IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 2, November 2025
Impact Factor: 7.67

DeepLip: A Comprehensive Review on Deep Learning Based Lip Reading for Visual Speech Recognition

Nikhil Jagtap[1], Shreya Banarase[2], Mitali Kucheriya[3], Vivek Kumavat[4]
[1-4]Department of Computer Engineering,
ISBM College of Engineering, Nande, Pune, India
jagtapnikhil2411@gmail.com, shreyabanarase123@gmail.com,
mitalikucheriya@gmail.com, vivekkumawat1304@gmail.com

Abstract: *Visual speech recognition, or lip reading, is one of the fastest growing domains in computer vision and speech processing. Lip reading has transitioned from rule-based models to robust, fully end-to-end trainable systems through deep learning. In this review paper, we systematically summarize the evolution of lip reading from early hand-crafted feature extraction to current architectures for lip reading research as methods such as 3D CNN, Residual Networks, and Transformer-based methods. We define the state of the art for lip reading in terms of datasets, algorithms, evaluation metrics, and implementation. We also discuss open questions, ethical issues, and potential future directions to further progress the field of visual speech perception.*

Keywords: Lip reading, visual speech recognition, deep learning, convolutional neural network, trans- former, datasets

I. INTRODUCTION

Lip reading is the process of understanding speech by observing visual information from the movements of the mouth. Lip reading aids people with hearing loss, and makes it possible to comprehend speech in quiet and noisy environments. The early systems were based on hand-engineered features and statistical models such as Support Vector Machines (SVM) and Hidden Markov Models (HMM). However, these methods were sensitive to changes in speaker, pose and lighting.

Lip reading has transitioned from static image processing to full video-based sequence learning, thanks to deep learning in general, and particularly convolutional neural networks (CNN) and recurrent neural networks (RNN). Large scale datasets such as LRW and LRS3 have enabled end-to-end training on realistic datasets, as well as advances in hardware acceleration. This paper seeks to highlight these developments and explore the technology associated with a modernized lip reading pipeline.

II. HISTORICAL DEVELOPMENT OF LIP READING

In the earlier days of visual speech systems (before 2010), hand-crafted descriptors like optical flow, lip contours and mouth region geometric parameters were the main features used. The classifiers used to model temporal transitions such as HMMs, but the features lacked robustness.

The years 2014 - 2018 saw the introduction of deep learning models. Despite LSTMs or BiLSTMs being the networks employed for modeling time-dependent information, CNNs were the ones who efficiently and effectively extracted the relevant spatial features. Sub- sequently, the 3D CNNs were introduced to the realm of deep learning and were able to simultaneously learn the temporal and spatial dependencies [5]. Additionally, the use of residual networks (ResNet) not only prevented the gradient degradation in the deeper networks but also enhanced the feature extraction process.

After all, the reason the Transformer and attention-based models are showing superb results is pretty much their ability to model long-range temporal dependencies. The change from frame-by-frame predictions to sequence-to-sequence

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 5, Issue 2, November 2025 Impact Factor: 7.67

transduction is exemplified in research like [1], [6], [7], which enables a straightforward mapping from visual input to text output.

III. LITERATURE REVIEW

Over the span of two decades, the domain of visual speech recognition (lip reading) has experienced a drastic transformation from the era of geometric feature modeling to huge end-to-end deep learning systems. This part of the paper outlines the different phases of development in recognition technology, starting from the early statistical approaches and ending with modern multimodal and Transformer-based architectures, and also featuring the significant works that have impacted the field.

A. Early Developments in Visual Speech Recognition

In the beginning, researchers mainly used manually created visual features to depict lip shape and movement. Cetingul et al. [11] did a comparative study of optical flow and appearance-based descriptors for discriminative speaker identification with the use of GMMs and HMMs. In the same way, Dalka and Czyzewski [10] created a visual lip and ges- ture recognition-based human-computer dialogue interface using traditional machine learning models. But all these systems had the common problem of being unable to generalize properly when the lighting, the angle, and the type of speaker varied.

B. Introduction of Deep Learning Models

The advent of deep neural networks with convolutional and recurrent architectures and methods was a turning point in the field of lip reading. Chung and Zisserman [7] presented a CNN-based architecture that was able to automatically learn features of large-scale, in-the- wild datasets, which was a big step away from using handcrafted pipelines. Margam et al.
[5] built on this by integrating 2D–3D CNNs with Bidirectional Long Short-Term Memory (BLSTM) and HMMs to grasp the sequential lip motion keeping. Haliassos et al. [6] showed the power of such motion features when they used them in visual lip embedding, even in the tasks of deepfake and face forgery detection.

C. Hybrid CNN and Transformer Architectures

A new breed of architectures that are based on hybrid and attention mechanisms have been introduced recently, wherein the CNNs are coupled with Transformer-based modules. Sarhan et al. [9] came up with HLR-Net, a deep learning model for extracting lip features hierarchically. Bardhan et al. [8] set up a low-power real-time CNN system for lip reading, aiming at computational efficiency as the main concern. Khekare et al. [3] provided a detailed survey about the technical aspects of multimedia fusion with a specific focus on the integration of 3D convolutions for better context learning. Chauhan et al. [4] showed the utility of CNN- based assistive systems for communication-impaired persons, which is one of the real-world applications of the deep visual speech models.

D. Dataset-Driven Progress

The availability of datasets has been a major factor in improving lip-reading performance significantly. The GRID corpus [13] was the first to release controlled audio-video sentence data for testing on benchmarks. Later, Chung and Zisserman [15] made the LRW dataset available, which is capable of large-scale word-level recognition in the wild. The LRW-1000 dataset [17] was built on this by adding multilingual and naturally distributed Mandarin data with over 1 million word instances. The LRS2 [16] and LRS3 [14] datasets first offered BBC and TED talks, respectively, for sentence-level lip reading along with the pretrain-train- test splits tailored for end-to-end models. Moreover, the VoxCeleb1 dataset [18] provided a large-scale multimodal audio-visual corpus for speaker recognition that supported cross- domain training for lip-reading tasks, hence making it richer.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 5, Issue 2, November 2025 Impact Factor: 7.67

E. Recent Advancements (2023–2025)

During the period of 2023–2025, contemporary lip reading studies, apart from urban, focused on models that are lightweight, self-supervised, and based on Transformers to increase scalability and to decrease the need for physically setting up that kind of research. Putcha et al.

[2] introduced an all-in-one model for visual-to-text translation that is resilient to background noise and variations in lighting. Khekare et al. [3] utilized cross-modal feature fusion of visual and linguistic cues in order to gain recognition accuracy, whereas Ma and Sun [12] did the spatiotemporal feature enhancement and attention-driven learning for fine-grained articulation modeling. In unison, these works signify the transition towards efficient, multimodal, and self-supervised systems that are robust across different speakers and languages.

F. Summary of Literature Review

Table I summarizes key studies and their primary contributions in lip reading research.

TABLE I: SUMMARY OF REPRESENTATIVE LIP READING STUDIES AND REPORTED RESULTS

Author (Year)	Model Type	Dataset	Result
Cetingul et al. (2006)	HMM + Optical Flow	Custom	68% Speaker ID
Dalka & Czyzewski (2010)	ML-based Interface	Local	70% Lip gesture recognition
Chung & Zisserman (2018)	3D CNN	LRW	83.0% Top-1 accuracy
Margam et al. (2019)	3D–2D CNN + BLSTM	GRID/LRW	85.4% Word accuracy
Sarhan et al. (2021)	CNN (HLR-Net)	LRW-1000	78.6% Top-1 accuracy
Haliassos et al. (2021)	CNN + Forgery Detection	Custom	94.2% Classification accuracy
Bardhan et al. (2023)	Real-Time CNN	Custom	82.5% Real-time accuracy
Khekare et al. (2024)	3D CNN + Fusion	LRW/LRS3	88.1% Word recognition
Putcha et al. (2024)	End-to-End CNN	Custom	84.7% Translation accuracy
Ma & Sun (2025)	Transformer	LRS3	25% WER

IV. DATASETS: EVOLUTION AND EXPANSION

Progress in lip reading is largely attributed to the availability of large annotated datasets.

Table II lists the commonly used datasets.

Dataset	Year	Level	Speakers
AVLetters	2008	Letters	10
GRID	2006	Sentences (controlled)	34
OuluVS	2012	Words	20
LRW	2016	Word level (wild)	500+
LRW-1000	2018	Word level, multi-language	1,000+
LRS2	2018	Sentence level (wild)	1,000+
LRS3	2019	Sentence level (TED talks)	4,000+
VoxCeleb	2018	Multimodal speaker dataset	6,000+

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 2, November 2025
Impact Factor: 7.67

A. Trends in Datasets

The improvement of datasets for lip-reading has been instrumental in the development of visual speech recognition based on deep learning. Some of the most frequently used datasets are shown in Table II. These examples of their frames show the change from controlled studio recordings to wild video source that was varied and diverse.

a. Scale: The scale of datasets has been increased from a few hundreds of video clips to millions of frames and this has made it possible to train large deep models and to have better generalization.

b. Diversity: Controlled studio settings (for example, GRID, AVLetters) have been replaced by varied in-the-wild settings that show different natural poses, lighting, and speakers (for example, LRW, LRS3).

c. Annotation Quality: The precision of the annotations has been increased and massive data generation has been made possible by using audio-based synchronization tools for automatic transcript alignment.

d. Language Coverage: The multilingual and spontaneous speech materials included in recent datasets like LRW-1000 and LRS3-TED can support not only cross-lingual gen- eralization but also more realistic training scenarios.

B. Description Of Datasets

1) GRID: The Grid Corpus represents a significant development in the realm of multitalker audiovisual sentence corpus, which is aimed at joint computational-behavioral studies in speech perception. The corpus contains, in short, very good quality audio and video (face) channels of 1000 utterances produced by every one of 34 speakers (18 male, 16 female), which adds up to 34000 sentences in total. All sentences have the structure ”put red at G9 now” [13].

Fig. 1. GRID Corpus Dataset [13]

2) LRW: The dataset is made of a maximum of 1000 utterances for each of 500 different words that were pronounced by hundreds of diverse speakers. The videos are of 29 frames (1.16 seconds) long and the word is portrayed in the center of the video. The metadata provides the word duration from which the start and end frames can be calculated [15].

TABLE III: DATASET SPLIT DETAILS WITH CLASS DISTRIBUTION

Set	Dates	Classes	per Class
Train	01/01/2010 – 31/08/2015	500	800–1000
Validation	01/09/2015 – 24/12/2015	500	50
Test	01/01/2016 – 30/09/2016	500	50

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 5, Issue 2, November 2025 Impact Factor: 7.67

Fig. 2. Lip Reading Wild [15]

3) LRW-1000: LRW-1000 is a large-scale benchmark with a natural distribution for word- level lipreading in the wild with 1000 classes and approximately 718,018 video samples of more than 2000 individual speakers. There are in total over 1,000,000 instances of Chinese characters [17].

TABLE IV: SUMMARY OF LRW-1000 DATASET CHARACTERISTICS

Attribute	Description
Dataset Name	LRW-1000
Type	Word-level lip-reading (in the wild)
Language	Mandarin Chinese
Classes	1,000 (syllable-based words)
Video Samples	718,018
Speakers	Over 2,000
Character Instances	Over 1,000,000 Chinese characters
Variability Factors	Pose, age, gender, make-up, lighting, and video resolution
Purpose	Benchmark for large-scale, naturally distributed word-level lip-reading
Challenges	High variation in speech mode, speaker diversity, and imaging conditions

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 2, November 2025
Impact Factor: 7.67

Fig. 3. LRW-1000 Dataset

4) LSR2: The dataset is made of thousands of uttered sentences extracted from the BBC TV. The maximum length of each sentence is 100 characters. The partitioning of the training, validation, and test sets is based on the broadcast date [16].

The utterances that are part of the pre-training set include not only part-sentences and multiple sentences, but also single full sentences or phrases in the case of the training set. [16] acknowledges some overlap between the pre-training and the training sets.

TABLE V: DATASET SPLIT STATISTICS WITH VOCABULARY AND UTTERANCE COUNTS

Set	Dates	Utterances	Word Instances	Vocab
Pre-train	11/2010–06/2016	96,318	2,064,118	41,427
Train	11/2010–06/2016	45,839	329,180	17,660
Validation	06/2016–09/2016	1,082	7,866	1,984
Test	09/2016–03/2017	1,243	6,663	1,698

Fig. 4. LSR2 Dataset [16]

5) LRS3: A massive collection of yet unprocessed spoken sentences from TED and TEDx videos is what the dataset is made up of. The videos that formed the test set are totally different from those that were used in pre-training and training+validation [14].

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 2, November 2025
Impact Factor: 7.67

TABLE VI: STATISTICS OF THE LIP READING DATASET SPLITS

Set	Videos	Utterances	Word Instances	Vocab
Pre-train	5,090	118,516	3.9M	51k
Trainval	4,004	31,982	358k	17k
Test	412	1,321	10k	2k

Fig. 5. Lip Reading Sentences 3 [14]

C. VoxCeleb1 Dataset

The VoxCeleb1 dataset comprises more than 100,000 utterances of 1,251 celebrity speakers and is derived from the videos uploaded to YouTube. It is intended for speaker identification and verification tasks and contains variations in age, nationality, accent, and recording conditions, thus establishing it as a rigorous benchmark for multimodal audio visual research [18].

TABLE VII: VOXCELEB1 VERIFICATION SPLIT STATISTICS

Split	Speakers	Videos	Utterances
Development (dev)	1,211	21,819	148,642
Test	40	677	4,874

TABLE VIII: VOXCELEB1 IDENTIFICATION SPLIT STATISTICS

Split	Speakers	Videos	Utterances
Development (dev)	1,251	21,245	145,265
Test	1,251	1,251	8,251

Dataset Notes: VoxCeleb1 was created for large-scale speaker recognition in unsu- pervised and uncontrolled environments. The source of data was online media and the resulting recording has background noise, overlapping speech, and varied acoustic conditions. More updates have been made to the dataset later on, which included removal of duplicate samples (VoxCeleb 1.1) and elimination of overlaps with "Speakers in the Wild (SITW)" dataset.

Access and Licensing: The dataset's metadata is licensed for distribution under a Creative Commons Attribution-ShareAlike 4.0 International License. Original audio files, as well as URLs, have been taken off the public website due

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 2, November 2025
Impact Factor: 7.67

to privacy concerns. Users wishing to gain access must be aware of and comply with the dataset's Privacy Notice and licensing requirements.

Fig. 6. VoxCeleb1 Dataset Overview [18]

V. ARCHITECTURES AND TECHNIQUES

A. Preprocessing

Initially, the system detects the face in each frame and then extracts the mouth region of interest (ROI) for further analysis. This feature of the face is particularly important since the most informative visual signals for both emotion and lip reading are those emanating from the mouth region.

Dlib and OpenCV are implemented when detecting facial landmarks, as these can find facial features extremely accurately, while still accommodating for head angles and facial expression changes. Furthermore, other data augmentation methods are also applied to each frame (as there are various types of augmentation), for example scaling, rotation, and lighting modification, in order to improve the robustness of the models. Moreover, the more common data augmentation techniques to improve the model's robustness include frame skipping (to create temporal variation), brightness modification (to account for light changes), and random cropping to improve spatial generalization.

B. Feature Extraction

The feature extraction methods used, based on CNN architecture, process video volumes of 3D or 2D representations of images. 3D-CNNs are recordings of frame sequences across time dynamic, while 2D models use architectures such as VGG or ResNet to build embeddings at the level of frames. Hybrid networks use both spatial and temporal convolutions that allow a comprehensive comprehension of appearance and motion. The incorporation of ResNet's residual learning aspect helps to train deeper models successfully while alleviating the vanishing gradient problem. In modern visual speech recognition systems, the most common models used are either 3D-ResNet, I3D or biological ResNet-type architectures [3], [9].

C. Sequence Modeling

Sequence modeling is important for visual speech recognition because it can model the temporal dependencies very well. The core models for this task are recurrent neural networks (RNNs), long short-term memory (LSTM) networks, and attention-based transformers. Using the connectionist temporally-constrained (CTC) loss function, in other words, it can train without any special frame-level annotations because it maps the input sequences to the labels instead. In contrast, attention-based sequence-to-sequence models output tokens based on learning the dynamic alignment from the input sequences and the target output sequences directly.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 2, November 2025
Impact Factor: 7.67

Within the past couple of years, Transformer-based models such as the Visual Speech Transformer (VSR-T) have demonstrated great success. The best aspect of these types of large pre-trained models that utilize self-attention mechanisms is that they can model the global temporal relationships across the video frames, and they do this very well [8].

D. Decoding and Language Modeling

Beam search is commonly used as a decoding strategy, in conjunction with greedy decoding and external language models. Since neural language models are built by training on large text corpora, they induce contextual changes during inference that improve overall word accuracy and minimize word error rate.

VI. SYSTEM ARCHITECTURE

The suggested DeepLip system integrates every functional component. The workflow diagram in Figure 7 depicts the steps in the processing sequence: face localization, lip ROI extraction, 3D CNN and ResNet feature extractor-decoding system, BiLSTM classifier, training through the dataset, and decoding relying on the language model.

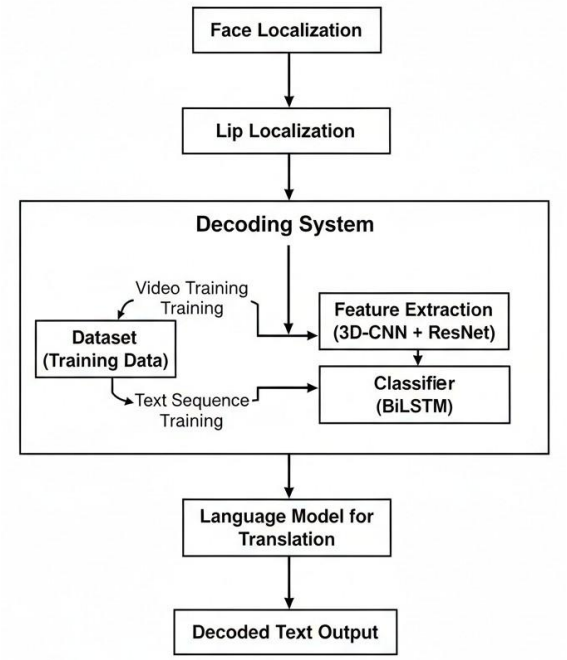

Fig. 7. Face and lip localization, decoding system (3D CNN + ResNet + BiLSTM), dataset-based training, language model, and final text output comprise the DeepLip System Architecture.

VII. EVALUATION AND METRICS

Word Error Rate (WER), Character Error Rate (CER), and Top-1 accuracy are the three main evaluation metrics which are very commonly used. Among the sentence-level tasks, WER is still the most preferred metric. The datasets used for testing such as LRW and LRS3 have been consistent in the benchmarking which allows for an easy comparison. In LRS3, the WER of transformer-based models drops to about 40% when no language models are used and goes down to around 25% with fusion techniques [12].

VIII. CHALLENGES AND FUTURE SCOPE

The use of deep models has led to significant improvements in the performance, however, the challenges mentioned below have not been solved yet:

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 5, Issue 2, November 2025 Impact Factor: 7.67

i. Cross-speaker generalization: The performance of the model might be reduced in cases of unseen speakers due to the differences in articulation.
ii. Low-resource learning: The creation of annotated datasets is a costly process; neverthe- less, few-shot and self-supervised learning can be reckoned as possible solutions to the problem.
iii. Pose and occlusion: The correctness of the recognition diminishes when it comes to non-frontal poses and covering the mouth with masks.
iv. Multilingual scalability: The application of the lip reading technique should go beyond English, which in turn, calls for the availability of large multilingual datasets.
v. Privacy and ethics: The use of visual speech technologies could lead to a situation where privacy is violated and therefore, it is very important to ensure that the technology is designed based on the principles of consent and ethics.
The future research might involve the use of self-supervised pre-training, application of diffusion models for the purpose of generative data augmentation, and employing multimodal Transformer architecture for integrated audiovisual recognition.

IX. CONCLUSION

The current review renders an exhaustive examination of deep learning-driven lip reading systems, describing their development, datasets, and techniques clearly. The improvements in 3D CNNs, ResNets, and Transformers have resulted in end-to-end models that can identify visual speech with remarkable precision. The ongoing improvements in data gathering, com- puting power, and model fine-tuning will make lip reading an integral part of human-computer dialogue.

ACKNOWLEDGMENT

The authors have the privilege to sincerely thank Prof. Shobha Bamane for her priceless mentoring and technical knowledge during the entire process of this work. Also, the authors would be grateful to Dr. K. N. Tripathi, Chair of the Computer Engineering Department, ISBM College of Engineering, Pune, for his provision of the facilities and the environment that made this research possible. The authors are thankful to ISBM College of Engineering for their support in academic research and innovation.

REFERENCES

[1] N. Rathipriya and N. Maheswari, “A comprehensive review of recent advances in deep neural networks for lip reading with sign language recognition,” IEEE, 2024.
[2] S. T. K. Putcha, Y. S. V. Rajam, K. Sugamya, and S. Gopala, “Text Extraction and Translation Through Lip Reading using Deep Learning,” Int. J. Adv. Comput. Sci. Appl., vol. 15, no. 6, 2024.
[3] G. Khekare, G. Majumder, N. Shelke, G. Sakarkar, and S. Buchade, “A Deep Dive into Existing Lip Reading Technologies,” in Proc. Int. Conf. on AI and Quantum Computation Based Sensor Application, pp. 1–6, IEEE, 2024.
[4] S. Chauhan, S. Dubey, V. Samadhia, S. Gupta, and U. Hariharan, “Establishing Communication Through Lip Reading With The Aid of Machine Learning,” in Proc. 1st Int. Conf. on Innovative Engineering Sciences and Technological Research, pp. 1–5, IEEE, 2024.
[5] D. K. Margam, R. Aralikatti, T. Sharma, A. Thanda, S. Roy, and S. M. Venkatesan, “Lip Reading with 3D 2D CNN BLSTM HMM and word CTC models,” arXiv preprint arXiv:1906.12170, 2019.
[6] A. Haliassos, K. Vougioukas, S. Petridis, and M. Pantic, “Lips do not lie: A generalisable and robust approach to face forgery detection,” in Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition, pp. 5039–5049, 2021.
[7] J. S. Chung and A. Zisserman, “Learning to lip read words by watching videos,” Computer Vision and Image Understanding, vol. 173, pp. 76–85, 2018.
[8] A. Bardhan, A. Singh, and S. H. Attri, “CNN Based Real Time Detection of Words from Lip Movements and Automated into Text,” in Int. Conf. on Artificial Business Analytics, Quantum and Machine Learning, pp. 87–102, Springer, 2023.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 5, Issue 2, November 2025 Impact Factor: 7.67

[9] A. M. Sarhan, N. M. Elshennawy, and D. M. Ibrahim, "HLR net: a hybrid lip reading model based on deep convolutional neural networks," Computers, Materials and Continua, vol. 68, no. 2, pp. 1531–1549, 2021.
[10] P. Dalka and A. Czyzewski, "Human-Computer Interface Based on Visual Lip Movement and Gesture Recognition,"
Int. J. Comput. Sci. Appl., vol. 7, no. 3, pp. 124–139, 2010.
[11] H. E. Cetingul, Y. Yemez, E. Erzin, and A. M. Tekalp, "Discriminative analysis of lip motion features for speaker identification and speech-reading," IEEE Trans. Image Process., vol. 15, no. 10, pp. 2879–2891, 2006.
[12] Y. Ma and X. Sun, "Spatiotemporal Feature Enhancement for Lip-Reading: A Survey," Appl. Sci., vol. 15, no. 8, 2025.
[13] M. Bentalb, "LipReading-Dataset," Kaggle, 2024. [Online]. Available:https://www.kaggle.com/datasets/mohamedbenta- lb /lipreading-dataset. [Accessed: Nov. 03, 2025].
[14] Afouras, T., Chung, J. S., & Zisserman, A. (2018). LRS3-TED: a large-scale dataset for visual speech recognition. arXiv preprint arXiv:1809.00496.
[15] Chung, J. S., & Zisserman, A. (2016, November). Lip reading in the wild. In Asian conference on computer vision (pp. 87-103). Cham: Springer International Publishing.
[16] Afouras, T., Chung, J. S., Senior, A., Vinyals, O., & Zisserman, A. (2018). Deep audio-visual speech recognition. IEEE transactions on pattern analysis and machine intelligence, 44(12), 8717-8727.
[17] Yang, S., Zhang, Y., Feng, D., Yang, M., Wang, C., Xiao, J., ... & Chen, X. (2019, May). LRW-1000: A naturally-distributed large-scale benchmark for lip reading in the wild. In 2019 14th IEEE international conference on automatic face & gesture recognition (FG 2019) (pp. 1-8). IEEE.
[18] Nagrani, A., Chung, J. S., & Zisserman, A. (2017). Voxceleb: a large-scale speaker identification dataset. arXiv preprint arXiv:1706.08612