip movements into text.

6. Comparison

In a highly intricate and specific use case, a groundbreaking approach has achieved remarkable results. With an accuracy of 87%, this method surpasses previous benchmarks by a significant margin. By combining various techniques, it outperforms deep learning models, achieving a substantial accuracy improvement of 62.1%. Furthermore, in the sequential prediction task of lip-reading, it boasts an impressive accuracy of 88.8%.

Table 2: Comparison of proposed system with existing system

Paper ID	Model accurately detect	Features, Optimize Weights	Particular Use
[1]	87%	Complex	particular use case
[2]	62.1%	Combined	Outperformed deep
[6]	Not Discussed	Dimensionality	Existing baseline
[4]	88.8%	Not Discussed	Not Discussed
[8]	Not Discussed	Redundancy	Not Discussed
[10]	65%	Not Discussed	Prediction task of lip-read
[12]	48.10%	Dynamic	Performs sequentially
Proposed Work	**58.20%**	**Dynamic**	**Strong performance specific case study**

This approach effectively minimizes redundancy, leading to a strong performance with an accuracy of 65%. In contrast, traditional methods relying solely on dynamic approaches achieve a comparatively lower accuracy of 48.10% than proposed work as 58.20%. This stark comparison highlights the effectiveness of the novel method in addressing the complexities of lip-reading tasks, showcasing its potential to significantly advance communication accessibility for the deaf and mute community.

7. Conclusion

The lip reading to text for deaf and mute individuals represents a significant step towards improving their communication and inclusion. By leveraging machine learning algorithms, this project can accurately convert lip movements into text, enabling these individuals to understand and respond to spoken language. This technology has the potential to enhance their quality of life by facilitating easier interaction in various settings, such as classrooms, workplaces, and social gatherings. Despite its current limitations, such as accuracy in complex environments and with different accents, further development and refinement of this project could lead to a transformative impact on the lives of deaf and mute individuals, empowering them to engage more fully with the world around them.

Future project work involves refining adaptability, expanding language support, and exploring additional modalities for comprehensive communication solutions.

References

Saeed, Umer, et al. "Extracting visual micro-Doppler signatures from human lips motion using UoG radar sensing data for hearing aid applications." IEEE Sensors Journal (2023).

Fenghour, Souheil et al. “Deep Learning-Based Automated Lip-Reading: A Survey.” IEEE Access 9 (2021): 121184-121205.

Kanamaru, T.; Arakane, T.; Saitoh, T. Isolated single sound lip-reading using a frame-based camera and event-based camera. Front. Artif. Intell. 2023, 5, 298. [Google Scholar] [CrossRef] [PubMed]

Zhang, Gaoyan and Yuanyao Lu. “Research on a Lip Reading Algorithm Based on Efficient-GhostNet.” Electronics (2023): n. pag.

Shirakata,T.Saitoh, T. Lip Reading using Facial Expression Features. Int. J. Comput. Vis. Signal Process. 2020, 10, 9–15. [Google Scholar]

Ivanko, Denis et al. “AUTOMATIC LIP-READING OF HEARING IMPAIRED PEOPLE.” The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences (2019): n. pag.

Pingxian, Yang et al. “Research on lip detection based on Opencv.” Proceedings 2011 International Conference on Transportation, Mechanical, and Electrical Engineering (TMEE) (2011): 1465-1468.

Rekik, Ahmed et al. “An adaptive approach for lip-reading using image and depth data.” Multimedia Tools and Applications 75 (2016): 8609-8636.

Shrestha, Karan et al. “Lip Reading using Neural Network and Deep learning.” (2019).

Chung, J.S and Zisserman, A. Lip Reading in the Wild. In Proceedings of the Asian Conference on Computer Vision (ACCV), Taipei, Taiwan, 20–24 November 2016. [Google Scholar]

K. Neeraja, K. Srinivas Rao and G. Praneeth, "Deep Learning based Lip Movement Technique for Mute," 2021 6th International Conference on Communication and Electronics Systems (ICCES), Coimbatre, India, 2021, pp. 1446-1450, doi: 10.1109/ICCES51350.2021.9489122.

C. Neti, G. Potamianos, J. Luettin, I. Matthews, H. Glotin, D. Vergyri, J. Sison, and A. Mashari, ‘‘Audio visual speech recognition,’’ IDIAP, Martigny, Switzerland, Tech. Rep., 2019.

T. Afouras, J. S. Chung, and A. Zisserman, ‘‘LRS3-TED: A large-scale dataset for visual speech recognition,’’ 2018, arXiv:1809.00496. Available:https://arxiv.org/abs/1809.00496.

G. Zhao, M. Barnard, and M. Pietikainen, ‘‘Lipreading with local spatiotemporal descriptors,’’ IEEE Trans. Multimedia, vol. 11, no. 7,pp. 1254–1265, Nov. 2018.

Q. M. Rizvi, ‘‘A review on face detection methods,’’ J. Manage. Develop. Inf. Technol., vol. 11, no. 2, pp. 1–11, 2011

Akbari et al., 2018. H. Akbari, H. Arora, L. Cao, N. Mesgarani, LIP2AUD-SPEC: Speech reconstruction from silent lip movements video. Proceedings of ICASSP2018, pp. 2516-2520.

Cheok et al., 2017. M. J. Cheok, Z. Omar, M. H. Jaward, A review of hand gesture and sign language recognition techniques. International Journal of Machine Learning and Cybernetics, vol. 10, no. 1, pp. 131-153.

Darrell et al., 1993. T. Darrell, A. Pentland, Space-time gestures, Proceedings on Computer Vision and Pattern Recognition, IEEE computer society conference, pp. 335-340.

Howell et al., 2016. D. Howell, S. Cox, B. Theobald, Visual units and confusion modelling for automatic lip-reading. In: Image and Vision Computing, vol. 51, pp. 1-12.

Petajan, 1984. E. D. Petajan, Automatic lipreading to enhance speech recognition. In: IEEE Communications Society Global Telecommunications Conference, Atlanta, USA.

Yang et al., 2010. R. Yang, S. Sarkar, B. Loeding, Handling movement epenthesis and hand segmentation ambiguities in continuous sign language recognition using nested dynamic programming, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 32, no.3, pp. 462-477.

Sumby, W.H. Visual Contribution to Speech Intelligibility in Noise. J. Acoust. Soc. Am. 1954, 26, 212–215. [CrossRef]

Aleksic, P.S.; Katsaggelos, A.K. Comparison of low- and high-level visual features for audio-visual continuous automatic speech recognition. In Proceedings of the 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing, Montreal, QC, Canada, 17–21 May 2004; p. V-917. [CrossRef]

Bakry A, Elgammal A (2013) Mkpls : Manifold kernel partial least squares for lipreading and speaker identification. In : International Conference on Computer Vision and Pattern Recognition, pp 684–691

Gogoi UR, Bhowmik MK, Saha P, Bhattacharjee D, De BK (2015) Facial mole detection : An approach towards face identification. Procedia Computer Science 46 :1546–1553

Lan Y, Theobald BJ, Harvey R (2012) View independent computer lip-reading. In : International Conference on Multimedia and Expo, pp 432–437

Lucey PJ, Sridharan S, Dean DB (2008) Continuous pose-invariant lipreading. In : Interspeech, Casual Productions, pp 2679–2682

Arnold, Thomas. A Method of Teaching the Deaf and Dumb Speech, Lip-reading, and Language. Vol. 43. Smith, Elder, 1881.

Mehta, Abhishek, Kamini Solanki, and Trupti Rathod. "Automatic translate real-time voice to sign language conversion for deaf and dumb people." Int. J. Eng. Res. Technol.(IJERT) 9 (2021): 174-177.

Kolb, Rachel. "Deaf People's" Subtile Art": Mabel Bell, Textual Deduction, and Cultural Representations of Lipreading." Journal of Literary & Cultural Disability Studies 15.2 (2021): 133-149.