iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 **Issue:** IX **Month of publication:** September 2021

DOI: https://doi.org/10.22214/ijraset.2021.38216

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue IX Sep 2021- Available at www.ijraset.com

Lip Reading: Delving into Deep Learning

Rishabh Nevatia[1], Nishant Dave[2], Meet Shah[3], Stevina Correia[4]
[1, 2, 3, 4]*Information Technology Department, Dwarkadas J. Sanghvi College of Engineering*

Abstract: ***Lip reading is the visual task of interpreting phrases from lip movements. While speech is one of the most common ways of communicating among individuals, understanding what a person wants to convey while having access only to their lip movements is till date a task that has not seen its paradigm. Various stages are involved in the process of automated lip reading, ranging from extraction of features to applying neural networks. This paper covers various deep learning approaches that are used for lip reading***

Keywords: Automatic Speech Recognition, Lip Reading, Neural Networks, Feature Extraction, Deep Learning

I. INTRODUCTION

Humans are social animals and to fulfil this social need they have always found ways to interact with each other in a plethora of forms. Communication has thus been given utmost importance for this reason. The interaction is based on several factors, such as body movements, expressions, tone, pitch, loudness of the voice, etc. The primitive humans communicated with signs and symbols and later invented languages for efficient communication with each other. The most common technique used is free speech wherein humans have learned a common tongue to understand each other. Over time, humans have become so used to this form of communication that several unique languages have been developed which has become a must for everyone to learn.

However, there are circumstances where speech becomes futile and other modes of communication have to be used. Lip reading or visual speech recognition (VSR) can supplement speech in such circumstances. Lip reading is executed by visual cues such as the movement of lips and then perceiving it as complete words, phrases or sentences. Basically, a visual stimulus is provided as an input and an interpreted word or sentence is retrieved as the output.

Most lip-reading systems have a streamlined flow of processing. The first step is the Input data, where the input format is of a video or photo. The lips of the speaker are the region of interest so the input data is pre-processed accordingly using different techniques to get the lips in the frame of reference, so that the insignificant parts of the input are already removed before feature extraction.

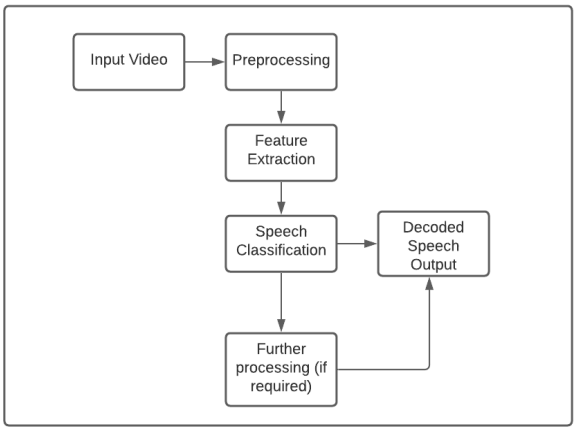


Figure 1 : Automated Lip Reading - General Architecture

Then the model moves on to feature extraction, in which the visual features from the input data are extracted. Finally, there is a classification step in which the extracted features are then used to predict the decoded text, along with a further processing of the output if required.

Lip reading has various applications. The usage of lip reading has been enormous and its primary usage is to recognize speech in noisy environments [1]. When the audio signal is corrupted it becomes difficult for audio-based speech recognition systems to correctly infer the text. In such scenarios lip reading is extremely helpful. The corrupted part of the audio signal can be successfully guessed from the lip movements and the context from the surrounding words [2]. For instance, the smart assistant for cars which uses lip reading to recognise phrases in a noise traffic environment [3].

It is also used in various departments such as in the entertainment industry for the dubbing of silent films. The application has also been extended to the security aspects such as for surveillance [4] and public safety domains. Using such techniques can help determine a long-distance conversation that could provide valuable insights.

Moreover, not all humans are fortunate enough to be able to speak freely. People can be muted either by birth or due to some unfortunate accidents they encounter in their lives. Many diseases such as inflammation, aphonia, hypothyroidism, laryngitis, vocal cord paralysis, etc. don't let people communicate freely due to partial or complete loss of voice. In such cases lip reading becomes a lifesaver and provides an alternate way for communication [5].

Lip reading is also useful in multi-talker speech recognition [6]. The recognition performs the task to separate speech in a multi-talker environment. From a sample of audio having multiple voices, the speech of each speaker can be separated using the lip movements of that corresponding speaker. This is very useful in legislature meetings.

Traditional approaches of automated lip reading involve pre-processing, feature extraction and recognition. They use machine learning methods such as SVM, HMM, KNN etc [7], [8], [9], [10]. These methods require careful feature selection and pre-processing to achieve their goals. Despite careful feature selection the quality of the features may be suboptimal. This results in lower accuracy for machine learning based lip-reading systems.

Recently there has been a lot of progress in the area of deep learning. Many techniques have been invented that can allow for more efficient feature extraction as well as recognition. Some examples are CNN, ANN and RNN. Deep learning techniques solve a given task in a complete start to finish manner. This saves the need to manually select the features. Deep learning has shown promising results in performing lip reading. In this paper, various reviews of deep learning techniques for lip reading have been presented.

II. RELATED WORKS

The work by J. S. Chung, A. Senior et al. is one of the newest research works in the lipreading domain, which focuses mainly on solving the apparent discrepancy at the word level because of homophones- phrases or words that have similar lip movements, such as ‘two’ and ‘too’ [11]. Although, such equivocacy can be solved to some level using an approach where the setting of the words after or before a part in a sentence is considered. This research’s aim was to recognize sentences and phrases being spoken by a face that was speaking, while having access to the audio. While previous research concentrated on recognizing a short amount of words or phrases, this research handled lip reading in a novel approach– using videos from the open world. Their key contributions are a WLAS model network, which is a model that learns to generate text for the videos consisting of the movement of the mouth by watching, learning, attending and spelling. It deploys a new learning strategy to accelerate training and to reduce overfitting. For the dataset, the LRS dataset for visual speech recognition was used, which consisted of over hundred thousand sentences in natural form, sourced from the Britain television. The model used, which was the WLAS model, had its training on the aforementioned dataset which had an exemplary result over the others. This approach presented that the information that had the audio unavailable in it, independently had a positive effect on speech recognition.

Y. Assael, B. Shillingford et al. showcased that the lipreading performance by the humans was not optimal or up to the standards required [12]. The accuracy reached by the hearing impaired was only up to 17% with a margin of 12% for a very diminutive subspace of thirty single syllable words and 21% with 11% margin positive or negative for thirty coalesced words.

Lipnet mimics the concept of ASR which stands for automatic speech recognition. Lipnet uses convolution neural networks which are stacked spatially over an image hence using the spatiotemporal convolutions (STCNN) [13]. It also uses an advanced form of recurrent neural network known as the Gated Recurrent Unit (GRU) [14]. Furthermore, the connectionist temporal classification loss (CTC), [15] which helps to compute the probability of a sequence based on discrete distribution of the token class, is used in lipnet. A CTC blank character is added to the sentences in the database at appropriate positions to convert it into a suitable format for CTC loss.

The lipnet uses the dataset known as the GRID [16] corpus dataset which contains words under six categories, which when combined produces sentences that makes it a sentence level dataset. The various permutations and combinations of the words tends to make the dataset containing 32746 usable videos from which corrupted as well as empty sets were removed. The sentence level accuracy that the Lipnet achieved was about 95.2% which exceeds the performance of the human lip readers and the previous attempts made. It also had an accuracy of 88.6% for the unseen speakers. This helped lipnet to become an end to end lip reading model which performed quite well on sentence level predictions.

TABLE I
Comparison of Existing Works

Paper Reference	Classification Algorithm / Method	Accuracy / Conclusion
Chung, J.S. et al. 2017 [11]	CNN, LSTM, MLP	A performance in Lip reading that surpassed a human, who was an outstanding lip reader. Both of them performed their predictions on videos that were provided by BBC television.
Wand, M. et al. 2016. [17]	SVM, LSTM	LSTM cells work better and increase accuracy instead of a normal NN cell.
Y. Assael, B. et al. [12]	STCNN, RNN, CTC Loss	94.2% accuracy.
Chung, J.S. and Zisserman, A. 2017 [18]	SyncNet + LSTM	94.1% accuracy for Short Phases
Thangthai, K. and Harvey, R. 2018[19]	Deep Autoencoder, CNN, RNN	6000 words vocabulary with 59% accuracy
Stafylakis, T. et al. 2018 [20]	Spatiotemporal ResNet with BiLSTM	11.92% MCR on LRW dataset
Petridis, S. et al. 2020 [21]	BLSTM	93.6% accuracy on OuluVS2 87.3% accuracy on CUAVE 66.3% accuracy on AVLETTERS 36.8% accuracy on AVLETTERS2

M. Wand, J. Koutnik et al. aimed at shortening and compacting the speech recognition architecture by using a neural network [17]. Feed forward and LSTM layers were used which would help to execute this aim for them. Features were extracted using PCA decomposition and HOG which was used for object recognition.

This paper proposed the use of a LSTM cell which stands for Long short term memory which eliminated the problem of the traditional neural network which caused gradients to explode or vanish. Three gates that were the input, output and the forget gate regulated the information flow in the LSTM. This paper used the GRID corpus dataset [16] which had a fixed organisational structure. Combined with the SVM classifier the overall accuracy for this paper amounted to 82 percent.

J. S. Chung and A. Zisserman Proposed a solution for audio-video synchronisation using two-stream ConvNet to learn embeddings from unlabelled data which is the end-to-end solution for audio-video synchronization [18]. The trained network can be extended to the problem of lip reading. The dataset used for the purpose of lip reading to measure the performance in this paper was BBC news videos.

The accuracy for the network along with LSTM amounts to the accuracy of 94.1% for the short phrases and 95.2% for the digits.

K. Thangthai and R. Harvey constructed a lip reading system with large vocabulary and attempted to solve the problem of speaker independence using an approach based-on Deep Neural Network Hidden Markov Models (DNN-HMMs) [19]. The dataset for this system was TCD-TIMIT audio visual corpus which had a 6000 word vocabulary and it attained an accuracy of 50% and managed to achieve 53.83% accuracy in speaker independent cross validation setting.

T. Stafylakis, M. H. Khan et al. proposed two sections, a frontend and a backend. The former consisted of a ResNet architecture with spatiotemporal inputs [20]. The architecture had eighteen layers. Training the network from scratch using a BiLSTM backend was unsuccessful. To solve this problem, a backend composed of temporal convolutional was used to train the network. The weights from this network were used to initialize the ResNet frontend for the BiLSTM backend which yielded good results. The BiLSTM backend used here was different from the traditional BiLSTM. Here, the two-directional outputs were concatenated at the last layer of the BiLSTM which obtained significantly better results. From various experiments, it was concluded that the BiLSTM backend works better than the temporal backend. It was also concluded that spatiotemporal ResNet works better than spatial

ResNets even when combined with more powerful backends. The reason being that temporal correlations provide necessary information for the task of lip reading. Experimentation on different configurations of ResNet yielded the result that a 18 layer ResNet Frontend with fully connected layers and BiLSTM backend achieved the misclassification rate of 17.01% which was the lowest. By incorporating optical flow with ResNet frontend and temporal convolution backend the lowest miss-classification rate achieved was 30.28%. Thus it can be concluded that spatiotemporal features work sufficiently well. After trying out different variations in the backend part and selecting the 3d ResNet as the frontend, the 2-layer BiLSTM architecture had the lowest misclassification rate which amounted to a mere 11.92%. The architecture consisted of dropouts, batch normalization, and average pooling layer to combine temporal information.

Overall, the lowest miss-classification rate achieved was 11.92% using 3d ResNets and BiLSTM.

S. Petridis, Y. Wang presented an approach for small-scale datasets [21]. In this approach, an end-to-end was presented which simultaneously learns the feature extraction and classification stages. The model used a BLSTM. One of the streams extracted features from the raw image and another stream extracted the local temporal difference of the mouth region. Both the streams are then combined via another BLSTM. For training, each stream is initialized independently and the bottleneck stream is initialized using Glorot initialization. On the OuluVS2 this model achieved a mean accuracy of 93.6%. The same model got a mean accuracy of 87.3% on the CUAVE dataset. For the AVLETTERS and AVLETTERS2, the mean accuracy reported was 66.3% and 36.8% respectively. Overall, this model beats the previously achieved cutting-edge accuracy obtained on these datasets.

III.DATASETS

The datasets span across being as small as images of 15 speakers speaking 10 phrases to being as exhaustive as 1000 sentences spoken by 34 speakers. Every research done for the lipreading domain used these datasets according to the need and availability of the datasets.

TABLE III
Summary of Datasets

Datasets	Information	Availability
Miracl-VC1 [22]	Ten words and ten sentences are spoken by a group of 15 people that comprised 5 men and 10 women. Every specimen of the dataset has a synchronized sequence of images that are colored and have depth, both of 640 by 480 pixels.	Yes
VIDTIMIT [23]	There are 43 people in this dataset whose video and audio are taken. Each narrator speaks ten sentences. Every speaker narrates the same words for the first two sentences and the rest are arbitrary, depending upon the speaker. There is a head rotation that is performed in a particular sequence which follows the order of moving the head first towards left and then towards right and then to the center. Later on the head moves up and then down and then to the center.	Yes

Grid Corpus[16]	GRID is a huge dataset that was made for the studies in speech observations which was used for social conduct studies using computer science. The recordings contain speech which were uttered by 34 speakers which had the demographic split of 18 males and 16 females who narrated 1000 sentences.	Yes
LRW[24]	Contains clips of speakers from BBC television.	Yes
AVICAR[25]	This corpus consists of videos of people being recorded in a car which was moving and there were cameras which captured the users lip movement from four different directions giving a separate view of the user.	Yes
AVLETTERS [26]	This dataset is composed of letters of the alphabet where each letter is repeated 3 times by each speaker. It was recorded by 10 speakers consisting of 5 males and 5 females. The resolution of the video captured was 376 × 288 pixels at 25fps.	Yes
AVLETTERS 2[27]	This dataset is composed of letters of the alphabet where each letter is repeated 7 times by each speaker. It was recorded by 5 speakers. The resolution of the video captured was 1920 × 1080 pixels at 50 fps.	Yes
AVTIMIT	The dataset contains 233 speakers of which 117 are male and 106 female. It has 4 h of AV data. The TIMIT sentences that were phonetically balanced were the sources from where the spoken utterances were chosen. The dataset consisted of speakers who spoke 20 sentences and it was made in such a way that every sentence was spoken at least 9 times. There was a common statement that was spoken by all the speakers in the dataset for uniformity. The recording of the video was done with the resolution of 720 x 480 pixels at 30 fps and the audio had 16 kHz.	No
CUAVE[28]	The CUAVE corpus consists of the speakers who speak about the digits separately as well as in a combined manner. It amounts up to 7,000. CUAVE is a speaker dependent corpus. The people in the dataset spoke fifty independent digits in a normal manner as well as thirty independent digits while making a sideways or forward-backward motion, or tilting the head.	Yes
IBMIH	It consisted of infrared videos over the ones captured under natural conditions such as vision of what a human eye would see. The speaker was given a pair of headphones, and a microphone along with a video camera that had infrared vision, having it placed in front of the mouth. The main focus was on the chin-mouth area, and only that part was captured in the frame. The corpus consists of seventy-nine speakers, who spoke digit strings in a continuous manner and another hundred and thirteen speakers who were	No

IV. CONCLUSIONS

A variety of deep learning techniques have been reviewed that are used in video speech recognition (VSR) or better known as lip reading. Different techniques and methodologies are used and various concepts are applied in this field which has progressed over the years such as sequence models and CTC. Various datasets have been used which have different combinations of words and letters, which started off as simple digits and alphabets to phrases, and then gradually shifted to form complete sentences over the years.

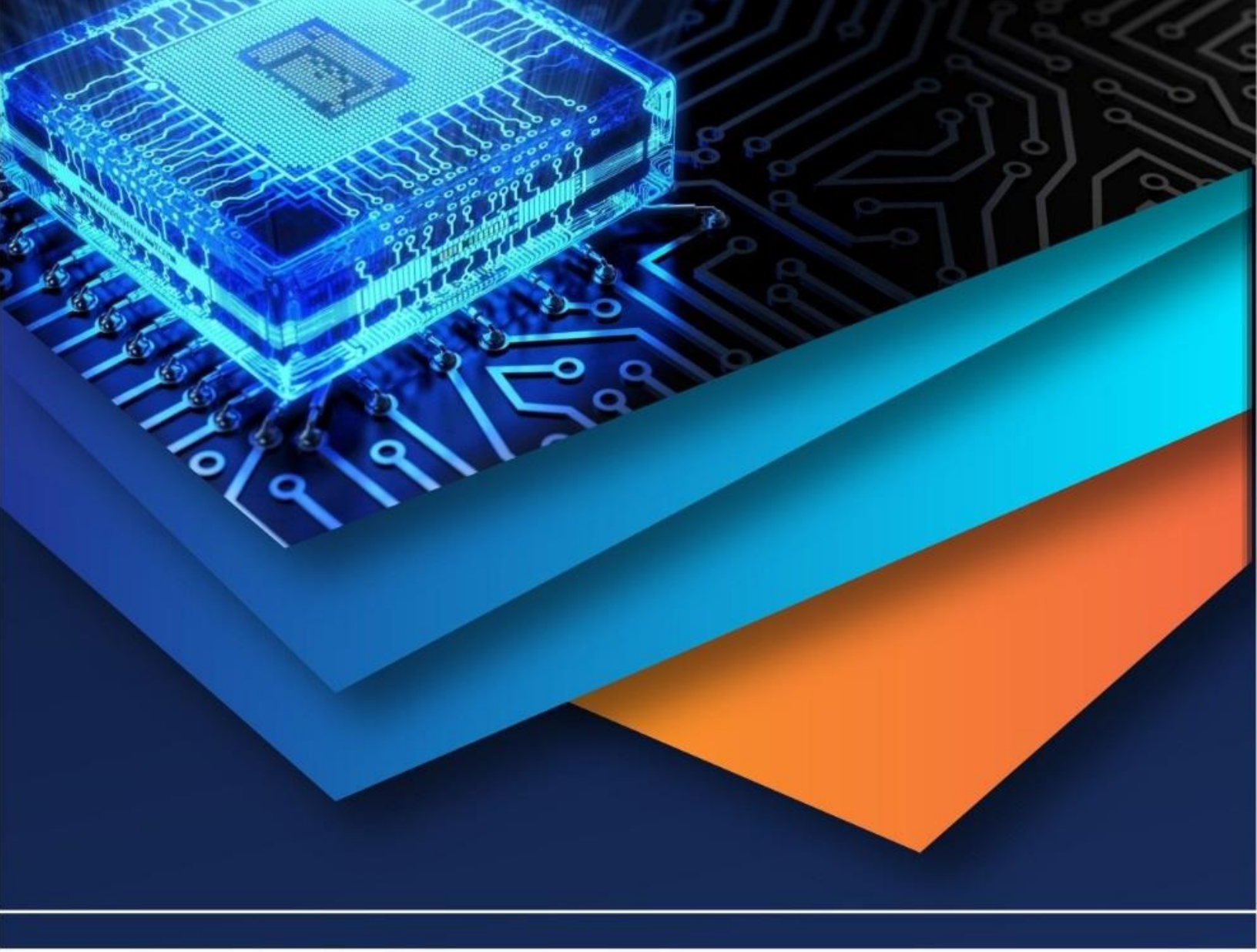
Increasing complexity of recognized phrases came along with challenges such as homophones. Whereas humans could easily distinguish between the applications of similar sounding words, imparting this sense of logic to a machine learning model was a daunting task.

Considering the recent development in the lip-reading domain, the artificial intelligence models with their novel approaches have been observed surpassing even the professional lip readers. Limitations such as homophones have been addressed and many different techniques have been combined to obtain promising results.

REFERENCES

[1] K. Palecek and J. Chaloupka, “Audio-visual speech recognition in noisy audio environments,” Jul. 2013, Accessed: Aug. 26, 2021. [Online]. Available: http://dx.doi.org/10.1109/tsp.2013.6613979.
[2] S. Petridis, T. Stafylakis, P. Ma, F. Cai, G. Tzimiropoulos, and M. Pantic, “End-to-End Audiovisual Speech Recognition,” Apr. 2018, Accessed: Aug. 26, 2021. [Online]. Available: http://dx.doi.org/10.1109/icassp.2018.8461326.
[3] R. Navarathna, P. Lucey, D. Dean, C. Fookes, and S. Sridharan, “Lip detection for audio-visual speech recognition in-car environment,” May 2010, Accessed: Aug. 26, 2021. [Online]. Available: http://dx.doi.org/10.1109/isspa.2010.5605429.
[4] L. Rothkrantz, “Lip-reading by surveillance cameras,” May 2017, Accessed: Aug. 26, 2021. [Online]. Available: http://dx.doi.org/10.1109/scsp.2017.7973348.
[5] B.-S. Lin, Y.-H. Yao, C.-F. Liu, C.-F. Lien, and B.-S. Lin, “Development of Novel Lip-Reading Recognition Algorithm,” IEEE Access, vol. 5, pp. 794–801, 2017, doi: 10.1109/access.2017.2649838.
[6] S. Rennie, J. Hershey, and P. Olsen, “Single-Channel Multitalker Speech Recognition,” IEEE Signal Processing Magazine, vol. 27, no. 6, Nov. 2010, doi: 10.1109/msp.2010.938081.
[7] G. Pomianos, C. Neti, G. Gravier, A. Garg, and A. W. Senior, “Recent advances in the automatic recognition of audiovisual speech,” Proceedings of the IEEE, vol. 91, no. 9, pp. 1306–1326, Sep. 2003, doi: 10.1109/jproc.2003.817150.
[8] T. Zhang, L. He, X. Li, and G. Feng, “Efficient End-to-End Sentence-Level Lipreading with Temporal Convolutional Networks,” Applied Sciences, vol. 11, no. 15, p. 6975, Jul. 2021, doi: 10.3390/app11156975.
[9] S. Dupont and J. Luettin, “Audio-visual speech modeling for continuous speech recognition,” IEEE Transactions on Multimedia, vol. 2, no. 3, pp. 141–151, 2000, doi: 10.1109/6046.865479.
[10] Z. Zhou, G. Zhao, X. Hong, and M. Pietikäinen, “A review of recent advances in visual speech decoding,” Image and Vision Computing, vol. 32, no. 9, pp. 590–605, Sep. 2014, doi: 10.1016/j.imavis.2014.06.004 .
[11] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, “Lip Reading Sentences in the Wild,” Jul. 2017, Accessed: Aug. 26, 2021. [Online]. Available: http://dx.doi.org/10.1109/cvpr.2017.367 .
[12] Y. Assael, B. Shillingford, S. Whiteson, and N. Freitas, “LipNet: End-to-End Sentence-level Lipreading,” 2016, [Online]. Available: https://arxiv.org/abs/1611.01599v2.
[13] Z. He, C.-Y. Chow, and J.-D. Zhang, “STCNN: A Spatio-Temporal Convolutional Neural Network for Long-Term Traffic Prediction,” Jun. 2019, Accessed: Aug. 26, 2021. [Online]. Available: http://dx.doi.org/10.1109/mdm.2019.00-53.
[14] K. Cho et al., “Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation,” 2014, Accessed: Aug. 26, 2021. [Online]. Available: http://dx.doi.org/10.3115/v1/d14-1179.
[15] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, “Connectionist temporal classification,” 2006, Accessed: Aug. 26, 2021. [Online]. Available: http://dx.doi.org/10.1145/1143844.1143891.
[16] J. Barker, M. Cooke, S. Cunningham, and X. Shao, “The GRID audiovisual sentence corpus,” University of Sheffield, 2013. http://spandh.dcs.shef.ac.uk/gridcorpus/.
[17] M. Wand, J. Koutnik, and J. Schmidhuber, “Lipreading with Long Short-Term Memory,” 2016, [Online]. Available: https://arxiv.org/abs/1601.08188.
[18] J. S. Chung and A. Zisserman, “Out of Time: Automated Lip Sync in the Wild,” in Computer Vision – ACCV 2016 Workshops, Cham: Springer International Publishing, 2017, pp. 251–263.
[19] K. Thangthai and R. Harvey, “Building Large-vocabulary Speaker-independent Lipreading Systems,” Sep. 2018, Accessed: Aug. 26, 2021. [Online]. Available: http://dx.doi.org/10.21437/interspeech.2018-2112.
[20] T. Stafylakis, M. H. Khan, and G. Tzimiropoulos, “Pushing the boundaries of audiovisual word recognition using Residual Networks and LSTMs,” Computer Vision and Image Understanding, vol. 176–177, pp. 22–32, Nov. 2018, doi: 10.1016/j.cviu.2018.10.003.
[21] S. Petridis, Y. Wang, P. Ma, Z. Li, and M. Pantic, “End-to-end visual speech recognition for small-scale datasets,” Pattern Recognition Letters, vol. 131, pp. 421–427, Mar. 2020, doi: 10.1016/j.patrec.2020.01.022.
[22] A. Ben-Hamadou, “MIRACL-VC1 - Achraf Ben-Hamadou,” Achraf Ben-Hamadou. https://sites.google.com/site/achrafbenhamadou/-datasets/miracl-vc1
[23] C. Sanderson, “Conrad Sanderson - VidTIMIT dataset,” Conrad Sanderson. https://conradsanderson.id.au/vidtimit/
[24] “Lip Reading in the Wild (LRW) dataset.” https://www.robots.ox.ac.uk/~vgg/data/lip_reading/lrw1.html (accessed Aug. 27, 2021).

[25] "AVICAR Project: Audio-Visual Speech Recognition at the University of Illinois at Urbana-Champaign." http://www.isle.illinois.edu/sst/AVICAR/
[26] "AV Letters Database." http://www.ee.surrey.ac.uk/Projects/LILiR/datasets/avletters1/index.html
[27] S. Cox, R. Harvey, Y. Lan, J. Newman, and B. Theobald, "CiteSeerX — The challenge of multispeaker lip-reading." http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.399.9242
[28] Patterson, Gurbuz, Tufekci, and Gowdy, "CUAVE: a new audio-visual database for multimodal human-computer interface research," 2002, [Online]. Available: http://dx.doi.org/10.1109/icassp.2002.1006168
[29] Guoying Zhao, M. Barnard, and M. Pietikainen, "Lipreading With Local Spatiotemporal Descriptors," IEEE Transactions on Multimedia, vol. 11, no. 7, pp. 1254–1265, Nov. 2009, doi: 10.1109/tmm.2009.2030637.
[30] "The XM2VTSDatabase." http://www.ee.surrey.ac.uk/CVSSP/xm2vtsdb/
[31] N. Harte and E. Gillen, "TCD-TIMIT: An Audio-Visual Corpus of Continuous Speech," IEEE Transactions on Multimedia, vol. 17, no. 5, pp. 603–615, May 2015, doi: 10.1109/tmm.2015.2407694.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY