AI DRIVEN LIP-READING SYSTEM FOR ASSISTIVE COMMUNICATION

Rashaz Rafeeque
Apex Institute of Technology (CSE)
Chandigarh University
Punjab, India
rashazrafeeque@gmail.com

Jeevan A J
Apex Institute of Technology (CSE)
Chandigarh University
Punjab, India
jeevanaj2003@gmail.com

Rhishitha T S
Apex Institute of Technology (CSE)
Chandigarh University
Punjab, India
rhishithats002@gmail.com

Dr. Preet Kamal
Apex Institute of Technology (CSE)
Chandigarh University
Punjab, India
preetkgarcha@gmail.com

Abstract*— The potential of new assistive communication systems has been made possible by deep learning models' enhanced performance efficiency and precise lip-reading capabilities. People who have speech impairments use traditional communication methods which operate both slowly and with limited accessibility to produce ineffective communication outcomes. The research developed a deep learning system to convert spoken lip movements into audible voice signals for speech communication with people who have speech problems. The LipNet model architecture provides processing of silent video inputs through its Convolutional Neural Networks (CNNs) which perform feature extraction and Recurrent Neural Networks (RNNs) for sequence prediction through Connectionist Temporal Classification (CTC). Natural language processing technologies facilitate the text processing modules which identify both errors along with suitable contextual terms to generate precise and understandable transcriptions. The method shows remarkable potential to transform speech conversion into a faster and more accessible system that delivers accurate contextual results.

***Keywords*—LipNet, Recurrent Neural Network, Visual Speech Recognition, Text-to-Speech, Convolutional Neural Network, GRID Corpus.**

I. INTRODUCTION

Human communication becomes challenging for individuals with speech impediments because they face difficulties in expressing themselves effectively. Traditional assistive communication systems that include sign language together with text-based devices and speech therapy fail to provide efficient solutions which are accessible at all times. Through AI and deep learning processing Noirbec enables developers to design sophisticated assistive technology that works excellently for AI-based lip-reading needs. Users can interpret spoken language when acoustic signals are unavailable through lip reading which also goes by the name of Visual Speech Recognition (VSR). The research established LipNet as an essential tool because Convolutional Neural Networks and Recurrent Neural Networks have helped in Connectionist Temporal Classification automation through the extraction of features and sequence prediction which produces more robust performance than traditional lip-reading methods limited by accuracy and speaker factor variation.

The proposed research presents a lip-reading system powered by AI which generates audible speech from silent lip expressions to enhance speech communication for those who are impaired. The system unifies three components making up LipNet for lip-reading and Natural Language Processing (NLP) text refinement together with Text-to-Speech (TTS) synthesis to create efficient communication technology. The system supports people with speech problems through a real time and independent communication. Thus, the system will increase accessibility in everyday life and provide equal participation in studies and work for everyone. The system will support people with speech problems, non-verbal conditions, neurological problems and people having surgery or trauma related to speech, by providing an accessible and real-time communication.

Real-time application of this technology helps users stay accessible through social inclusion while offering independence because it solves a longstanding communication challenge.

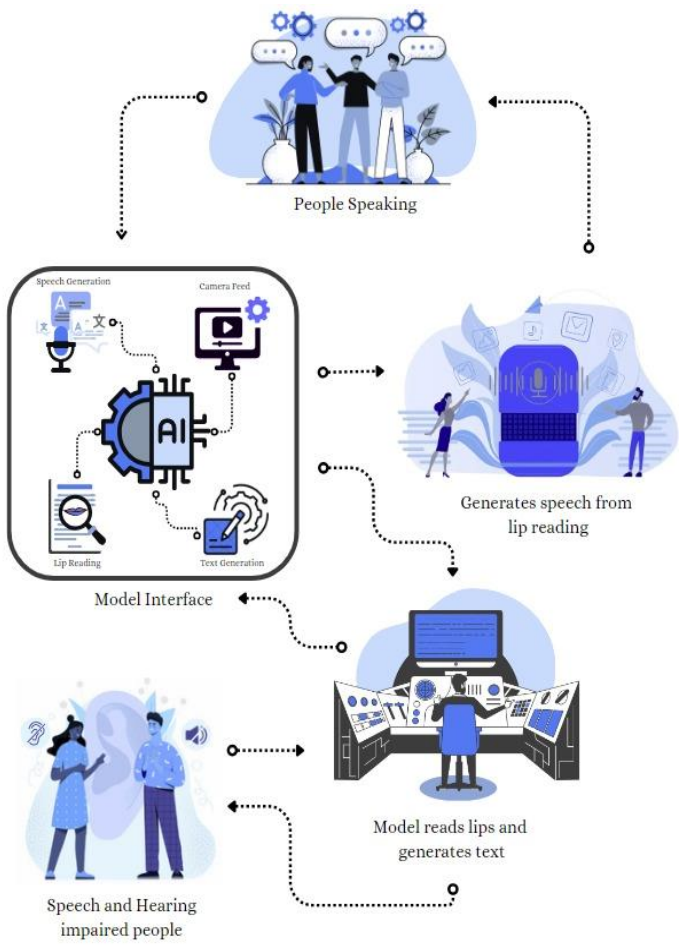

Figure 1: End-to-end Process Flowchar

II. LITERATURE SURVEY

A. LIP READING AND VISUAL SPEECH RECOGNITION:

Through visual speech recognition (VSR) methods known as lip reading one can interpret speech through observation of lip movements while disregarding acoustic signals. Supervisory communication requires this tool to help those who struggle with their speech and hearing abilities.[1] The previous methods for lip reading depended on human-designed features together with rule-based programming but demonstrated limited ability to work consistently for various voices under different circumstances.[3] Deep learning approaches namely Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have delivered improved accuracy together with enhanced efficiency in lip-reading systems.[6] Deep learning models show their ability to extract complex patterns between temporal and spatial dependencies leading to real-time performance under any speaker condition according to studies. AI systems benefit from the integration of these techniques through which they can convert silent lip expressions into significant speech output that serves as the core goal of this project.

B. SPEECH AND HEARING IMPAIRMENTS:

Speech and hearing impairment sufferers use text-based devices along with sign language and speech therapy to communicate because they encounter daily communication problems.[4] The current communication approaches face restrictions which affect both general availability and quick accessibility and immediate engagement. Studies demonstrate how insufficient communication resources produce negative consequences on the social integration opportunities and educational development and work possibilities of people with disabilities.[12] The implementation of AI-powered lip-reading technology enables non-verbal individuals to demonstrate their natural communication abilities by closing communication barriers. Recognized research in assistive technology field demonstrates the necessity for real-time communication products that are intuitive and independent which supports the core targets of this project. [19]

C. LIPNET MODEL:

The lip-reading model LipNet represents a groundbreaking advancement over phoneme-based approaches since it extracts sentences from video frames.[5] The system employs CNNs to extract features while using RNNs with Connectionist Temporal Classification for sequence prediction which makes it an effective solution for continuous speech recognition.[9] Benchmark data evaluation demonstrates that studies of LipNet achieve exceptional recognition accuracy levels while exhibiting enhanced generalization capabilities. The purpose of applying LipNet to this project relies on its capacity to execute sequential processing of lip expressions which leads to time-sensitive accurate silent speech transcription. The project builds a solid basis for lip-reading technology through its adoption of LipNet. [14]

D. ASSISTIVE COMMUNICATION:

The main purpose of assistive communication technologies includes providing daily assistance to people with speech and hearing difficulties in their social interactions.[2] The current Augmentative and Alternative Communication (AAC) devices which include text-to-speech software together with sign language recognition tools represent existing solutions. The current design of these systems depends on human inputs or training by specialists which limits their usefulness for immediate application.[13] Deep learning stands out in AI-powered assistive communication research as it enables creators to build responsive and user-friendly solutions. The project uses advanced lip-reading with speech synthesis to provide hands-free communication since it does not require manual input. [18]

E. TEXT-TO-SPEECH SYNTHESIS

The conversion of recognized text to natural-sounding speech functions as a vital part of Text-to-Speech (TTS) synthesis.[11] TheApiController software produced by wavenet and tacotron deep learning models synthesizes audio that imitates natural human speech patterns. Research advancements in TTS technology have improved voice quality and real-time processing capabilities alongside better intonation functions thus making TTS an ideal technology for lip-reading systems.[7] The combination of TTS functionality grants transcribed text dual viewing options between texts and audio that allows users with non-verbal abilities to speak fluently during conversations.[20]

F. MULTIMODAL LEARNING:

Multimodal learning in AI performance enhancement combines several information types by unifying visuals with texts and audios.[10] Visual speech recognition systems gain improved performance quality when combined with additional text and audio features for better efficiency in lip-reading applications.[13] Research on multimodal artificial intelligence proves that it offers benefits during sensor noise and different speaker variations and partially obstructed mouth conditions. The project utilizes multiple methods to increase speech recognition dependability to enable different real-world environments to connect with each other.[18]

G. COGNITIVE LEARNING IN VISUAL SPEECH RECOGNITION:

Through visualizing speech cues human beings understand and interpret spoken words according to cognitive learning theories.[17] The research contains a description of how the brain processes verbal language by examining face expressions and mouth actions and their relationship to environmental signs. This section investigates the link between cognitive learning theories such as observational learning and pattern recognition with automatic lip-reading system development.[19] The analysis discusses deep learning models which try to mimic human learning approaches through self-supervised and reinforcement learning methods.[20]

H. SEQUENTIAL LEARNING IN LIP READING:

The fundamental requirement for AI-driven lip-reading systems is sequence learning because lip reading relies on identifying patterns in visual speech.[1] The chapter discusses various sequence learning approaches starting with Recurrent Neural Networks (RNNs) through Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRUs) which enable efficient modeling of temporal dependencies within lip movement data.[5] Through sequence learning strategies lip-reading systems reach better accuracy at recognizing continuous speech which makes them appropriate for assistive communication usage.[14]

III. PROBLEM FORMULATION

The essential need to communicate efficiently exists for all people but individuals with impaired speech encounter major obstacles in expressing their thoughts. The traditional communicative tools including sign language alongside text-based systems and speech therapy demonstrate reduced flexibility for real-time use while needing human assistance which reduces their available application. Existing technical solutions do not offer autonomous natural communication modes which prevents non-verbal people from fully participating in social relationships and educational settings as well as professional opportunities.

Visual Speech Recognition (VSR) known as lip reading serves as an important tool in assistive communication when it detects speech through lip movement interpretation. The accuracy levels as well as speaker variability together with environmental elements such as lighting conditions and occlusions present significant hurdles to standard lip-reading systems. Deep learning models specifically Combustional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) show better performance for automatic lip-reading yet their application as real-time practical assistive systems remain restricted.

A gap exists in modern assistive technologies thus this research creates an AI-enabled lip-reading translation system to convert lip movements into spoken texts. The system leverages LipNet for real-time lip-reading, Natural Language Processing (NLP) for text refinement, and Text-to-Speech (TTS) synthesis for speech generation. The research objective establishes a complete automated tool for instant real-time communication which aims to boost accessibility and improve self-reliance and inclusivity for users with speech disabilities.

IV. PROPOSED SYSTEM

The proposed system helps speech-impaired individuals by transforming their lip movements into real-time synthesized speech. The system combines deep learning-based lip reading with natural language processing and text-to-speech (TTS) analysis to create an interface that needs no additional hardware or manual text entry for communication.

Video analysis through LipNet deep learning model detects lip movements in real-time and interprets their actions within the system's main operational setup. Real-time video frames enter the model where it detects mouth movements prior to its neural network training stage that generates written output from these detected movements. A text processing module improves the correctness and coherence of output text following the development of lip movement generation. The text processing results in speech output through a TTS engine which produces natural sounding speech. The available interface enables people with speech communication challenges to convey their thoughts effortlessly while uniting non-verbal expression with spoken words.

Real-time processing allows system users to receive immediate feedback thus making the system appropriate for typical daily interactions. The system provides easy communication accessibility by combining its lip-reading functionalities with Natural Language Processing text improvement features along with high-quality speech generation capabilities. Technical functionalities of this system support meaningful social benefits that enhance its practical application. This system enables people with speech disabilities to become independent in their communication which results in better social participation that enhances their capacity to speak in various personal and professional contexts. Users from various groups can benefit from this tool because it does not require manual input which enlarges its utility range.

V. METHODOLOGY

A real-time lip-reading system needs data preparation and model training and evaluation as mandatory steps to obtain operational capability for real-time inference. The training model utilizes the GRID corpus dataset that contains videos of speakers delivering predefined sentences. The video files with *.mpg extension have corresponding *. align files which provide text transcriptions for each video. The model receives data after multiple preprocessing operations take place. The model extracts video frames with a set interval to maintain time consistency in the data. The system performs face detection through OpenCV while cropping specifically the mouth section to concentrate on essential visual information.

The LipNet model functions as a deep learning system which receives image sequences to generate associated text sequences. The first layer accepts a sequence of frames showing the mouth region. Spatial and temporal features become meaningful because of the application of 3D convolutional layers. The model then uses Bidirectional LSTM layers to track extended lip movement dependencies that help it interpret word sequences. CTC loss serves as the alignment technique to match input frames with output characters because these sequences do not always have equivalent lengths. The model uses Adam optimization to compile its operations with a learning rate that adjusts dynamically.

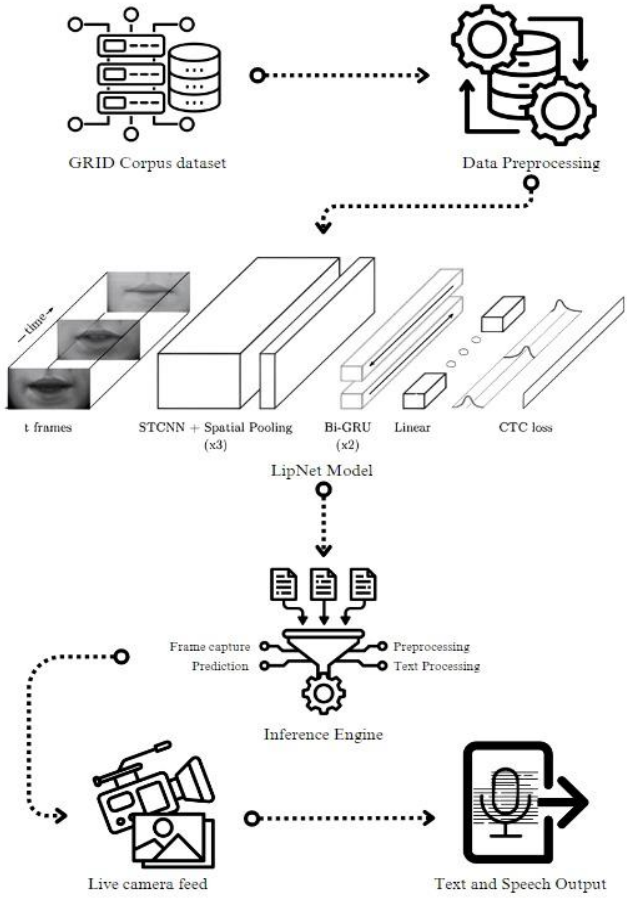

Figure 2: Methodology Flowchart

The model checkpointing mechanism saves weights according to the lowest validation loss for overfitting prevention. The learning rate scheduler controls dynamic adjustments of learning rates for achieving optimal convergence. The training stops automatically through early stopping when validation loss shows no further improvement. OpenCV enables the integration of a real-time lip-reading system which captures video from a webcam. The webcam continuously records video frames which the system uses to detect and crop the mouth region before processing them into the training format. The LipNet model performs inference on preprocessed frames that have been obtained from the training process. The system converts predicted text through gTTS (Google Text-to-Speech) before playing it with pygame. Through this real-time feedback system users obtain voice output which matches their lip movements in real time.

VI. RESULT

The LipNet model trained over 50 training epochs (in Figure 3) with data from the GRID corpus dataset. The model demonstrated a substantial 65.78 training loss which indicated that predicted text sequences differed significantly from actual sequences.

```
500/500 ━━━━━━━━━━━━━━━━━━━━ 465s 931ms/step - loss: 7.1929 - learning_rate: 1.6530e-05
Epoch 49/50
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 192ms/stepstep - loss: 7.08
Original: lay white by z four now
Prediction: lay white by z four now
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
Original: place white with k six please
Prediction: place white with k six please
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
500/500 ━━━━━━━━━━━━━━━━━━━━ 465s 929ms/step - loss: 7.0846 - learning_rate: 1.4957e-05
Epoch 50/50
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 194ms/stepstep - loss: 7.023
Original: bin white by t eight sp please
Prediction: bin white by t eight please
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
Original: set green by c seven soon
Prediction: set gren by seven son
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
500/500 ━━━━━━━━━━━━━━━━━━━━ 466s 932ms/step - loss: 7.0234 - learning_rate: 1.3534e-05
```

Figure 3: Training Epochs

The training process resulted in progressive reduction of loss values. The model achieved effective learning and better generalization through its final epoch training which resulted in training loss of 4.57 and validation loss of 4.21 (in Figure 4).

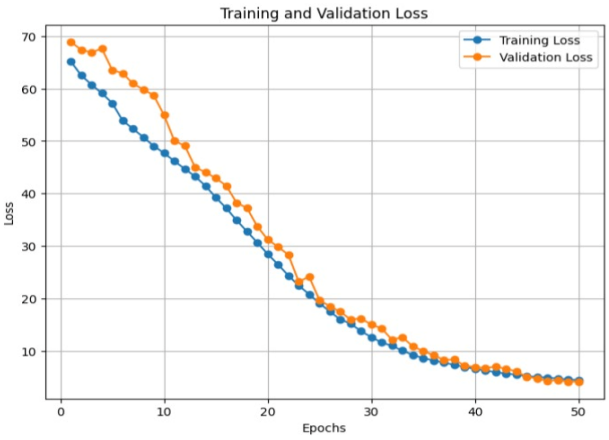

Figure 4: Training and Validation Loss

The developed model was integrated to the real-time lip-reading system which processed live webcam video streams. OpenCV operated in real time to track continuous lip movements for the system and text outputs were generated by the system. The system then converted the final textual output into spoken words through the gTTS (Google Text-to-Speech) tool. Through this process the system received silent lip movements as inputs and produced real-time audible outputs. The combination of real-time video processing with deep learning prediction models and speech generation technology produced an interactive communication system that served people with speech difficulties.

The lip-reading model output (in Figure 5) displays video processing results that present predicted numerical features with accompanying text transcription, video frame dimensions and a "Speak Prediction" button for playing the generated speech which shows the entire lip-to-speech conversion process.

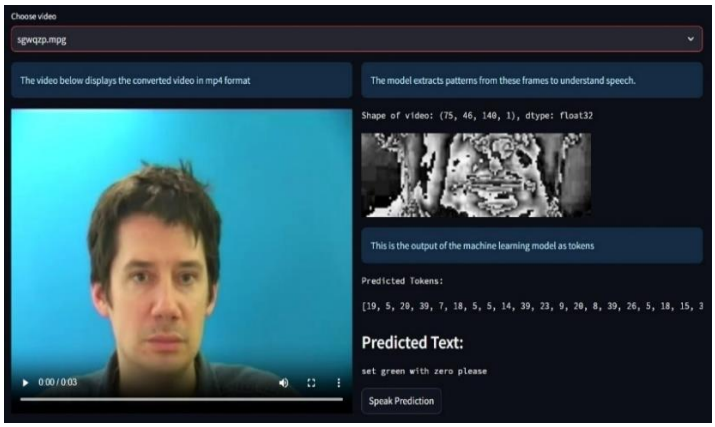

Figure 5: Model Output

The system proved its operational effectiveness in real-life conditions through its quick response mechanism to recorded lip movements. The system proved its practical potential through its performance even though it displayed occasional errors under difficult lighting and speed variations. Further expansion of the system can include dataset expansion along with the integration of facial expressions and context processing to improve accuracy and operational strength.

VII. CONCLUSION

Real-time lip movement processing by the system creates text transcriptions and audible speech output which enables smooth independent communication to users. LipNet processing by the system presents live lip movements to produce both transcription output and spoken audio which grants users accessibility and independence. NLP technology enables the system to extract key meanings from information while TTS technology converts these meaningful inputs into speech that sounds natural to users. The system functioned smoothly during practical tests which demonstrated its efficiency towards enabling users without voice capabilities for communication purposes. This approach is different from standard systems which do not support speech generation because it integrates transcription with real-time speech synthesis to provide users a full communication solution.

Although it currently delivers positive results a few problems persist such as changing lighting conditions alongside variable speech speeds and inaccurate detection of lip movements. The following stage of development requires work on three fundamental factors by expanding training data bases and implementing multimodal processing methods and incorporating facial expressions to detect context. The system requires additional enhancements to reach high accuracy levels and become a dependable tool for various uses. Through the use of AI-driven lip-reading technology human beings with speech difficulties can now communicate inclusively while engaging better with their environment in both social and professional interactions.

VIII. FUTURE SCOPE

The proposed system provides many promising directions for future development and research. One of the areas of development is the training dataset which can include broader range of languages, accents, facial structure and many other features which will help a lot in improving the accuracy of the model among all types of users. Another area of development is the integration of lip movements along with sign languages and hand gestures that will help the system in creating a more natural and expressive speech. Integration of emotion detection that give information of speaker's tone and mood, will make the communication more human-like. Enhancing the system for various mobile platforms and Internet of

Things (IoT) devices will increase the range of accessibility to users in everyday life. Finally, enhancing of system's resistance against many challenges like occlusions, lighting and various camera angles will provide higher performance in real-world conditions. These future improvements will help the system to provide more intelligent, user-friendly and human-like assistive communication solutions for people with speech impairments.

REFERENCE

[1] Exarchos, Themis, Georgios N. Dimitrakopoulos, Aristidis G. Vrahatis, Georgios Chrysovitsiotis, Zoi Zachou, and Efthymios Kyrodimos. "Lip-Reading Advancements: A 3D Convolutional Neural Network/Long Short-Term Memory Fusion for Precise Word Recognition." BioMedInformatics 4, no. 1 (2024): 410-422.

[2] Prajwal, K.R., Mukhopadhyay, R., Namboodiri, V.P. and Jawahar, C.V., 2020. Learning individual speaking styles for accurate lip to speech synthesis. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 13796-13805).

[3] Chandra, A., Paruchuri, C., Karthika, A. and Yogitha, P., 2024. Lip Reading Using Neural networks and Deep Learning. Available at SSRN 4825936.

[4] Paul, Suraj, Dhanesh Lakhani, Divyanshu Aryan, Shudhashekhar Das, and Rohit Varshney. "Lip Reading System for Speech-Impaired Individuals."

[5] Jishnu, T. S., and Anju Antony. "LipNet: End-to-End Lipreading." Indian Journal of Data Mining (IJDM) 4, no. 1 (2024): 1-4.

[6] Kholiev, V. O., and O. Yu Barkovska. "Improved Speaker Recognition System Using Automatic Lip Recognition." Control systems & computers 1 (2024): 38-49.

[7] Shahed, Md Tanvir Rahman, Md Tanjil Islam Aronno, Hussain Nyeem, Md Abdul Wahed, Tashrif Ahsan, R. Rafiul Islam, Tareque Bashar Ovi, Manab Kumar Kundu, and Jane Alam Sadeef. "LipBengal: Pioneering Bengali Lip-Reading Dataset for Pronunciation Mapping through Lip Gestures." Data in Brief (2024): 111254.

[8] Wang, Huijuan, Gangqiang Pu, and Tingyu Chen. "A lip reading method based on 3D convolutional vision transformer." IEEE Access 10 (2022): 77205-77212.

[9] Sarhan, Amany M., Nada M. Elshennawy, and Dina M. Ibrahim. "HLR-net: a hybrid lipreading model based on deep convolutional neural networks." Computers, Materials and Continua 68, no. 2 (2021): 1531-49.

[10] Al-Qurishi, Muhammad, Thariq Khalid, and Riad Souissi. "Deep learning for sign language recognition: Current techniques, benchmarks, and open issues." IEEE Access 9 (2021): 126917-126951.

[11] Guliani, Dhruv, Françoise Beaufays, and Giovanni Motta. "Training speech recognition models with federated learning: A quality/cost framework." In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 3080- 3084. IEEE, 2021.

[12] Reddy, V. Madhusudhana, T. Vaishnavi, and K. Pavan Kumar. "Speech-to-Text and Text-toSpeech Recognition Using Deep Learning." In 2023 2nd International Conference on Edge Computing and Applications (ICECAA), pp. 657-666. IEEE, 2023.

[13] Matsui, Kenji, Kohei Fukuyama, Yoshihisa Nakatoh, and Yumiko O. Kato. "Speech enhancement system using lip-reading." In 2020 IEEE 2nd International Conference on Artificial Intelligence in Engineering and Technology (IICAIET), pp. 1-5. IEEE, 2020.

[14] Prashanth, B. S., MV Manoj Kumar, B. H. Puneetha, R. Lohith, V. Darshan Gowda, V. Chandan, and H. R. Sneha. "Lip Reading with 3D Convolutional and Bidirectional LSTM Networks on the GRID Corpus." In 2024 Second International Conference on Networks, Multimedia and Information Technology (NMITCON), pp. 1-8. IEEE, 2024.

[15] G. C, R. J. D, S. K. A and S. S. V. P. Reddy, "AI Lip Reader Detecting Speech Visual Data with Deep Learning," 2024 4th International Conference on Intelligent Technologies (CONIT), Bangalore, India, 2024, pp. 1-6, doi: 10.1109/CONIT61985.2024.10627537.

[16] E. de la Guía, V. L. Camacho, L. Orozco-Barbosa, V. M. Brea Luján, V. M. R. Penichet and M. Lozano Pérez, "Introducing IoT and Wearable Technologies into Task-Based Language Learning for Young Children," in IEEE Transactions on Learning Technologies, vol. 9, no. 4, pp. 366-378, 1 Oct.-Dec. 2016, doi: 10.1109/TLT.2016.2557333.

[17] Mevlüde Akdeniz, Fatih Özdinç, Maya: An artificial intelligence based smart toy for preschool children, International Journal of Child-Computer Interaction, Volume 29, 2021, 100347, ISSN 2212-8689.

[18] Khondaker A. Mamun, Rahad Arman Nabid, Shehan Irteza Pranto, Saniyat Mushrat Lamim, Mohammad Masudur Rahman, Nabeel Mahammed, Mohammad Nurul Huda, Farhana Sarker, Rubaiya Rahtin Khan, Smart reception: An artificial intelligence driven bangla language based receptionist system employing speech, speaker, and face recognition for automating reception services, Engineering Applications of Artificial Intelligence, Volume 136, Part A, 2024, 108923, ISSN 0952-1976

[19] Amara, K., Boudjemila, C., Zenati, N., Djekoune, O., Aklil, D., & Kenoui, M. (2022). AR Computer-Assisted Learning for Children with ASD based on Hand Gesture and Voice Interaction. IETE Journal of Research, 69(12), 8659–8675.

[20] arXiv:2401.05459 (cs) [Submitted on 10 Jan 2024 (v1), last revised 8 May 2024 (this version, v2)] Personal LLM Agents: Insights and Survey about the Capability, Efficiency and Security Yuanchun Li, Hao Wen, Weijun Wang, Xiangyu Li, Yizhen Yuan, Guohong Liu, Jiacheng Liu, Wenxing Xu, Xiang Wang, Yi Sun, Rui Kong, Yile Wang, Hanfei Geng, Jian Luan, Xuefeng Jin, Zilong Ye, Guanjing Xiong, Fan Zhang, Xiang Li, Mengwei Xu, Zhijun Li, Peng Li, Yang Liu, Ya-Qin Zhang, Yunxin Liu.