Web Based Interface for Lip-Reading and Lip-Synchronization Using Deep Learning

1st Ramachandra C G
G*CSE(AI&ML) Department*
Keshav Memorial Institute of Technology
Hyderabad, Telangana, 500029, India

2nd K.Kaviarasu
CSE(AI&ML) Department
Keshav Memorial Institute of Technology)
Hyderabad, Telangana, 500029, India

3rd S.M.Prasad
CSE(AI&ML) Department
Keshav Memorial Institute of Technology)
Hyderabad, Telangana, 500029, India

4th Dinesh Kumar
CSE(AI&ML) Department
Keshav Memorial Institute of Technology)
Hyderabad, Telangana, 500029,India

5th V.Vinay Krishna
CSE(AI&ML) Department
Keshav Memorial Institute of Technology)
Hyderabad, Telangana, 500029, India

***Abstract*— Lip reading and lip synchronization are two key technologies that bridge the communication gap in various domains, such as assistive tools, multimedia content creation, and real-time applications. The current advancements, though impressive in terms of accuracy and performance, lack a coherent system that seamlessly integrates these functionalities. This paper proposes an interactive, web-based solution combining LipNet for lip reading and Wav2Lip for lip synchronization, leveraging state-of-the-art deep learning techniques. The system has employed datasets such as GRID and LRS2 for high precision in both transcription and audiovisual synchronization. Modular integration with the platform supports user-friendly access towards lip-reading transcription and creation of synchronized videos for adaptability towards diverse real-world applications. Multilingual support, on-demand real-time capacities, extended datasets, and mobile deployment mechanisms are discussed to enhance versatility, accessibility, and effectiveness. The proposed solution provides a basis for the transformation of communication tools, with a strong platform for assistive technologies, global education, and multimedia production.**

Keywords— Lip reading, lip synchronization, deep learning, LipNet, Wav2Lip, audiovisual synchronization, GRID dataset, LRS2 dataset, real-time applications, multilingual support, assistive technologies, multimedia content creation.

I. INTRODUCTION

Communication is one of the fundamental qualities of every human being; to accomplish day-to-day activities we need to hold a conversation with each other. Every 1 out of 5 individuals face difficulties with hearing, which makes it even more challenging for them to perform routine tasks in their everyday life. Another common problem is the noisy videos that are being utilized for security footage which do not even have clear audio. Lip reading is the art of following what a person is saying by watching their lips, face, and tongue; Lip synchronization is the ability to move your mouth, not talk or sing, so that its movements are somehow jive with what is happening on a recorded audio.

Lip synchronization is the process of moving your mouth to match the sound of a recorded audio recording without speaking or singing. It can be utilized to create visually appealing and decadent multimedia content for the recipient. It is essential for live multilingual broadcasting and dubbing in films and other media. It can also be used to make engaging avatars for educational and video game characters.

We employ the LipNet model by Assael et al. [1] for lip reading and the Wav2Lip model by Prajwal et al. [14] for lip synchronization. LipNet has demonstrated great accuracy in sentence-level lipreading challenges and is designed to decipher speech by analyzing spatiotemporal cues from video sequences and achieves a remarkable sentence-level word accuracy of 95.2% on The GRID[20] dataset. Conversely, Wav2Lip uses a conditional GAN structure to accurately synchronize lip motion and audio in videos.

Although current models have produced amazing models and attained good accuracy, there is still a lack of an interface that seamlessly transitions in real-time between lip synchronization and lip reading. An integrated system like this might close large gaps in usability and accessibility while addressing a variety of real-world applications in a cohesive way. The proposed system aims to improve communication of hearing-impaired people and streamline the process of media creation in industries like entertainment and education.

Our interactive web-based interface assists users in lip-reading and lip-syncing. By integrating these technologies, we hope to further the science of audiovisual processing and offer a user-friendly tool that can revolutionize communication synchronization and interpretation in a variety of contexts.

II. LITERATURE REVIEW

In the paper by Assael et al. present LipNet, the first end-to-end deep learning model that can lipread sentences. by the use of bidirectional Gated Recurrent Units (GRUs), spatiotemporal convolutional layers, and Connectionist Temporal Classification (CTC) loss. LipNet outperformed the prior word-level accuracy of 86.4% with a sentence-level accuracy of 95.2% on the GRID corpus.

With a Word Error Rate (WER) of 11.4% on unseen speakers. LipNet's rating was based only on the GRID corpus, which has a fixed grammar and a limited vocabulary even if it is sentence-level. Performance is still unknown on bigger, more varied datasets. The paper by Zhang et al. [3] introduces an end-to-end sentence-level lipreading model

leveraging Temporal Convolutional Networks (TCNs) that overcome the issues in recurrent architectures like GRUs. They achieved a Word Error Rate (WER) of 1.1% on seen speakers and 6.2% on unseen speakers, outperforming LipNet and other state-of-the-art models.

Although TCN handles gradients it may struggle with very long sequences. Research by Stafylakis & Tzimiropoulos [4] proposes a novel architecture that combines Residual Networks (ResNets) with Long Short-Term Memory networks (LSTMs) for sentence-level lipreading. On the GRID corpus, it attained a Word Error Rate (WER) of approximately 1.3% for seen speakers and 8.6% for unseen speakers. For the LRW (Lip Reading in the Wild) dataset, which presents a more challenging real-world scenario, the model achieved a WER of 23.8%. While effective, LSTMs can face issues like slow training and vanishing gradients for very long sequences.

All You Need Is a Lip Sync Expert for Speech to Lip Generation in the Wild Prajwal et al. [14], 2020 paper discusses a model designed specifically for real-world application to do lip syncing with an audio-visual synchronization discriminator along with the facial encoder, trained beforehand to make the mouth look more realistic to what the words say.A temporal convolutional network captures speech-lip alignment over time, and maintains temporal coherence. Training the model on a large-scale wild dataset enables generalization across speakers, languages, and noisy conditions. On benchmarks of LRS2 and LRS3 datasets, it delivers state-of-the-art results with improved lip-sync accuracy and naturalness compared to prior methods. Qualitative tests also confirm the model's robustness in handling varied facial movements and head poses. Li et al. [16] proposed speech-driven lip-sync with CNNs that extract spatial audio features and LSTMs modeling temporal dependencies for the architecture focusing on accurate, temporally coherent lip movements. All experiments were performed by training on paired audio-visual datasets and evaluated based on synchronization accuracy and naturalness metrics. Experimental results on the GRID and LRS datasets demonstrate that the proposed method outperforms the baseline models, in terms of higher lip-sync precision and realism, and better speech dynamics over time, since it uses LSTMs. This paper by Thikekar et al. [18] proposes a GAN-based video dubbing architecture that aligns the synchronized audio dub with visual lip movements. A speech input to the generator synthesizes lip movements, while an adversarial training of the discriminator ensures highly realistic synchronization. It was then tested on several datasets to obtain higher synchronization scores as well as more visually realistic results than existing systems. Its ability to generalize across different languages and speakers highlights its practical application for multilingual video dubbing.

III. METHODOLOGY

A. *Datasets*

The LipNet uses the GRID [20] corpus, which comprises 34 speakers (18 men and 16 women), each of whom narrates 1000 sentences from a total of 34,000 video recordings. The videos are synced with high-quality audio and captured at 25 frames per second with a 720x576 pixel resolution. A textual transcription of the sentence that the person in the video speaks is included with every video in the dataset.

In order to produce realistic lip movements that are in sync with audio input.

The primary dataset utilized by Wav2Lip is LRS2, a large-scale dataset with more than 29 hours of audiovisual data that was created by the University of Oxford and includes data sourced from BBC television Broadcasts. LRS2[21] contains unrestricted sentences spoken in natural environments, in contrast to the GRID dataset. Each video in the dataset is synchronized with an audio track in WAV format (16 kHz sampling rate) and is accompanied by accurate transcriptions that match the audio and video frames.

B. *Lip Reading Mechanism*

Lip reading lets visual indicators of speech such as lip movements get translated to meaningful text information. One of the best solutions regarding this has been proposed recently based on the LipNet model, which is primarily based on deep learning techniques for sequential video frame processing. LipNet is mainly based on a STCNN model which includes a combination of spatial and temporal processing. These networks capture precise lip patterns with high efficiency but there would be a drop in efficiency as lip movements may change over time. A Bi-LSTM is then used for fine-tuning the system's temporal dependencies to let it have better context over past and future frames.

The input pipeline of LipNet will ensure proper preprocessing. The system starts with facial detection. Libraries, such as DLib and Haar cascades, are employed to detect and align the face of the speaker in every frame. Then there is isolation, cropping, and normalization of the mouth region so that the size of the input will be uniform. This will remove the background noise and focus only on the significant lip movements. These are the inputs to the STCNNs and Bi-LSTM. CTC loss makes the model handle variable-length inputs. The CTC loss enables the network to predict transcriptions directly from raw video data. This framework is not only more simplified for the training process but also increases the accuracy of the model. Its strength lies in the fact that it can predict sentences at a sentence level as opposed to word-level lip-reading models used traditionally. The contextual variation of spoken sentences is captured by LipNet through analysis of consecutive frames.

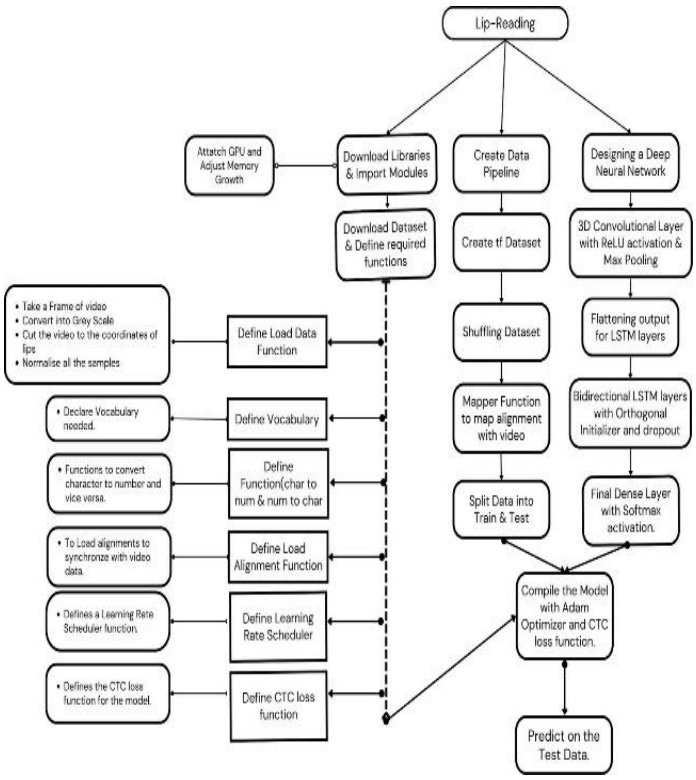

Fig. 1. Workflow of Lip-Reading

The above figure provides a pipelined framework for deep learning in lip reading, starting from the download libraries and datasets. Under data preprocessing: video frames' extraction, resizes, normalize, vocabulary is defined, convert characters, before training the model. The main constituent is

the combination of a 3D CNN with the LSTM layers, which enables the model to capture both temporal and spatial characteristics from video frames. The Connectionist Temporal Classification (CTC) loss is used during training,

making it possible for flexible sequence outputs without the pre-alignment data. The model performs optimization using the optimizer and scheduler for learning. The model would then be evaluated on unseen data after training in order to ensure that the correctness of the prediction is within appropriate limits. Alignment mapping of the video is significant to ensure proper synchronization of the lip movements to the transcriptions.

The pipeline gets the video data processed, hence rightly predicting lip-reading. Using 3D CNNs, LSTMs, and CTC loss, the model can predict words from visual input, hence useful in other applications such as speech recognition in video.

C. Lip Syncing Mechanism

Lip syncing is the process of synchronizing lip movements with audio. Wav2Lip solves this problem by using the Conditional Generative Adversarial Network (cGAN) framework. The cGAN architecture is divided into three modules known as Face Encoder, Audio Encoder, and Face Decoder. These modules ensure that the generated lip movements are oriented with the input audio while preserving the facial attributes of the speaker. The Face Encoder processes reference video frames to extract facial features. This ensures the visual consistency of the speaker's identity throughout the generated video. In parallel, it uses 2D convolutional layers to encode input audio in order to capture the audio representations of the audio. This encoding will allow the model to capture the required temporal and audio features important for good lip synchronization. These two encoders are combined in the Face Decoder to produce video frames whose lips are in synchronization.

An important property of Wav2Lip is its use of an evaluation discriminator that pre-trained SyncNet performs for quality evaluation. SyncNet considers the temporal frames of video with corresponding audio for ensuring that lip movements are synchronized with speech in exact timing. The model optimizes reconstruction loss for pixel accuracy, sync loss for better alignment of audio-visual, and adversarial loss for natural frame generation to ensure that output has high-quality visuals, accurate time, and realistic appearance.

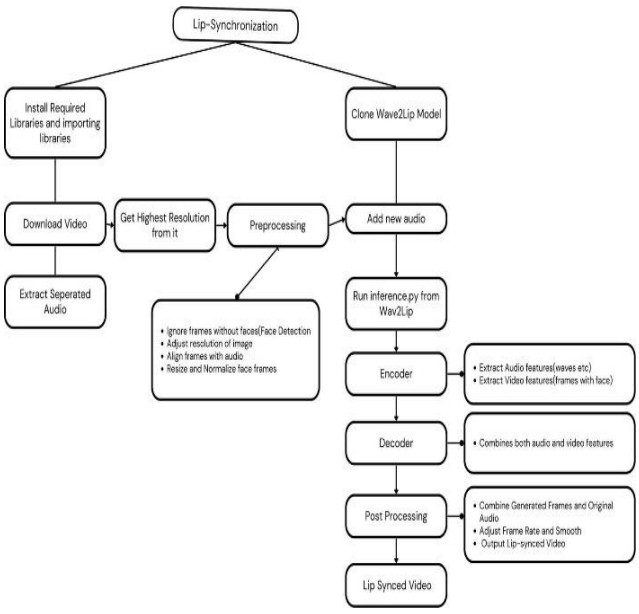

Fig. 2. Workflow of Lip-Synchronization

The above figure describes a pipeline for lip synchronization using the Wave2Lip model. It starts with the cloning of the Wave2Lip repository and the installation and importing of required libraries, downloading a video file, extracting its audio, and accounting for adding new audio to the video before proceeding to do any inference. During preprocessing, frames without any visible faces are ignored using a face detection mechanism. The face frames detected are resized, normalized, and aligned to audio for perfect

matching. Along with video features such as waveforms and mel-spectrograms, the audio features are extracted from the sample. The encoder fetches audio and video features one after another, gathering information to allow synchronization.

It then allows the decoder to gather those and create frames, so that perfect alignment of lips is done on audio. It involves taking a frame generated to the original audio and getting it framed while ensuring transitions are smooth. This finally gives out the output-a video with lip-syncing and accurate synchronization. Advanced feature extraction, inference mechanisms, and post-processing techniques will be used for all high-quality requirements of audio-visual synchronization that applications may mandate.

D. Integration

The system offers a dashboard where users browse through each of the models provided and select their desired one. For LipNet, users upload their video file into the system which then processes for extracting, resizing, and normalizing frames of the videos to prepare those frames for its analysis. Bi-LSTMs, a subset of recurrent neural networks, process sequential input for LipNet. These networks are very good at identifying the temporal patterns of lip movements, which allows them to predict words of the video with great accuracy, eventually providing a transcription of the spoken words - ideal for generating subtitles or text outputs. Wav2Lip addresses lip synchronization.

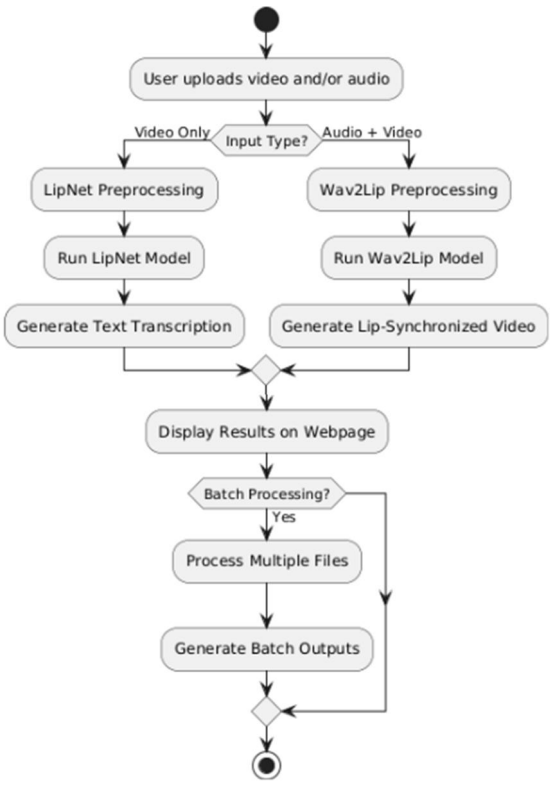

Fig. 3. Integration of Lip-Reading and Lip-Synchronization

Users can upload a video file with an audio file, or simply upload a video that they let the system generate synthesized audio on the same. Wav2Lip processes the video frames and

converts the accompanying audio into mel-spectrograms that are used to align audio with the lip movements. The model synchronizes audio with video frames to generate a seamless output in which lip movements perfectly match the words being spoken. This module is most useful for using these models in tasks such as dubbing or fixing mismatched audio-visual content because the system is designed in a modular fashion where both models operate independently. In this way, the user can opt for lipreading or lip synchronization based on his requirements without any redundancy in functionality.

The independent operation of the two models ensures that each one does its job with optimal efficiency. LipNet focuses on accurate transcription, while Wav2Lip handles precise audio-visual synchronization. The interface is built with user-friendliness in mind, which gives a nice, intuitive way to interact with the models. Through this homepage, users will be educated on LipNet and Wav2Lip capabilities and then easily upload their files to begin processing. For LipNet, the system previews the processed video frames and shows the text transcription. For Wav2Lip, the system displays the synchronized video, so the perfect alignment between audio and lip movements is brought out.

Preprocessing is significant in both the models. For LipNet, frames of video are extracted and normalized so that the model can only focus on the lip movements to provide an accurate transcription. For Wav2Lip, preprocessing includes the conversion of audio into mel-spectrograms, which are then aligned with the video frames to create synchronization. It is the critical stage where the output of both models is of quality. The independence of LipNet and Wav2Lip ensures that one model can be updated or modified without affecting the other. The system is designed to handle real-world challenges, such as low-quality audio or poor lighting in videos, and can still deliver reliable results. This makes the system suitable for tasks that include content creation, tools for accessibility, and real-time communication. This allows one to bind both LipNet and Wav2Lip into one system so that users could switch between lipreading and lip synchronization based on their requirements, creating a more versatile yet efficient solution for an array of operations.

The system will require the following software: Python 3.9.12 as the primary programming language for out interface, deep learning frameworks such as PyTorch or TensorFlow, OpenCV for image and video processing, NumPy and SciPy for numerical computations, and Torchvision for vision-related tasks. Extra dependencies include Ffmpeg for LipSync audio-visual processing, matplotlib and related functions for data plots, imageio for image and video I/O, librosa for audio analysis, torch to implement models, tqdm for tracking progress, numba for numeric functions, gdown for downloading files, and Streamlit for creating the interface. The system utilizes various pretrained models like Wav2Lip for lip synchronization and LipNet for lip reading, thereby ensuring a very efficient and accurate performance in their various functions.

IV. RESULTS

Our integrated system utilizing LipNet achieved a Word Error Rate (WER) of 11.4% and a Character Error Rate (CER) of 6.4% on unseen speakers in the GRID corpus, surpassing the baseline models in sentence-level lipreading accuracy and the Wav2Lip module demonstrated superior lip synchronization performance with an LSE-D of 6.512 and an LSE-C of 7.490 on the LRW dataset, showcasing its effectiveness in generating accurate lip movements synchronized to audio.

Table 2.1 results of LipNet

Method	Dataset	Size	Output	Accuracy
Fu et al. (2008)	AVICAR	851	Digits	37.9%
Hu et al. (2016)	AVLetter	78	Alphabet	64.6%
Papandreou et al. (2009)	CUAVE	1800	Digits	83.0%
Chung & Zisserman (2016a)	OuluVS1	200	Phrases	91.4%
Chung & Zisserman (2016a)	OuluVS2	520	Phrases	94.1%
Chung & Zisserman (2016a)	BBC TV	>400000	Words	65.4%
Gergen et al. (2016)	GRID	29700	Words	86.4%
LipNet	GRID	28775	Sentences	95.2%

Table 2.2 results of Wav2Lip

Method	LRW [8]			LRS2 [1]			LRS3 [3]		
	LSE-D ↓	LSE-C ↑	FID ↓	LSE-D ↓	LSE-C ↑	FID ↓	LSE-D ↓	LSE-C ↑	FID ↓
Speech2Vid [17]	13.14	1.762	11.15	14.23	1.587	12.32	13.97	1.681	11.91
LipGAN [18]	10.05	3.350	2.833	10.33	3.199	4.861	10.65	3.193	4.732
Wav2Lip (ours)	**6.512**	**7.490**	**3.189**	**6.386**	**7.789**	**4.887**	**6.652**	**7.887**	**4.844**

Our Streamlit-based interface combines Wav2Lip to synchronise lip movements with audio and LipNet to predict text from video frames. Users can easily submit video recordings, process them for lipreading, and produce synchronised outputs with this user-friendly web program.

V. CONCLUSION

In general, with a single application for lip reading and lip synchronization, the results would offer good ground for development, and development could be very rewarding with regards to bettering effectiveness as well as increasing the accessibility. Its most valuable change would have support for different languages. This upgrade will make the system attractive on a global level since it would work with a higher percentage of diverse languages, accents, and dialects for it to be more versatile and varied. Multilingual support may unlock further applications in global industries like video productions and tools in multilingual education which demand a highly precise synchronization of lip recognition and speech detection.

There is the necessity for improving real-time capabilities, and therefore this is an important area of development. The system would be apt at handling live applications like video conferencing, real-time translation, and broadcasting, by reducing latency. This would be invaluable in situations wherein speed and accuracy are required and, thus, be very useful in live news reporting, virtual meetings, and online learning. This may also enhance its role in gaming and virtual reality applications where immediate responsiveness and natural interactions are critical to immersive experiences. Another important step forward is enlarging the dataset upon which the models are trained. With more extensive datasets, such as including possibilities of various facial features, different levels of lighting, unique speaker identification, the system can learn and adapt to a whole range of possible scenarios. It would then be more robust, and capable in challenging environments, making the approach more applicable and useful. It would make sure that it is adaptive toward unusual or very complex use cases, like in forensic video analysis or in case silent or low-quality content videos are part of the data.

One more crucial addition would be its implementation on mobile devices and edge computing platforms. This might allow the running of the application with a less powerful device such as a smartphone or a tablet smoothly and would

thus enhance the use of the tool much more. Hence, it would provide the technology as a portable user-friendly tool, accessible at any time or in any rural location. For instance, in education-based applications, mobile deployment will be helpful because the students can use their personal devices to learn new languages or develop communication skills. The next generations of models might enhance visual quality even more when it comes to lip syncing. Lip movements are rendered as much more realistic and lifelike than current models allow for, guaranteeing high-quality results in any usage scenario. It would help particularly in entertainment areas, especially with video dubbing or the animation of characters where the goal is achieving believable realism visually.

Advanced generative models would allow the generation of realistic avatars for virtual meetings and online games. These future developments can only enhance the integrated software further and bring it to a position where it will be more useful in the widest sectors of industry and activity. Actually, with more support for other languages, the addition of real-time capabilities, large datasets, mobile access, and greater visual realism, this particular set of tools may eventually revolutionize how people communicate, create assistive technologies, and produce multimedia content.

VI. REFERENCES

1] Assael, Y., Shillingford, B., Whiteson, S., & De Freitas, N. (2016). LipNet: End-to-End Sentence-level Lipreading. arXiv: Learning.

[2] Wang, H., Guo, P., Wan, X., Zhou, H., & Xie, L. (2024). Enhancing Lip Reading with Multi-Scale Video and Multi-Encoder. 2024 IEEE International Conference on Multimedia and Expo Workshops (ICMEW), 1-6. https://doi.org/10.1109/ICMEW63481.2024.10645400.

[3] Zhang, T., He, L., Li, X., & Feng, G., 2021. Efficient End-to-End Sentence-Level Lipreading with Temporal Convolutional Networks. Applied Sciences. https://doi.org/10.3390/app11156975.

[4] Stafylakis, T., & Tzimiropoulos, G., 2017. Combining Residual Networks with LSTMs for Lipreading. ArXiv, abs/1703.04105. https://doi.org/10.21437/Interspeech.2017-85.

[5] Li, D., Gao, Y., Zhu, C., Wang, Q., & Wang, R. (2023). Improving Speech Recognition Performance in Noisy Environments by Enhancing Lip Reading Accuracy. Sensors (Basel, Switzerland), 23. https://doi.org/10.3390/s23042053.

[6] Zhu, M., Wang, Q., & Luo, J. (2019). Lip-Reading Based on Deep Learning Model. Trans. Edutainment, 15, 32-43. https://doi.org/10.1007/978-3-662-59351-6_4.

[7] Fenghour, S., Chen, D., Guo, K., Li, B., & Xiao, P. (2021). Deep Learning-Based Automated Lip-Reading: A Survey. IEEE Access, 9, 121184-121205. https://doi.org/10.1109/ACCESS.2021.3107946.

[8] Mesbah, A., Berrahou, A., Hammouchi, H., Berbia, H., Qjidaa, H., & Daoudi, M. (2019). Lip reading with Hahn Convolutional Neural Networks. Image Vis. Comput., 88, 76-83. https://doi.org/10.1016/J.IMAVIS.2019.04.010.

[9] Sri, N., Akhil, R., Prasad, S., Jayanth, V., & Jyothi, K. (2023). Lip Reading Using Neural Networks and Deep Learning. INTERANTIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT. https://doi.org/10.55041/ijsrem18765.

[10] Chung, J., & Zisserman, A. (2017). Lip Reading in Profile. https://doi.org/10.5244/C.31.155.

[11] Journal, I. (2022). Lip Reading using Convolutional Neural Network. INTERNATIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT. https://doi.org/10.55041/ijsrem13521.

[12] Fernandez-Lopez, A., & Sukno, F. (2018). Survey on automatic lip-reading in the era of deep learning. Image Vis. Comput., 78, 53-72. https://doi.org/10.1016/j.imavis.2018.07.002.

[13] Fernandez-Lopez, A., & Sukno, F. (2018). Survey on automatic lip-reading in the era of deep learning. Image Vis. Comput., 78, 53-72. https://doi.org/10.1016/j.imavis.2018.07.002.

[14] Prajwal, KR, Mukhopadhyay, R, Namboodiri, V & Jawahar, CV 2020, A Lip Sync Expert Is All You Need for Speech to Lip Generation In the Wild. in 28th ACM International Conference on Multimedia (ACM MM). Association for Computing Machinery, Seattle, USA, pp. 484–492. https://doi.org/10.1145/3394171.3413532 DOI: 10.1145/3394171.3413532

[15] Kadam, A., Rane, S., Mishra, A., Sahu, S., Singh, S., & Pathak, S. (2021). A Survey of Audio Synthesis and Lip-syncing for Synthetic Video Generation. EAI Endorsed Trans. Creative Technol., 8, e2. https://doi.org/10.4108/EAI.14-4-2021.169187.

[16] Li, X., Wang, X., Wang, K., & Lian, S. (2021). A Novel Speech-Driven Lip-Sync Model with CNN and LSTM. 2021 14th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), 1-6. https://doi.org/10.1109/CISP-BMEI53629.2021.9624360.

[17] Guan, J., Zhang, Z., Zhou, H., Hu, T., Wang, K., He, D., Feng, H., Liu, J., Ding, E., Liu, Z., & Wang, J. (2023). StyleSync: High-Fidelity Generalized and Personalized Lip Sync in Style-Based Generator. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 1505-1515. https://doi.org/10.1109/CVPR52729.2023.00151.

[18] Thikekar, A., Menon, R., Telge, S., Tolamatti, G., & L, P. (2022). Generative Adversarial Networks based Viable Solution on Dubbing Videos With Lips Synchronization. 2022 6th International Conference on Computing Methodologies and Communication (ICCMC), 1671-1677. https://doi.org/10.1109/ICCMC53470.2022.9753796.

[19] Shalev, Y., & Wolf, L. (2020). End to End Lip Synchronization with a Temporal AutoEncoder. 2020 IEEE Winter Conference on Applications of Computer Vision (WACV), 330-339. https://doi.org/10.1109/WACV45572.2020.9093490.

[20] GRID Dataset https://paperswithcode.com/dataset/grid

[21] Yu, J., Zhang, S., Wu, J., Ghorbani, S., Wu, B., Kang, S., Liu, S., Liu, X., Meng, H., & Yu, D. (2020). Audio-Visual Recognition of Overlapped Speech for the LRS2 Dataset. ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 6984-6988. https://doi.org/10.1109/ICASSP40776.2020.9054127.