Existing Work in the Field of Video to Text Conversion

Gaurav Boob [1], Somya Maheshwari [1], Abhishek Jain [1], Dr. R Sreemathy [1]
[1] Department of Electronics, SCTR's Pune Institute of Computer Technology, Pune, India
gauravboob5@gmail.com, som.maheshwari2002@gmail.com, abhishekjainindore24@gmail.com

Abstract **This survey offers evaluations and contrasts between the various elements comprising automated lip-reading systems. These elements encompass audio-visual databases, methods for extracting features, classification networks, and classification frameworks. The objective of the paper is to provide a more detailed analysis of areas where significant gaps or weaknesses exist in research on shaders. The main objectives of this survey paper are: 1) A thorough evaluation of the benefits associated with Attention-Transformers and Temporal Convolutional Networks (TCNs), Long short term memory (LSTM), Gated recurrent unit (GRU). 2) A review of the most current lip-reading systems available as of early 2023.**

I INTRODUCTION

Lipreading is the task of decrypting text from the movement of a speaker's mouth. Traditional approaches separated the problem into two stages: designing or learning visual features, and vaticination. In the realm of lip-reading classification, the backend systems are tailored to anticipate sequential speech components like words or sentences. To do this, systems often use sequences of processing networks such as Recurrent Neural NetworksRNNs which can be described in terms of Long Short Term Memory NetworksTMLS or Gated Recurrent UnitsGRUs. Besides RNNs, lip-reading backends also use alternative classification networks like Attention-based Transformers (ATT) and Temporal Convolution Networks (TCN).

The initial approaches for automating lip-reading primarily relied on non-deep learning techniques including methods like Markov Models.

This survey paper concentrates on lip-reading architectures for classification, with a particular emphasis on the advantages of Attention-Transformers and TCNs over RNNs. It offers distinctive insights by delving deeper into the comparative advantages of alternative frontend networks like feedforward neural networks and autoencoders, as opposed to the conventional CNNs. Additionally, it provides a comprehensive comparison of various classification schemas commonly employed in lip-reading. Importantly, the paper encompasses the latest approaches up to the late stages of 2022 and early 2023, ensuring the inclusion of the most current developments in the field.

Lip-reading systems typically adhere to a structured framework that encompasses several stages.

FIGURE 1. FRAMEWORK FOR AUTOMATED LIP READING

These stages include an initial preprocessing step, a frontend responsible for feature extraction, and a backend for classification. The automated lip-reading process is delineated in Figure 1.

II DATASETS

The design and development of lip-reading systems has inevitably been influenced by accessible data as a data-driven approach. The data should ideally be vocabulary-rich, with variations in stance and illumination. Large data corpuses such as LRS2 [1], LILiR Twotalk Corpus [2] and LRW [3] have been constructed from hours of BBC, TED-X, and YouTube streaming. Table 1 lists some of the main audio-visual datasets that have been utilized for lip-reading. These corpuses are made up of hundreds of recordings of humans saying phrases using thousands of different words.

In addition to the type of speech segments, another advancement of lip-reading data corpuses is the capacity to train lip-reading algorithms to classify speech from people speaking at varied angles (profile views), as opposed to frontally facing the cameras (frontal views). Furthermore, datasets such as LRW [3], LRS2 [1] and LRS3 [4] have moved on to collecting videos from multiple speakers rather than individual speakers, as one of the challenges facing the success of automated lip-reading systems is the inability to generalize to different people - particularly unseen speakers who did not appear in the training phase.

In addition to the type of speech segments, another advancement of lip-reading data corpuses is the capacity to train lip-reading algorithms to classify speech from people speaking at varied angles (profile views), as opposed to frontally facing the cameras (frontal views). Furthermore, datasets such as LRW [40], LRS2 and LRS3 have moved on to collecting videos from multiple speakers rather than individual speakers, as one of the challenges facing the success of automated lip-reading systems is the inability to generalize to different people - particularly unseen speakers who did not appear in the training phase.

In real-life circumstances, people may be speaking at various distances from cameras, in various lighting conditions, and with cameras of differing quality. These factors can cause considerable changes in video resolution. To ensure that audio-visual models can properly handle these issues, datasets have included movies of varying resolutions. The LRW [3] dataset, for example, is intended to imitate the variety observed in real-world lip reading events. It comprises films with resolutions ranging from

TABLE 1: DATASETS

Dataset Name	Number of Classes	Number of Videos	Segment
LRW	51	500,000 +	Sentences
GRID	51	33	Sentences
OuluVS2	6	80	Words
LRS2	N/A	100,000+	Words
LRS3	N/A	3,000,000+	Sentences
LRW - 1000	51	1000	Sentences
GRID - 480	51	480	Sentences
MVLRS	N/A	41,222	Sentences
IBMViaVoice	290	24,325	Sentences

high-definition to lower-quality visuals, akin to what you could see in surveillance footage or recordings shot with various sorts of equipment. Researchers hope to improve the resilience and adaptability of their models by integrating movies with varied resolutions. Models trained on such data learn to recognise and understand lip movements and speech cues over a wide range of visual features. This is critical for practical lip reading applications such as voice recognition, accessibility technology, and spying.

III. LITERATURE REVIEW

The study done by Joon Son Chung et al. in [1] compared the performance of the 'Watch, Listen, Attend and Spell' (WLAS) network with the LRS dataset, which demonstrated that visual cues can significantly improve speech recognition in combination with audio data. Furthermore, the WLAS model was able to outperform prior benchmarks, even outperforming a professional lip reader in BBC videos.

In [2] LipNet, an end-to-end trainable model for sentence-level lipreading, is presented in this research. Spatial-temporal convolutions, a recurrent network, and the connectionist temporal classification loss are used. The model converts a variable-length video frame sequence to text, capturing temporal context in lipreading and allowing sentence-level sequence prediction. The model is tested using the GRID corpus, which is a dataset for sentence-level lipreading. The GRID corpus is made up of videos of speakers saying sentences containing various words and phrases. It offers a difficult and realistic testbed for evaluating lipreading models. LipNet performs admirably on the GRID corpus, achieving 95.2% accuracy in sentence-level lipreading. This outperforms skilled human lip readers and exceeds previous state-of-the-art word-level accuracy. In lipreading, the model highlights the necessity of spatiotemporal feature extraction and efficient temporal aggregation. It also emphasizes the value of end-to-end models for applications like quiet dictation and audiovisual speech recognition. The report makes no mention of any potential biases or limits of employing a character-level approach for lipreading, such as the effect of homophones or similar lip movements for distinct phonemes.

A thorough examination of various approaches to improve lip-reading system performance, such as data augmentation, temporal modeling, self-dissecting methods, and word boundary indicators was done [3]. The outcome of the analysis was highly encouraging, as the combination of these approaches resulted in an increase in accuracy of 93.4% compared to the current performance on the LRT dataset (4.6%). These findings demonstrate the potential for incorporating these methods to significantly improve the accuracy and effectiveness of automated lip reading systems.

A 3D convolutional vision transformer (3DCvT)-based lip reading system was introduced in [4] that combines a vision transformer with 3D convolution to extract spatio-temporal characteristics from continuous images. Following that, the collected features are passed to a Bidirectional Gated Recurrent Unit (BiGRU) for sequence modeling. The approach is tested using the LRW and LRW-1000 large-scale lip reading datasets. On the LRW dataset, the best accuracy is 88.5%, whereas on the LRW-1000 dataset, it is 57.5%. Due to the intricacy of the Chinese language, the accuracy of the Chinese lip reading dataset LRW-1000 is relatively poor. To extract more robust spatio-temporal features from continuous images, the suggested 3D convolutional vision transformer (3DCvT) can be improved and optimized.

This research [6] methodology involved a thorough examination of various approaches to improve lip-reading system performance, such as data augmentation, temporal modeling, self-dissecting methods, and word boundary indicators. The outcome of the analysis was highly encouraging, as the combination of these approaches resulted in an increase in accuracy of 93.4% compared to the current performance on the LRT dataset (4.6%). These findings demonstrate the potential for incorporating these methods to significantly improve the accuracy and effectiveness of automated lip reading systems.

Xing Zhao et al. focused on the implementation of mutual information constraints on both local and global lip-reading features [5]. This novel approach was assessed using two large-scale benchmark datasets. The results were highly promising, as it led to a new level of performance on both benchmarks. This demonstrates the effectiveness of mutual information limitations in improving lip-reading accuracy and capabilities, which is a major step forward in the field. From a visual point of view, there are still issues with speaker dependency, especially when attempting to use lip-reading on individuals who were not included in the training dataset. Additionally, generic lip-reading systems must be able to handle videos with varying spatial resolution and frame rates, which contain varying amounts of temporal data.

A deep learning approach for audiovisual speech recognition was employed in [7], though the dataset used was not specifically specified. The results of the study were remarkable, as they revealed a significant improvement in performance. Specifically, the word error rate of the ASR system decreased by 6.59% when transcribing spoken language, and the lip reading model reached an impressive 95% accuracy rate, demonstrating the capability of deep learning methods to effectively combine audio and visual signals for speech recognition.

For the first time, the research [8] introduces an

attention-based pooling approach to aggregate visual speech representations and employs sub-word units for lip reading. It also introduces a Visual Speech Detection (VSD) model trained on top of the lip reading network. The suggested models are trained and evaluated using the LRS2, LRS3, and AVA-ActiveSpeaker benchmarks. When trained on public datasets, it obtains state-of-the-art performance on the LRS2 and LRS3 benchmarks, and even outperforms models trained on large-scale industrial datasets with an order of magnitude less data. On the LRS2 dataset, the top model obtains a word error rate of 22.6%, dramatically narrowing the performance gap between lip reading and automatic speech recognition. Furthermore, in the AVA-ActiveSpeaker test, the VSD model beats several current audio-visual techniques and excels all visual-only baselines. The Visual Speech Detection (VSD) model built on top of the lip reading system provides cutting-edge results, exceeding visual-only baselines and even numerous audio-visual techniques.

In this research [9], a novel Alternating Spatiotemporal and Spatial Convolutions (ALSOS) module was introduced to the methodology, combining spatiotemporal convolutions with spatial convolutions to enhance lip-reading performance. The module was integrated with both a Greek and a LRW-500 language dataset, and the results were promising, with the ALSOS module significantly improving lip-reading system performance, particularly in terms of spoken word accuracy. This integration highlights the potential of the ALSOS module as a valuable component of lip-reading technology that can improve results in a variety of language datasets.

The idea of using deep 3D Convolutional Neural Networks (CNNs) as the front-end for visual feature extraction in word-level lipreading was introduced in the paper composed by Xinshuo Weng and Kris Kitani [10]. The authors specifically replace the shallow 3D CNNs + deep 2D CNNs front-end with a two-stream I3D network composed of grayscale video and optical flow streams. Different combinations of front-end and back-end modules are evaluated using the LRW dataset. Pre-training on large-scale image and video datasets, such as ImageNet and Kinetics, is performed to improve classification accuracy. The LRW dataset, which contains short video clips extracted from BBC TV broadcasts, is used for training and evaluation. The dataset consists of 488,766 training videos, 25,000 validation videos, and 25,000 testing videos. The two-stream I3D front-end with a Bi-LSTM back-end achieves an absolute improvement of 5.3% over the previous state-of-the-art on the LRW dataset

A combination of HPConv (Hierarchical Pyramidal Convolution) and Self-attention (Self-Attention) methods in a new way was accomplished in the recent study. The study used the lip-reading in-wild dataset (LRW). The results were remarkable, with the proposed method achieving an impressive 86.83% accuracy rate, which is a significant improvement from the current lip-reading system (1.53%). These results demonstrate the power of HPConV and Self-Attention to significantly improve the accuracy and performance of lip reading technology, demonstrating its potential for progress in the field.

In order to achieve its goals, the study [14] used a combination of Computer Vision (CV) and CNN (Deep Convolutional Network) models to transcribe spoken sentences into text. To test this method, the researchers used the GRID (Audiovisual Language Interpreter Interpreter) dataset, which consists of 1000 spoken sentences from 34 different speakers. The results showed that the system was able to generate an output string that accurately represented the spoken sentence as written. This success highlights the potential of combining CV and CNN models to rewrite spoken sentences into text, which has applications in speech recognition, accessibility technologies, and more.

In a recent study conducted by Liang Lin et al. [13] proposes a novel reconfigurable part-based model called the And-Or graph model for object shape detection in images. The model is divided into four layers: leaf-nodes for recognising contour fragments, or-nodes for activating leaf-nodes, and-nodes for capturing holistic shape deformations, and a root-node for handling global changes. The authors propose a structural optimization algorithm to train the And-Or model from weakly annotated data. The authors publish a new shape database with annotations that contains over 1500 difficult form cases for recognition and detection. On various challenging datasets, the proposed model outperforms current state-of-the-art algorithms in robust shape-based object detection against background clutter. On various challenging datasets, the proposed model outperforms current state-of-the-art algorithms in robust shape-based object detection against background clutter.

The system utilized in [24] included working out Signal-to-Clamor Proportions (SNR) for both the spotless sound (SNRi) and the sound with foundation commotion (SNRw) utilizing a solitary mouthpiece. The dataset used for this examination was the Methodology corpus, which contains high-goal, high-framerate video transfers and sound accounts caught under uproarious circumstances. The review's outcomes fixated on the show and inside and out conversation of the results got from the execution of a general media programmed discourse acknowledgment (ASR) motor, which was prepared and thoroughly tried utilizing the information from the Methodology corpus. This exploration contributes significant bits of knowledge into the field of general media ASR and its exhibition in genuine world, loud conditions.

To progress in the field of lip-perusing, [18] tackled the force of the MS-TCN technique for lip-perusing related to the AV-HuBERT model for highlight extraction. The exploration utilized a significant dataset comprising of Persian word-level lipreading materials, incorporating a tremendous storehouse of roughly 244,000 video tests. Quite, the results of this examination exhibited a noteworthy improvement in exactness while utilizing the AV-HuBERT highlight extraction approach. These discoveries highlight the urgent job that include extraction plays in supporting the exhibition of lip-understanding frameworks, especially for the acknowledgment of Persian words.

Driving the way in the field of word-level lipreading, [25] presented an imaginative methodology that rotates around a two-stream model. This model was painstakingly created to catch both static and dynamic elements by utilizing explicit CNN streams, each finely tuned with successful convolutional structures at the front-end. A careful assessment was done on two broadly perceived lipreading datasets to survey the presentation of the proposed model. The consequences of this thorough examination uncovered a remarkable result - another best in class execution level on these difficult lipreading datasets, meaning a critical progression in word-level lipreading innovation.

TABLE 2. A SUMMARY OF RESEARCH REVIEWED IN LITERATURE REVIEW

Work	ML Algorithm	Features	Dataset
Joon Chung, Andrew Zisserman [1]	Watch, Listen, Attend and Spell (WLAS) network	13-dimensional MFCC features	Lip Reading Sentences' (LRS) dataset
Yannis Assael, Brendan Shillingford, Shimon Whiteson, Nando De Freitas [2]	LipNet (end-to-end trainable model)	Spatiotemporal convolutions, recurrent network	GRID corpus (contains entire sentences)
Pingchuan Ma et al [3]	Densely-Connected Temporal Convolutional Networks (DC-TCN)	Frame-wise features from a mouth Region-Of-Interest (ROI) encoder	LRW dataset
Huijuan Wang et al [4]	3D Convolutional Vision Transformer (3DCvT)	Spatio-temporal features of continuous images	LRW and LRW-1000 (large-scale lip reading datasets)
Xing Zhao et al [5]	Mutual Information Maximization	Local features and global sequence features	LRW (Lip Reading in the Wild) dataset
Daqing Chen et al. [6]	For visual data processing, 3D convolution is used first, followed by 2D ResNet. For phoneme recognition, Transformers with multi-headed attention are used. For sequential data processing, an RNN is used as the language model.	Lip-reading issues include visual ambiguity, inadequate temporal resolution, efficient storing of spatial-temporal information, speaker reliance, head posture fluctuation, and lighting circumstances.	LRS2 (BBC Lip Reading Sentences)
L Kumar et al. [7]	Deep learning models used for audio visual speech recognition - Stack of CNN layers used for lip image classification	Spectrogram features used for audio signals representation. - Mel-Frequency Cepstral coefficient (MFCC) and spectrogram commonly used.	LibriSpeech and Grid datasets
K Prajwal et al. [8]	Pooling process based on attention, Visual Speech Detection (VSD) model	Visual speech representations, sub-word units	LRS2, LRS3, AVA-ActiveSpeaker
Dimitrios Tsourounis et al [9]	ALSOS (Alternating Spatiotemporal and Spatial Convolutions) ResNet: Residual Networks	Spatiotemporal and spatial convolutions	Lip reading datasets in Greek and English languages
Xinshuo Weng et al. [10]	Two-stream I3D, Bi-LSTM	Grayscale video, optical flow	LRW (Lip Reading in the Wild)
Tasuya Shirakata et al. [11]	GRU model, auto-encoder neural network	HP, Shape, Exp, AU (Action Unit-based)	OuluVS, CUAVE, CENSREC-1-AV
Hang Chen et al. [12]	Hierarchical pyramidal convolution (HPConv), self-attention	Multi-scale processing, spatial feature extraction	Lip Reading in the Wild (LRW) dataset
Liang Lin et al. [13]	Discriminatively trained And-Or	Global shape features,	Shape database

	graph model	spatial contextual features	
Munender Varshney et al [14]	VAE (Variational Auto-Encoder) - Transformer model	Models using MFCC feature perform better. - Proposed approach outperforms recent Lip2wav models.	Grid dataset
Saakshi Bhosale et al [15]	Deep Convolutional Neural Network (CNN) Model End-to-end deep learning architecture models.	CNN model for image-based problem Dataset includes distance of lips for pronouncing each word	Dataset includes words and lip distances for pronunciation. - Lip distances used for word segmentation and classification.
Guangxin Xing et al. [16]	LSTM encoder-decoder architecture, spatiotemporal convolutional neural network (STCNN), Word2Vec, Attention model	Mouth shapes, homophones	GRID, LRW, LRW-1000
Javad Peymanfard et al. [18]	AV-HuBERT model	Embedding vectors obtained from AV-HuBERT model	Persian word-level lipreading dataset (244,000 videos), OuluVS dataset (phrases), LRW dataset (word-level lip-reading)
Xubo Liu et al. [19]	Speech-driven lip animation model, VSR model	Lip images, speech utterances	LRS3, Librispeech, CelebA
Marzieh Oghbaie et al. [20]	LCANet, Highway Networks, 3D CNN, Bidirectional Gated Recurrent Units (Bi-GRU)	Spatial and sequential feature extractors	Wild LRRo, Lab LRRo
Souheil Fenghour et al. [26]	Weighted Finite State Transducers (WFSTs), Hidden Markov Models (HMMs)	Visemes	Lip Reading Sentences in the Wild (LRS2)
Ümit Atila et al [21]	Bi-LSTM (Bidirectional Long Short-Term Memory)	Features are extracted from video frames using pre-trained CNN models. - Feature vectors are obtained from specific layers of ResNet-18 and GoogLeNet models.	Two new datasets were created, one with 111 words and the other with 113 sentences, for Turkish lip-reading research
Souheil Fenghour et al [22]	Neural network-based lip reading system Transformer with a unique topology for classification of visemes	Purely visual cues, visemes as classes	BBC LRS2 dataset with 45839 sentences for training and 1243 sentences for testing
Shashidhar Rudregowda et al [23]	VGG16 Convolutional Neural Network	Visual features derived from image sequences	Kannada Language dataset with five words (Avanu, Bagge, Bari, Guruthu, Helida)
Souheil Fenghour [27]	Deep learning network model with an Attention based Gated Recurrent Unit	Conversion of visemes to words, discriminating between homophone words, robustness to incorrectly classified visemes	LRS2 and LRS3 datasets
Munender Varshney et al [25]	VAE with transformer, metric learning	MFCC, raw audio, LSP	GRID, Lip2Wav, Chemistry

IV. CONCLUSION

This survey provides an overview of the evolution of automated lip reading systems over the period 2007-2021. It outlines the shift from traditional algorithms used for letter and digit classifying to deep neural networks for word and sentence prediction. New datasets cover not only a larger vocabulary of thousands of words, but also speakers in different poses, lighting, and resolution. Lip-reading systems are composed of two main components: feature extractions and classification. CNNs are the most prominent frontend network, as they are able to capture both temporal and spatial features. Autoencoder, however, has the advantage of being able to map high-dimensional data into a smaller space without the need for labeled classification. As for classification networks, recurrent neural networks (RNN) are the dominant type, although Transformers and temporal convolutional networks (TCN) have been replacing RNNs in recent years due to their capability for efficient parallel computing, the ability to capture long-term relationships, and the shorter training periods. Automated lip-reading technology is still facing a number of obstacles. These include the ability to anticipate spoken words that were not included in the training data, as well as visual ambiguities. For example, some words may appear to be the same when spoken, but have different semantic and syntactic characteristics. Additionally, there is still a problem with speaker dependency when applying lip-reading to non-trained individuals. Finally, generalizing lip-reading systems must be able to handle videos with varying spatial resolution and frame rates, which contain varying amounts of temporal data.

REFERENCES

[1] Chung, J.S., Zisserman, A. (2017). Lip Reading in the Wild. In: Lai, SH., Lepetit, V., Nishino, K., Sato, Y. (eds) Computer Vision – ACCV 2016. ACCV 2016. Lecture Notes in Computer Science(), vol 10112. Springer, Cham.

[2] Yannis, M., Assael., Brendan, Shillingford., Shimon, Whiteson., Nando, de, Freitas. "LipNet: End-to-End Sentence-level Lipreading." arXiv: Learning, undefined (2016).

[3] P. Ma, Y. Wang, S. Petridis, J. Shen and M. Pantic, "Training Strategies for Improved Lip-Reading," ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, Singapore, 2022, pp. 8472-8476, doi: 10.1109/ICASSP43922.2022.9746706.

[4] H. Wang, G. Pu and T. Chen, "A Lip Reading Method Based on 3D Convolutional Vision Transformer," in *IEEE Access*, vol. 10, pp. 77205-77212, 2022, doi: 10.1109/ACCESS.2022.3193231.

[5] X. Zhao, S. Yang, S. Shan and X. Chen,"Mutual Information Maximization for Effective Lip Reading," 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020), Buenos Aires, Argentina, 2020, pp.420-427, doi: 10.1109/FG47880.2020.00133.

[6] El-Bialy, R., Chen, D., Fenghour, S., Hussein, W., Xiao, P., Karam, O. H., & Li, B. (2023). Developing phoneme-based lip-reading sentences system for silent speech recognition.*CAAI Transactions on Intelligence Technology*, *8*(1), 129-138. https://doi.org/10.1049/cit2.12131

[7] L Ashok Kumar, D Karthika Renuka, S Lovelyn Rose, M C Shunmuga priya, I Made Wartana, Deep learning based assistive technology on audio visual speech recognition for hearing impaired, International Journal of Cognitive Computing in Engineering, Volume 3, 2022.

[8] Prajwal, K., R., Triantafyllos, Afouras., Andrew, Zisserman. (2021). Sub-word Level Lip Reading With Visual Attention. arXiv: Computer Vision and Pattern Recognition.

[9] Lip Reading by Alternating between Spatiotemporal and Spatial Convolutions. *J. Imaging* 2021, *7*, 91. https://doi.org/10.3390/jimaging7050091.

[10] Weng, X., & Kitani, K. (2019). Learning Spatio -Temporal Features with Two-Stream Deep 3D CNNs for Lipreading. ArXiv. /abs/1905.02540.

[11] Shirakata, Tasuya, and Takeshi Saitoh. "Lip reading using facial expression features." Int. J. Comput. Vis. Signal Process 1.1 (2020): 9-15.

[12] Chen, H., Du, J., Hu, Y., Dai, L., Lee, C., & Yin, B. (2020). Lip-reading with Hierarchical Pyramidal Convolution and Self-Attention. *ArXiv*./abs/2012.14360

[13] L. Lin, X. Wang, W. Yang and J. -H. Lai work on "Discriminatively Trained And-Or Graph Models for Object Shape Detection," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 5, pp. 959-972, 1 May 2015, date of issue (doi) : 10.1109/TPAMI.2014.2359888.

[14] N. Deshmukh, A. Ahire, S. H. Bhandari, A. Mali and K. Warkari, "Vision based Lip Reading System using Deep Learning," *2021 International Conference on Computing, Communication and Green Engineering (CCGE)*, Pune, India, 2021, pp. 1-6, date of issue (doi): 10.1109/CCGE50943.2021.9776430.

[15] Bhosale, Saakshi, et al. "An Application to Convert Lip Movement into Readable Text."

[16] Xing, G., Han, L., Zheng, Y., & Zhao, M. (2023). Application of deep learning in Mandarin Chineselip-reading recognition. *EURASIP Journal on Wireless Communications and Networking*, *023*(1), 1-14. https://doi.org/10.1186/s13638-023-02283-y

[17] Jin Ting, Chai Song, Hongyang Huang, Taoling Tian, A Comprehensive Dataset for Machine-Learning-based Lip-Reading Algorithm, Procedia Computer Science, Volume 199, 2022.

[18] J. Peymanfard, A. Lashini, S. Heydarian, H. Zeinali and N. Mozayani, "Word-level Persian Lipreading Dataset, "*2022 12th International Conference on Computer and Knowledge Engineering (ICCKE)*, Mashhad, Iran, Islamic Republic of, 2022, pp. 225-230, date of issue : 10.1109/ICCKE57176.2022.9960105.

[19] X. Liu *et al*., "SynthVSR: Scaling Up Visual Speech

Recognition With Synthetic Supervision," *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Vancouver, BC, Canada, 2023,pp. 18806-18815, doi:10.1109/CVPR52729.2023.01803.

[20] Oghbaie, M., Sabaghi, A., Hashemifard, K., & Akbari, M. (2021). Advances and Challenges in Deep Lip Reading . ArXiv. /abs/2110.07879.

[21] Ü. Atila, F. Sabaz, Turkish lip-reading using Bi-LSTM and deep learning models. Eng. Sci. Technol. Int. J. 35, 101206 (2022).

[22] S. Fenghour, D. Chen, K. Guo and P. Xiao, "Lip Reading Sentences Using Deep Learning With Only Visual Cues, " in *IEEE Access*, volume. 8, pp. 215516-215530, 2020, date of issue (doi): 10.1109/ACCESS.2020.3040906.

[23] Rudregowda S, Patil Kulkarni S, H L G, Ravi V, Krichen M. Visual Speech Recognition for Kannada Language UsingVGG16 Convolutional Neural Network *Acoustics*. 2023; 5(1):343-353. Refer the link below : https://doi.org/10.3390/acoustics5010020

[24] Czyzewski, A., Kostek, B., Bratoszewski, P. et al. An audio-visual corpus for multimodal automatic speech recognition. J Intell Inf Syst 49, 167–192 (2017).

[25] Li H, Yadikar N, Zhu Y, Mamut M, Ubul K. Learning the Relative Dynamic Features for Word-Level Lipreading. *Sensors*. 2022; 22(10):3732. https://doi.org/10.3390/s22103732

[26] Fenghour, S., Chen, D., Guo, K., & Xiao, P. (2020). Disentangling Homophemes in Lip Reading using Perplexity Analysis. ArXiv. /abs/2012.07528

[27] K. Thangthai and R. Harvey, ‘‘Improving computer lipreading via DNN sequence discriminative training techniques,’’ in Proc. Interspeech, Aug. 2017, pp. 1–5.

[28] F. Tao and C. Busso, "End-to-End Audiovisual Speech Recognition System With Multitask Learning,"in *IEEE Transactions on Multimedia*, vol. 23, pp. 1-11,2021, doi: 10.1109/TMM.2020.2975922.

[29] G. Sterpu, C. Saam and N. Harte, "How to Teach DNNs to Pay Attention to the Visual Modality in Speech Recognition," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, volume .28, pp.1052-1064, 2020, date of issue (doi): 10.1109/TASLP.2020.2980436.

[30] Akhter N, Ali M, Hussain L, Shah M, Mahmood T, Ali A, Al-Fuqaha A. Diverse Pose Lip-Reading Framework. *Applied Sciences*. 2022; 12(19):9532. https://doi.org/10.3390/app12199532

[31] Tao, Fei, and Carlos Busso. "Bimodal Recurrent Neural Network for Audiovisual Voice Activity Detection." INTERSPEECH. 2017.

[32] Stafylakis .T & Tzimiropoulos G. (2017). Combining Residual Networks with LSTMs for Lipreading. *ArXiv*. /abs/1703.04105

[33] Noda, Kuniaki, "Audio-visual speech recognition recognition using deep learning." *Applied intelligence* 42 (2015): 722-737.

[34] Matthews, I. et al. “Extraction of Visual Features for Lipreading.” IEEE Trans. Pattern Anal. Mach. Intell. 24 (2002): 198-213.

[35] Koller, Oscar et al. “Deep Learning of Mouth Shapes for Sign Language.” 2015 IEEE International Conference on Computer Vision Workshop (ICCVW) (2015): 477-483.