IAES International Journal of Artificial Intelligence (IJ-AI)
Vol. 15, No. 6, October 2026, pp. 5031~5039
ISSN: 2252-8938, DOI: 10.11591/ijai.v15.i6.pp5031-5039

Deep learning approaches for lip movement and gesture recognition in dialog generation: a comprehensive review

Mohanaradhya, Prashanth Keshavamurthy, Preethika
Master of Computer Applications, RV College of Engineering, Bengaluru, India

Article Info

Article history:

Received Oct 1, 2025
Revised Aug 24, 2026
Accepted Sep 20, 2026

Keywords:

Deep learning
Gesture recognition
Human-computer interaction
Lip reading
Multimodal systems
Visual speech recognition

ABSTRACT

This literature review explores recent advancements in lip movement and gesture recognition system aimed at enhancing human-computer interaction (HCI) and dialog generation. Unlike existing surveys that focus on isolated modalities, this review systematically analyzes multimodal frameworks, including visual speech recognition, gesture-based interaction, emotion-aware systems, and hybrid convolutional neural network-recurrent neural network (CNN-RNN) architectures. Recent advances involving transformers, diffusion models, and neural radiance fields (NeRFs) are critically examined with respect to accuracy, real-time performance, and deployment feasibility. The review highlights methodological trends, comparative performance insights, and practical deployment challenges, particularly in signer independence, multilingual generalization, and edge-device constraints. By consolidating fragmented research directions into a unified analysis, this work identifies key research gaps and future pathways for building robust, inclusive, and real-world deployable dialog generation systems. This review reveals steady progression of gesture recognition and lip-reading systems with accuracy of (91% – 95%) to (98.76% – 99%) from traditional feature-based approaches to advanced deep learning frameworks.

This is an open access article under the CC BY-SA license.

Corresponding Author:

Mohanaradhya
Master of Computer Applications, RV College of Engineering
RV Vidyanikethan Post, Mysuru Road, Bengaluru, Karnataka, India
Email: mohanaradhya@rvce.edu.in

1. INTRODUCTION

Human-computer interaction (HCI) has evolved significantly with the integration of computer vision and deep learning, enabling more natural and accessible modes of communication. Among these, lip movement recognition and hand gesture recognition have gained prominence for assisting hearing- and speech-impaired individuals and for enabling interaction in noise-free or hands-free environments. Recent advances in convolutional neural networks (CNNs) have made it possible to automatically learn discriminative visual features from lip and hand movement data, improving recognition accuracy and system robustness.

Deep learning-based lip movement detection systems have demonstrated the feasibility of converting visual lip movements directly into text without relying on audio signals. By leveraging CNN architectures, recent approaches have achieved effective lip localization, feature extraction, and character-level prediction, showing strong performance even under challenging visual conditions [1]. Such systems highlight the growing potential of visual-only speech understanding in real-world applications.

Hand gesture recognition has similarly benefited from deep learning, particularly CNN-based models trained on static and dynamic gesture datasets. Vision-based gesture recognition systems using deep neural networks have shown high accuracy in recognizing predefined gesture sets and have been applied to sign language interpretation, virtual interaction, and assistive technologies [2], [3]. These studies emphasize

the role of robust feature learning in handling variations in hand shape, orientation, and lighting conditions. To further enhance gesture recognition performance, feature-level integration strategies have been explored. By combining multiple feature groups extracted from gesture images, improved discrimination between similar gestures has been achieved, addressing challenges such as inter-class similarity and background complexity [4]. These techniques demonstrate that complementary visual features can significantly improve recognition outcomes when integrated effectively.

In parallel, sequence modeling approaches have been applied to lip recognition tasks to capture temporal dependencies in lip movements. Recurrent neural network (RNN) variants, particularly long short-term memory (LSTM) models, have been shown to effectively model the temporal evolution of lip shapes during speech, improving recognition accuracy in continuous lip-reading scenarios [5]. Temporal modeling is essential for understanding speech dynamics beyond isolated frames. Visual lip movement information has also been exploited for biometric applications such as automatic speaker recognition. By analyzing distinctive lip motion patterns, visual cues have been shown to complement traditional speaker recognition systems, especially in environments where audio signals are unreliable or unavailable [6]. This demonstrates the broader applicability of lip movement analysis beyond speech transcription.

Recent research has also focused on vision-based hand gesture recognition techniques for intelligent systems, emphasizing real-time performance and practical deployment. These systems rely on camera-based input and deep learning models to recognize gestures without the need for wearable sensors, making them suitable for natural interaction scenarios [7]. Such approaches contribute to scalable and user-friendly HCI solutions. Advancements in real-time gesture recognition have further enabled the integration of gesture understanding with higher-level tasks such as language generation. Hybrid CNN–LSTM frameworks have been proposed to recognize gestures and generate corresponding textual or semantic representations, enhancing the expressiveness of gesture-based interfaces [8]. These systems move beyond simple classification toward meaningful interaction.

Multimodal approaches have extended gesture recognition by incorporating emotional and expressive cues. Emotion-aware sign language recognition systems that jointly analyze hand gestures and facial expressions have demonstrated improved semantic understanding, particularly in expressive communication contexts such as sign language [9]. This highlights the importance of multimodal correlation in natural communication. More recently, multimodal generative frameworks have been introduced to synthesize co-speech gestures aligned with conversational content. By modeling the relationship between speech and gesture using advanced diffusion-based architectures, these approaches enable realistic gesture generation synchronized with spoken language, contributing to more natural human–machine interaction [10].

2. THEMATIC ANALYSIS OF SELECTED LITERATURE

To systematically analyze the existing body of work, the selected literature is organized into thematic categories based on methodological approaches, application domains, and technological advancements. This thematic classification enables a structured comparison of gesture recognition, lip-reading, and multimodal interaction systems, highlighting common trends and distinct contributions. Organizing the studies in this manner also facilitates the identification of research gaps and emerging directions relevant to intelligent human–computer interaction.

2.1. Visual speech recognition and lip-reading techniques

Visual speech recognition has been extensively explored using computer vision and deep learning techniques to interpret lip movements in silent or noisy environments. Early works focus on extracting geometric and appearance-based features from lip regions, followed by classification using CNNs to recognize spoken words without relying on audio input [11]. These approaches demonstrate the feasibility of lip reading for assisting hearing- and speech-impaired individuals.

Further improvements integrate sequential modeling techniques with spatial feature extraction to better capture temporal dependencies in lip movements. By combining convolutional layers with recurrent architectures, such methods improve recognition accuracy for continuous speech scenarios and complex word formations [12]. These models show robustness against variations in lighting and speaker pose.

Additional studies emphasize lightweight and real-time visual speech recognition systems, targeting practical deployment on low-resource devices. Deep learning-based lip-reading models trained on limited datasets still achieve reasonable performance, highlighting their applicability in assistive communication systems [13]. Collectively, these works establish visual speech recognition as a foundational component for multimodal communication systems [14].

2.2. Sign language and hand gesture recognition systems

Hand gesture and sign language recognition play a critical role in non-verbal communication systems. Several studies focus on recognizing static and dynamic hand gestures using deep learning models trained on image and video data, enabling real-time translation of sign language into text [15]. These systems significantly reduce communication barriers for deaf and mute users. Advanced gesture recognition techniques employ multi-feature extraction and machine learning-based classification to enhance recognition accuracy across diverse gesture sets. Fusion-based methods leverage spatial and motion features to improve robustness under varying environmental conditions [16]. Such approaches demonstrate improved generalization across users and backgrounds. Real-time gesture recognition has also been achieved using object detection-based deep learning models, allowing fast and accurate identification of hand gestures from live video streams [17]. These solutions highlight the effectiveness of deep learning in translating complex hand movements into meaningful representations for assistive applications.

2.3. Audio-visual and multimodal recognition approaches

Multimodal systems combining visual, audio, and sensor-based inputs have gained attention for improving recognition reliability. By jointly processing lip movements and hand gestures captured through mobile device sensors, such systems enhance recognition accuracy in real-world scenarios [18]. This multimodal integration compensates for the limitations of single-modality systems. Machine learning-based gesture recognition frameworks further contribute to multimodal understanding by utilizing pattern recognition techniques to analyze motion and spatial features [19]. These systems demonstrate improved adaptability across diverse interaction contexts. Comprehensive analyses of intelligent gesture recognition techniques provide insights into algorithmic trends, datasets, and evaluation strategies used across multimodal systems. Survey-based studies consolidate existing knowledge and highlight challenges such as scalability, real-time performance, and cross-user variability [20], making them valuable for comparative evaluation.

2.4. Talking face and speech-driven visual synthesis

Recent advancements extend beyond recognition toward generating realistic visual speech representations. Speech-driven facial synthesis models utilize neural rendering techniques to produce high-fidelity talking faces that are synchronized with spoken or predicted speech content [21]. These approaches enable visually expressive avatars for communication systems. Synchronization between audio cues and facial movements is critical for realism in generated talking faces. Advanced synchronization-aware models focus on aligning lip motion, facial expressions, and speech timing to produce natural and coherent facial animations [22]. Such methods significantly reduce perceptual artifacts. Neural radiance fields (NeRFs)-based approaches further enhance facial realism by jointly modeling facial expressions and audio guidance. These models achieve improved identity preservation and expressive accuracy, even in limited-data scenarios [23]. These developments open new possibilities for virtual communication interfaces.

2.5. Surveys and comparative

Comparative and survey-oriented studies provide a structured overview of gesture recognition and multimodal interaction systems. These analyses compare traditional machine learning approaches with deep learning-based methods, highlighting improvements in accuracy, robustness, and scalability [24]. Such evaluations guide system design choices. Context-aware gesture generation systems further explore the integration of gesture synthesis with communication agents. By modeling conversational context, these systems generate appropriate co-speech gestures, enhancing the naturalness of human–machine interaction [25]. This direction emphasizes the importance of semantics and timing in multimodal communication. Overall, survey and comparative works consolidate existing research, identify open challenges, and outline future directions for gesture recognition and visual communication systems. They play a crucial role in bridging theoretical advancements with practical deployment.

3. RELATED WORK

The growing field of HCI has been substantially enriched through concurrent advancements in gesture recognition, lip reading, and dialog synthesis systems. While this review focused on a curated set of 20 impactful papers, many additional studies contribute significantly to understanding emerging deep learning techniques, multimodal integration strategies, real-world deployments, and novel interaction paradigms. This section outlines some of the key auxiliary contributions that complement and expand the core themes reviewed earlier.

3.1. Survey and foundational studies

Gesture recognition research was initially analyzed through survey-based studies that reviewed static and dynamic gesture recognition systems. These works discussed feature extraction techniques, classification models, and application areas such as sign language interpretation and human–computer interaction. Limitations related to illumination changes, gesture variability, and real-time performance were clearly identified [26]. Three-dimensional facial motion studies provided biomechanical insights into lip movement during speech articulation. Lip displacement was quantified using spatial coordinates to construct average motion templates for different visemes. These templates serve as reliable references for validating visual speech recognition systems [27]. Geometric feature-based lip-reading systems focused on extracting lip shape parameters such as width, height, and surface area. These compact features enabled efficient representation of spoken words while reducing computational and storage requirements. The results demonstrated that simple geometric descriptors can support effective lip-reading performance [28].

3.2. Vision-based lip and gesture interaction systems

Lip gesture-based interaction systems were developed to enable hands-free human–computer interaction. Facial images captured using standard cameras were processed to locate the mouth region and track lip movements. Neural-network-based classification enabled recognition of predefined lip gestures for assistive interaction applications [29]. Lip movement analysis was further extended to detect inconsistencies between visual speech and audio signals. Temporal misalignment between lip motion and sound was exploited to identify manipulated media content. High detection accuracy demonstrated the reliability of lip-based synchronization analysis [30]. Multiview hand gesture recognition approaches addressed occlusion and viewpoint challenges by generating synthetic gesture views. Adversarial learning techniques were used to enrich training datasets and improve recognition robustness. Experimental evaluations showed significant accuracy improvements in complex scenarios [31].

3.3. Deep learning-based gesture and lip recognition

CNNs-based hand gesture recognition systems achieved high accuracy when trained on predefined gesture sets. These systems demonstrated effective learning of spatial hand features under controlled environments. However, dataset size and gesture diversity were identified as major constraints affecting generalization [32]. Vision-based lip-reading systems employing deep learning techniques focused on assisting hearing-impaired users. Lip regions were analyzed to extract discriminative visual features and map them to spoken words. The studies confirmed the feasibility of visual-only speech recognition for assistive communication [33]. Temporal modeling approaches enhanced lip-reading performance by integrating recurrent architectures with convolutional feature extractors. These systems captured speech dynamics and coarticulation effects across consecutive frames. Improved recognition accuracy was observed for continuous speech scenarios [34].

3.4. Feature fusion and multimodal recognition

Gesture recognition systems based on multialgorithm fusion combined outputs from multiple classifiers to improve recognition stability. Feature-level and decision-level fusion strategies reduced misclassification caused by gesture similarity. The results demonstrated improved robustness compared to single-model approaches [35]. Multialgorithm gesture recognition frameworks further explored the integration of complementary visual features. Texture, shape, and motion descriptors were combined to enhance classification performance. These approaches showed resilience against background complexity and lighting variations [36]. Visual speech recognition systems focused on lip motion analysis for word-level speech understanding. Accurate lip localization and motion tracking were emphasized as critical factors influencing recognition accuracy. Experimental results validated the effectiveness of vision-based speech recognition under controlled conditions [37].

3.5. Multimodal and mobile-based recognition systems

Multimodal recognition systems integrated audio, visual, and sensor data from mobile devices to recognize speech and gestures simultaneously. The use of complementary modalities improved robustness in noisy and unconstrained environments. These systems demonstrated feasibility for real-time mobile applications [38]. Machine learning-based gesture recognition models were designed with an emphasis on real-time performance and computational efficiency. Lightweight architectures were evaluated for interactive systems requiring low latency. The findings highlighted trade-offs between recognition accuracy and processing cost [38]. A large-scale survey analyzed intelligent gesture recognition systems across vision-based, sensor-based, and hybrid approaches. The study compared datasets, feature representations,

classifiers, and evaluation metrics. It concluded that deep learning methods dominate recent research but still face generalization challenges [39].

3.6. Talking face and gesture generation systems

Speech-driven talking-face synthesis systems employed neural rendering techniques to generate realistic facial movements synchronized with speech. These approaches modeled lip motion and facial expressions to achieve high visual fidelity. The results supported applications in virtual avatars and digital communication [39]. Synchronization-aware facial generation models emphasized precise temporal alignment between audio signals and lip movements. Improved synchronization significantly enhanced perceptual realism in generated faces. These systems demonstrated the importance of timing consistency in speech-driven synthesis [40]. NeRFs-based talking-face generation methods utilized audio guidance to produce expressive facial animations. Joint modeling of facial geometry and expression dynamics enabled identity preservation. These systems achieved high-quality synthesis even with limited training data [41]. One-shot talking-face synthesis frameworks explored deformable neural representations to generate high-fidelity facial motion. Minimal input data was required to personalize facial synthesis. The studies highlighted the efficiency and scalability of compact generative models [42].

3.7. Co-speech gesture generation

Lip-guided talking-face generation approaches incorporated visual speech cues to enhance speech–face alignment. Lip-reading information was shown to improve synchronization accuracy and facial realism. These systems reinforced the role of visual speech dynamics in generation tasks [43]. Context-aware co-speech gesture generation systems focused on producing gestures aligned with conversational context. Temporal and semantic cues were incorporated to generate meaningful gestures during interaction. Successful deployment on humanoid robotic platforms demonstrated practical applicability [44].

3.8. Recent multimodal interaction advances

Advanced multimodal interaction systems combined lip movement, gesture, and speech cues to enable natural communication interfaces. These systems demonstrated improved expressiveness and interaction quality through coordinated multimodal outputs. The results emphasized the convergence of recognition and generation paradigms [45]. Recent frameworks explored real-time multimodal interaction using integrated lip and gesture analysis. These systems focused on enhancing accessibility and communication effectiveness. Experimental evaluations highlighted the potential for assistive and intelligent interface applications [46].

Survey-oriented multimodal communication studies reviewed recent advances in lip and gesture-based interaction systems. The analyses highlighted trends in deep learning, multimodal fusion, and generative modeling. These works identified open challenges related to scalability and real-world deployment [47]. Comprehensive reviews of multimodal assistive communication frameworks examined system architectures and evaluation strategies. The studies emphasized the need for unified frameworks capable of robust performance across diverse environments. These insights provide guidance for future multimodal system design [48].

4. COMPARATIVE TABLE OF METHODOLOGIES AND OUTCOME

To visually summarize the comparative performance of the reviewed models, Figure 1 presents an accuracy-based comparison of representative gesture recognition, lip-reading, and multimodal systems discussed in this review. The comparison highlights the effectiveness of hybrid and multimodal architectures over unimodal approaches, particularly in real-time and noisy environments. The bar chart in Figure 1 illustrates the comparative accuracy of various deep learning models and hybrid systems used for gesture recognition, lip reading, and audio-visual speech recognition, based on data extracted from the reviewed literature. Each bar represents a specific system from the selected papers [P1], [P2], and their reported classification accuracy in percentage terms.

The accuracy chart clearly illustrates that modern deep learning architectures, particularly those leveraging multimodal fusion and temporal modeling significantly outperform traditional standalone models. Systems based on you only look once version 5 (YOLOv5) [P13] and audio-visual fusion models [P16] achieved the highest accuracies, reaching up to 99.00% and 98.76% respectively. This suggests that real-time recognition tasks benefit greatly from models that process spatial and temporal features simultaneously.

Additionally, systems integrating attention mechanisms, CNN-RNN hybrids, or optical flow analysis [P1], [P11], [P12], and [P15] consistently achieved accuracies around 97%, demonstrating their robustness in dynamic, real-world conditions. These models show the value of enhancing visual perception with temporal awareness, especially in lip reading and gesture interpretation tasks. Conversely, more

conventional models (e.g., basic CNN or classical feature-based systems like [P2] and [P6]) achieved moderate performance (91% – 95%), highlighting their limitations in handling real-time variability, emotional cues, and signer-independent accuracy.

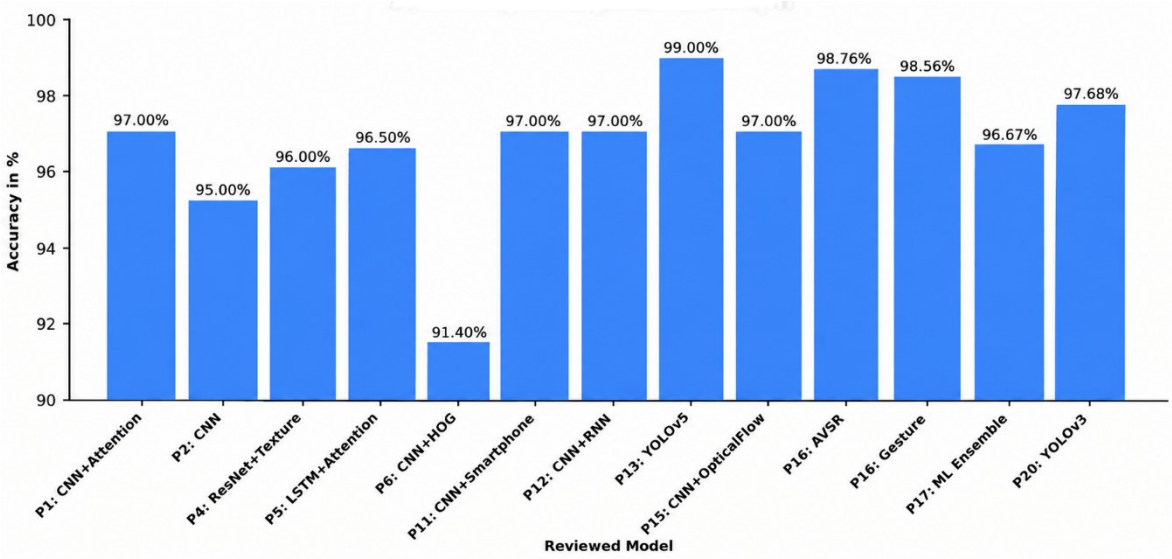

Figure 1. Accuracy comparison (%) of gesture recognition, lip movement recognition, and multimodal deep learning models

In conclusion, the chart supports the following key insights: multimodal and hybrid models outperform unimodal systems, especially in environments with noise, lighting variation, or real-time constraints. Attention mechanisms, sequential modeling (LSTM, bidirectional long short-term memory (BiLSTM)), and CNN-based fusion techniques are crucial for achieving high recognition accuracy. YOLO-based frameworks dominate in real-time detection, thanks to their speed and precision.

5. CONCLUSION

This literature survey shows how gesture recognition and lip-reading systems have steadily improved from conventional feature-based methods to sophisticated deep learning frameworks with accuracies ranging from (91% – 95%) to (98.76% – 99%). Early models based on convolutional and recurrent architectures enabled reliable extraction of spatial and temporal information from visual and audio cues, establishing a strong foundation for automated recognition systems. More recent studies highlight the importance of multimodal integration, where lip movements, hand gestures, and contextual cues are jointly modelled to enhance recognition accuracy, temporal consistency, and interaction naturalness. Such integrated approaches have expanded the applicability of these systems across domains including assistive communication, education, healthcare, virtual environments, and human–robot interaction. Despite these advancements, several technical and practical challenges persist. Many existing systems continue to face limitations in generalization across diverse users, environments, and hardware platforms, particularly under variations in lighting, viewpoint, and occlusion. The reliance on large, labelled datasets also restricts scalability, especially for low-resource sign languages and underrepresented user groups. Future research should therefore prioritize the development of lightweight and computationally efficient models suitable for real-time deployment, along with improved dataset diversity and robust multimodal fusion strategies. In parallel, addressing ethical considerations related to privacy, bias, data security, and responsible data usage will be essential to ensure the inclusive and trustworthy evolution of intelligent human–computer interaction systems.

FUNDING INFORMATION

Authors state no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Mohanaradhya	✓	✓		✓		✓	✓	✓		✓	✓	✓	✓	✓
Prashanth Keshavamurthy		✓		✓				✓		✓		✓		✓
Preethika	✓		✓				✓		✓		✓			

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

INFORMED CONSENT

This review paper does not involve individuals nor any personal identification information that could require any informed consent.

ETHICAL APPROVAL

This paper involves only the study of the publishes research papers and does not involve people or animals. Therefore, it does not seek approval from any institutional review board.

DATA AVAILABILITY

Data availability is not applicable to this paper.

REFERENCES

[1] A. B. Y. Awan, T. Zhang, N. Younas, X. Zhao, M. Fahad, and Y. Geng, "Lips movement detection and conversion into text using deep learning-based CNN model," *Signal, Image and Video Processing*, vol. 19, no. 9, 2025, doi: 10.1007/s11760-025-04178-5.

[2] S. Gupta, P. Bujade, V. Singh, and S. Tiwari, "Hand gesture recognition system using deep learning," *International Journal For Multidisciplinary Research*, vol. 6, no. 3, pp. 1–20, May 2024, doi: 10.36948/ijfmr.2024.v06i03.19602.

[3] T. Chandana, S. C. Venkateswarlu, and V. S. Nagaraju, "Convolutional neural network for hand gesture recognition using 8 different gestures," *International Journal of Scientific Research in Engineering and Management*, vol. 9, no. 5, pp. 1–9, 2025, doi: 10.55041/IJSREM48820.

[4] I. T. Ahmed, W. H. Gwad, B. T. Hammad, and E. Alkayal, "Enhancing hand gesture image recognition by integrating various feature groups," *Technologies*, vol. 13, no. 4, Apr. 2025, doi: 10.3390/technologies13040164.

[5] N. Liang, "LSTM algorithm for human machine interaction lip recognition," *Journal of Computational Methods in Sciences and Engineering*, vol. 25, no. 1, pp. 876–887, Jan. 2025, doi: 10.1177/14727978251321690.

[6] S. Nainan, S. Parmar, and K. Bakade, "Exploiting visual lip movement information for automatic speaker recognition," *Circuits, Systems, and Signal Processing*, vol. 44, no. 9, pp. 6752–6767, Sep. 2025, doi: 10.1007/s00034-025-03113-w.

[7] X. Lou, "Vision-based hand gesture recognition technology," *Applied and Computational Engineering*, vol. 141, no. 1, pp. 54–59, Mar. 2025, doi: 10.54254/2755-2721/2025.21696.

[8] M. Kavitha and S. B. Rajakumari, "Advancing human-computer interaction: real-time gesture recognition and language generation using CNN-LSTM networks," in *2024 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES)*, Dec. 2024, pp. 1–7, doi: 10.1109/ICSES63760.2024.10910557.

[9] H. Ranka, D. Sarda, H. Kunder, A. Rajesh, and A. Kore, "Emotion-aware Indian sign language recognition: a multimodal approach with sign-expression correlation analysis," *Educational Administration: Theory and Practice*, vol. 30, no. 6, pp. 4399–4407, Jun. 2024, doi: 10.53555/kuey.v30i6.7043.

[10] M. H. Mughal, R. Dabral, I. Habibie, L. Donatelli, M. Habermann, and C. Theobalt, "ConvoFusion: multi-modal conversational diffusion for co-speech gesture synthesis," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2024, pp. 1388–1398, doi: 10.1109/CVPR52733.2024.00138.

[11] M. R. Lokesh, C. L. Arpitha, Hemalatha, and K. G. Krupa, "Lip reading using computer vision techniques and deep learning algorithms for deaf and dumb people," *International Journal of Research Publication and Reviews*, vol. 5, no. 5, pp. 1858–1865, May 2024, doi: 10.55248/gengpi.5.0524.1130.

[12] R. R. Patil, D. A. Verma, S. B. Rathod, R. A. Mahajan, and P. Agrawal, "Enhancing lip reading: a deep learning approach with CNN and RNN integration," *Journal of Electrical Systems*, vol. 20, no. 2s, pp. 463–471, Apr. 2024, doi: 10.52783/jes.1367.

[13] A. B. Gavali, G. S. Pooja, K. R. Supriya, and K. A. Rajnandini, "Visual speech recognition using lips movement," *Journal of Signal Processing*, vol. 9, no. 2, pp. 1–7, Aug. 2023, doi: 10.46610/JOSP.2023.v09i02.001.

[14] P. Rathod, A. Yadav, K. Ranva, and P. Ahir, "Visual speech recognition by lip reading using deep learning," *International Research Journal of Modernization in Engineering Technology and Science*, vol. 3, no. 5, pp. 1115–1119, 2021.

[15] T. Fathima, A. Alam, A. Gangwar, D. K. Khetan, and K. Ramya, "Real-time sign language recognition and translation using deep learning techniques," *International Research Journal on Advanced Engineering Hub*, vol. 2, no. 2, pp. 93–97, 2024, doi: 10.47392/IRJAEH.2024.0018.

[16] Y. Zhu, Y. Wang, X. Gao, and L. Liu, "Research on gesture recognition based on multialgorithm fusion," in *Data Science: 9th International Conference of Pioneering Computer Scientists, Engineers and Educators (ICPCSEE 2023)*, 2023, pp. 62–76, doi: 10.1007/978-981-99-5971-6_5.

[17] A. Mujahid *et al.*, "Real-time hand gesture recognition based on deep learning YOLOv3 model," *Applied Sciences*, vol. 11, no. 9, May 2021, doi: 10.3390/app11094164.

[18] D. Ryumin, D. Ivanko, and E. Ryumina, "Audio-visual speech and gesture recognition by sensors of mobile devices," *Sensors*, vol. 23, no. 4, Feb. 2023, doi: 10.3390/s23042284.

[19] M. L. Hyder, S. H. Ahammad, K. Tarun, K. Saikumar, M. V. D. Prasad, and V. S. Reddy, "An adavanced gesture recognition using machine learning," in *2022 International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON)*, Dec. 2022, pp. 1–6, doi: 10.1109/SMARTGENCON56628.2022.10084168.

[20] J. J. O.-Castelo, M. D. L. M. C.-Uriarte, J. A. P.-Fernandez, and R. Ayala, "A survey on intelligent gesture recognition techniques," *IEEE Access*, vol. 10, pp. 87135–87156, 2022, doi: 10.1109/ACCESS.2022.3199358.

[21] B. Nandwana, S. Tazi, S. Trivedi, D. Kumar, and S. K. Vipparthi, "A survey paper on hand gesture recognition," in *2017 7th International Conference on Communication Systems and Network Technologies (CSNT)*, Nov. 2017, pp. 147–152, doi: 10.1109/CSNT.2017.8418527.

[22] H. Popat, S. Richmond, D. Marshall, and P. L. Rosin, "Facial movement in 3 dimensions," *Otolaryngology–Head and Neck Surgery*, vol. 145, no. 1, pp. 24–29, Jul. 2011, doi: 10.1177/0194599811401701.

[23] K. S. Talha *et al.*, "Extracting features point of lip movement for computer-based lip reading system," *International Journal of Mechanical and Mechatronics Engineering*, vol. 14, no. 2, pp. 48–53, 2014.

[24] P. Dalka and A. Czyzewski, "Lip movement and gesture recognition for a multimodal human-computer interface," in *2009 International Multiconference on Computer Science and Information Technology*, Oct. 2009, pp. 451–455, doi: 10.1109/IMCSIT.2009.5352685.

[25] M. Javed, Z. Zhang, F. H. Dahri, A. A. Laghari, M. Krajčík, and A. Almadhor, "Audio–visual synchronization and lip movement analysis for real-time deepfake detection," *International Journal of Computational Intelligence Systems*, vol. 18, no. 1, Jul. 2025, doi: 10.1007/s44196-025-00911-7.

[26] M. Garg, D. Ghosh, and P. M. Pradhan, "Generating multiview hand gestures with conditional adversarial network," in *2021 IEEE 18th India Council International Conference (INDICON)*, Dec. 2021, pp. 1–6, doi: 10.1109/INDICON52576.2021.9691527.

[27] S. Zhu, "Research on the dynamic viseme of the lip shape based on facial motion capture technology," *Frontiers in Neurorobotics*, vol. 16, Sep. 2022, doi: 10.3389/fnbot.2022.922756.

[28] M. Z. Ibrahim and D. J. Mulvaney, "Geometry based lip reading system using multi dimension dynamic time warping," in *2012 Visual Communications and Image Processing*, Nov. 2012, pp. 1–6, doi: 10.1109/VCIP.2012.6410805.

[29] P. Dalka and A. Czyzewski, "Human-computer interface based on visual lip movement and gesture recognition," *International Journal of Computer Science and Applications*, vol. 7, no. 3, pp. 124–139, 2010.

[30] W. Liu *et al.*, "Lips are lying: spotting the temporal inconsistency between audio and visual in lip-syncing deepfakes," in *38th International Conference on Neural Information Processing Systems*, 2024, pp. 91131–91155, doi: 10.52202/079017-2892.

[31] A.-D. Ho and H.-G. Doan, "Quantitation of new arbitrary view dynamic human action recognition framework," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 37, no. 3, pp. 1797–1803, Mar. 2025, doi: 10.11591/ijeecs.v37.i3.pp1797-1803.

[32] G. Park, V. K. Chandrasegar, and J. Koh, "Accuracy enhancement of hand gesture recognition using CNN," *IEEE Access*, vol. 11, pp. 26496–26501, 2023, doi: 10.1109/ACCESS.2023.3254537.

[33] S. Fenghour, D. Chen, K. Guo, and P. Xiao, "Lip reading sentences using deep learning with only visual cues," *IEEE Access*, vol. 8, pp. 215516–215530, 2020, doi: 10.1109/ACCESS.2020.3040906.

[34] P. Ma, Y. Wang, J. Shen, S. Petridis, and M. Pantic, "Lip-reading with densely connected temporal convolutional networks," in *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Jan. 2021, pp. 2856–2865, doi: 10.1109/WACV48630.2021.00290.

[35] T. Kanwal and S. Altaf, "Exploring sensor fusion techniques for enhanced dynamic hand gesture recognition: a comprehensive metadata analysis," *IEEE Sensors Reviews*, vol. 2, pp. 122–156, 2025, doi: 10.1109/SR.2025.3567259.

[36] C. Zhu, J. Yang, Z. Shao, and C. Liu, "Vision based hand gesture recognition using 3D shape context," *IEEE/CAA Journal of Automatica Sinica*, vol. 8, no. 9, pp. 1600–1613, Sep. 2021, doi: 10.1109/JAS.2019.1911534.

[37] C. Yu, X. Wang, and Z. Qian, "Silent speech recognition using visual cascading fusion of tongue-lip movements based on pre-trained and fine-tuned model," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2025, no. 1, Apr. 2025, doi: 10.1186/s13636-025-00403-8.

[38] Z. Peng *et al.*, "SyncTalk: the devil is in the synchronization for talking head synthesis," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2024, pp. 666–676, doi: 10.1109/CVPR52733.2024.00070.

[39] K. Vougioukas, S. Petridis, and M. Pantic, "Realistic speech-driven facial animation with GANs," *International Journal of Computer Vision*, vol. 128, no. 5, pp. 1398–1413, May 2020, doi: 10.1007/s11263-019-01251-8.

[40] J. Wang, X. Qian, M. Zhang, R. T. Tan, and H. Li, "Seeing what you said: talking face generation guided by a lip reading expert," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2023, pp. 14653–14662, doi: 10.1109/CVPR52729.2023.01408.

[41] T. V. T. Nguyen, V. Schmuck, and O. Celiktutan, "Gesticulating with NAO: real-time context-aware co-speech gesture generation for human-robot interaction," in *ICMI'23 Companion: Companion Publication of the 25th International Conference on Multimodal Interaction*, Oct. 2023, pp. 112–114, doi: 10.1145/3610661.3620664.

[42] W. Li *et al.*, "One-shot high-fidelity talking-head synthesis with deformable neural radiance field," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2023, pp. 17969–17978, doi: 10.1109/CVPR52729.2023.01723.

[43] S. J. Park, M. Kim, J. Hong, J. Choi, and Y. M. Ro, "SyncTalkFace: talking face generation with precise lip-syncing via audio-lip memory," *AAAI Conference on Artificial Intelligence*, vol. 36, no. 2, pp. 2062–2070, Jun. 2022, doi: 10.1609/aaai.v36i2.20102.

[44] N. T. V. Tuyen and O. Celiktutan, "It takes two, not one: context-aware nonverbal behaviour generation in dyadic interactions," *Advanced Robotics*, vol. 37, no. 24, pp. 1552–1565, Dec. 2023, doi: 10.1080/01691864.2023.2279595.

[45] Y. Yoon *et al.*, "Speech gesture generation from the trimodal context of text, audio, and speaker identity," *ACM Transactions on Graphics*, vol. 39, no. 6, pp. 1–16, Dec. 2020, doi: 10.1145/3414685.3417838.
[46] S. J. Arora and R. P. Singh, "Automatic speech recognition: a review," *International Journal of Computer Applications*, vol. 60, no. 9, pp. 34–44, Dec. 2012, doi: 10.5120/9722-4190.
[47] C. Zhang, Z. Yu, X. Wang, Z.-J. Chen, C. Deng, and S. Q. Xie, "Exploration of deep learning-driven multimodal information fusion frameworks and their application in lower limb motion recognition," *Biomedical Signal Processing and Control*, vol. 96, Oct. 2024, doi: 10.1016/j.bspc.2024.106551.
[48] W. Zhao, K. Gangaraju, and F. Yuan, "Multimodal perception-driven decision-making for human-robot interaction: a survey," *Frontiers in Robotics and AI*, vol. 12, Aug. 2025, doi: 10.3389/frobt.2025.1604472.

BIOGRAPHIES OF AUTHORS



Mohanaradhya completed Ph.D. from Visvesvaraya Technological University in the year 2018, M.CA. from Indira Gandhi National Open University in the year 2005 and has secured 5th rank in post diploma in Digital Communication Engineering from Electronics and Radar Development Establishment (LRDE) a defense organization. Having experience of more than 18 years in teaching at pre-university level to master's degree. Currently working as assistant professor in the Department of Master of Computer Applications at RV College of Engineering, Bengaluru. Presented and published research papers in international journals and conferences. He can be contacted at email: mohanaradhya@rvce.edu.in.



Prashanth Keshavamurthy holds a Master of Computer Applications degree from Visvesvaraya Technological University, Belagavi, Karnataka, India in 2005. He also received his B.Sc. from Gulbarga University, Karnataka, India in 1999. He is currently an assistant professor at Master of Computer Applications, in RV College of Engineering, Bengaluru, Karnataka, India, since January 2015. His research area is power optimization in WSN. He has published research papers in international journals and conferences. He can be contacted at email: prashanthk@rvce.edu.in.



Preethika holds a Master of Computer Applications from RV College of Engineering, affiliated to Visvesvaraya Technological University, Belagavi, Karnataka, India. Her areas of interest include artificial intelligence, machine learning, and deep learning. She has worked on projects involving human–computer interaction, gesture recognition, and visual speech recognition, with a focus on deep learning-based dialog generation systems. She can be contacted at email: preethika.mca23@rvce.edu.in.