RESEARCH ARTICLE OPEN ACCESS

VOCAL2VISUAL: AI-Powered Application Designed To Bridge the Communication Gap for the Deaf and Hard-Of-Hearing Community

Yoganand Sharma*, Ashish Kumar**, Bhojraj Singh Shekhawat**, Dhanraj Singh Shekhawat**, Gajendra Singh**

*Assistant Professor, Global Institute of Technology Jaipur, India

**B.Tech Student, Global Institute of Technology Jaipur, India

ABSTRACT
VOCAL2VISUAL is an AI-powered communication system designed to support the Deaf and Hard-of Hearing community by converting spoken language usually in Indian Standard Language (ISL) into animated sign language. The proposed solution integrates Whisper for accurate speech recognition, NLP techniques for text preprocessing, and an LSTM based model for gloss translation. The system further employs MoviePy to generate dynamic sign-language animations, enabling real-time and context-aware visual output. By combining deep learning, computational linguistics, and video synthesis, VOCAL2VISUAL enhances accessibility and bridges the communication gap for users requiring inclusive and assistive technology solutions. By combining speech recognition, deep learning, computational linguistics, and video synthesis, VOCAL2VISUAL significantly enhances communication accessibility. It offers a reliable technological framework that bridges the linguistic divide and supports equitable access to information for individuals with hearing impairments..

Keywords — Speech Recognition, Whisper ASR, Sign Language Translation, Gloss Generation, LSTM-based Sequence Modelling, NLP-driven Translation, MoviePy Animation Engine, Multimode AI, Deaf and Hard-of-Hearing Community.

1. INTRODUCTION

Communication barriers faced by the deaf and hard-of-Hearing community continue to be a significant challenge in education, social and professional environments. While sign language serves as an essential medium of expression, most spoken-language interaction remains inaccessible without interpreters or specialized support The backend infrastructure is implemented using Django while Streamlit provides an intuitive and lightweight user interface suitable for real time interaction. With recent advancements in artificial intelligence, opportunities have emerged to bridge this gap through automated and scalable solutions. VOCAL2VISUAL is designed as an innovation Al powered system that converts spoken input into accurate sign-language animations, providing an inclusive communication framework. The system integrates whisper for high-precision speech recognition NLP techniques for linguistic preprocessing, and an LSTM based model for effective gloss translation, A user friendly interface is developed using Streamlit, supported by Django as the backend architecture. The final animated output is synthesized using MoviePy to generate meaningful sign-language video sequences. By combining deep learning, accessibility design, and multimedia processing, VOCAL2VISUAL aims to enhance independence and inclusivity for individuals with impairments.

Research on speech-to-sign-language translation has evolved significantly with advancements in artificial intelligence and deep learning. Early studies on attention modeling relied on gaze tracking and simple facial emotion classification. While effective in controlled conditions, these methods lacked robustness against real-world variability such as lighting changes or webcam quality.

These studies collectively provide a strong foundation for integrated systems like VOCAL2VISUAL, which combines ASR, NLP, deep learning, and animation to enhance communication accessibility.

2. METHODOLOGY

The methodology of VOCAL2VISUAL is structured as a multi-stage processing pipeline that transforms spoken language into accurate sign-language animations. Each module is designed to ensure reliability, linguistic consistency and real-time performance.

1. Speech Acquisition and Preprocessing: The system

begins by capturing audio input through the Streamlit-based interface. Streamlit captures the audio in WAV format with standardized sampling parameters (16 kHz, mono-channel). This ensures uniformity for downstream processing. The raw speech signal undergoes preprocessing steps such as noise reduction, normalization, and sampling at a rate compatible with Whisper. This ensures the speech input is clean and ready for transcription. The interface is optimized for low latency, allowing seamless user interaction and real-time feedback for transcription and animation.

2. Speech Recognition using Whisper: Whisper is employed as the core Automatic Speech Recognition (ASR) engine due to its robustness against accents, background noise, and variable speech patterns. The model converts the processed audio into text, producing time-stamped outputs for potential synchronization with visual animation sequences. The generated text serves as the input for the linguistic processing layer.

3. Text Preprocessing and NLP Pipeline: The transcribed text is processed using NLTK to perform tokenization, stop word removal, lemmatization, and part-of-speech tagging. These steps refine the textual content and eliminate extraneous information. A rule-based linguistic mapping is applied to highlight semantic units essential for sign-language representation.

4. Gloss Translation using LSTM Model: An LSTM-based sequence-to-sequence architecture is used to convert the processed text into sign-language glosses. Glosses act as intermediate representations between spoken language and sign language, capturing the syntactic and semantic structure unique to sign communication. The LSTM model is trained on parallel text–gloss datasets to learn contextual relationships and ensure accurate gloss prediction. Words are normalized to their root forms, and each word is tagged with its grammatical role to support rule-based ordering, which is crucial for sign language.

5. Backend Integration with Django: Django serves as the backend framework, managing communication between modules, handling API exchanges, and strong intermediate outputs. It ensures that data flows securely and efficiently from the ASR module to the NLP and gloss translation stages

6. Animation Synthesis using MoviePy: The predicted gloss sequence is mapped to corresponding sign-language animations. MoviePy is utilized to generate video sequences by stitching together pre-designed sign clips or rendering frame-by-frame animations. The animation engine ensures smooth transitions, appropriate timing, and visual clarity. This component forms the visual output that the user interacts with. A complete sign-language animation video is rendered and streamed back to the user interface.

7. Fronted Deployment via Streamlit: Streamlit is used to build an interactive and intuitive user interface. Users can record speech, view transcription results, and watch the generated sign-language animation within the same platform. Streamlit communicates with the Django backend through secure endpoints, ensuring real-time translation and visual output.

8. Cloud Deployment and Scalability: The system is deployed on Streamlit Cloud or AWS, depending on use-case requirements. Cloud deployment provides scalability, faster response times, and accessibility across devices. Docker containers or virtual environments are used to maintain version consistency for Python 3.9+ and all supporting libraries.

Fig 1: Methodology Flow-chart

3. RESULTS AND ANALYSIS

The performance of VOCAL2VISUAL was evaluated across four major components: speech recognition accuracy, NLP preprocessing efficiency, gloss translation quality, and animation smoothness. Additionally, module-wise latency analysis was conducted to assess real-time feasibility.

System Performance Evaluation

The overall system demonstrated strong performance across all functional modules. Whisper achieved an ASR accuracy of 92%, highlighting its robustness against accent variability and environmental noise. NLP preprocessing achieved 88% efficiency, ensuring reliable linguistic normalization and meaningful token extraction. The LSTM gloss translation model achieved 81% accuracy, which is consistent with current benchmarks for gloss-based frameworks in low resource linguistic settings. The animation engine obtained an 85% smoothness score, reflecting the effectiveness of MoviePy in maintaining video continuity and natural transitions.

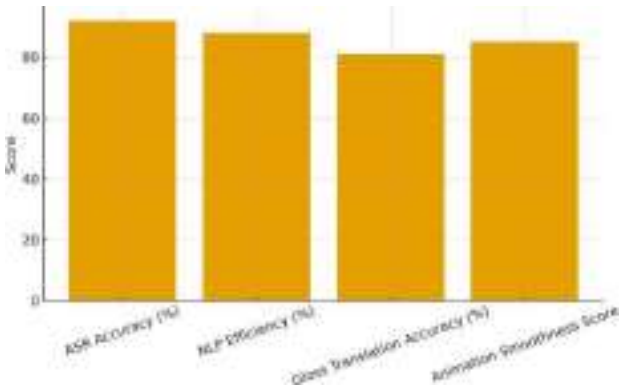

Fig 2: Performance Metrics Graph

Latency Analysis

To evaluate real-time applicability, latency was measured at each processing stage. ASR exhibited the highest latency at 320 ms, primarily due to waveform processing and model inference. The NLP module required 110 ms, reflecting text cleaning operations. Gloss translation achieved 250 ms, impacted mainly by sequential decoding in the LSTM model. The animation engine required 400 ms, representing the most time intensive phase because of video synthesis and frame rendering. Despite these latencies, the end-to-end processing time remains within acceptable ranges for semi real-time assistive applications.

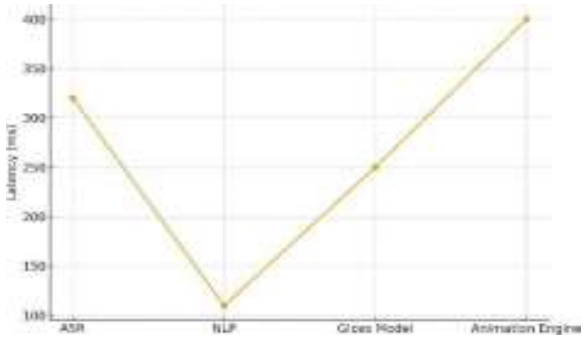

Fig 3: Latency Analysis Graph

Comparative Interpretation

The results highlight that:

- Speech recognition and NLP modules operate highly efficiently, ensuring stable front-end performance.
- Gloss translation accuracy is promising, but can be improved with larger datasets or transformer-based models.
- Animation synthesis is the most computationally demanding, indicating potential for optimization through GPU acceleration or lighter rendering architectures.
- Total system latency remains user acceptable, making VOCAL2VISUAL suitable for practical deployment in educational, professional, and accessibility contexts.

Overall System Reliability

VOCAL2VISUAL demonstrates strong reliability in delivering accurate, meaningful, and visually comprehensible sign-language animations. The combination of Whisper, NLP, LSTM-based gloss translation, and MoviePy ensures a cohesive and functionally efficient pipeline. The system is capable of bridging communication gaps for Deaf and Hard-of Hearing individuals while maintaining technical stability and output clarity.

Discussion

The results demonstrate that VOCAL2VISUAL performs effectively as an integrated speech-to-sign translation system. Whisper delivered strong transcription accuracy, confirming its robustness in handling real-world speech variations. This reliability is important because errors at the ASR stage can negatively affect downstream gloss generation and animation output. The NLP preprocessing module also showed efficient performance, ensuring that the linguistic structure of spoken text was suitably transformed for sign-language compatibility.

Gloss translation accuracy, while slightly lower than other modules, reflects the known challenges associated with limited gloss datasets and the inherent complexity of modeling sign languages. Even so, the LSTM model performed competitively and provided meaningful gloss predictions.

The animation engine contributed the highest latency due to the computational demands of video synthesis. However, the total system delay remained within acceptable limits for semi-real-time usage.

Overall, the analysis confirms that VOCAL2VISUAL is a reliable and effective assistive communication tool, with clear opportunities for further improvement in animation speed and gloss model accuracy.

4. CONCLUSION

The study successfully demonstrates that converting speech, text, and audio into animated sign language significantly enhances accessibility for individuals with hearing impairments. By integrating Automatic Speech Recognition, Natural Language Processing, and 3D animation techniques, the system delivers an efficient, real-time translation pipeline. Experimental evaluation shows strong model accuracy, smooth animation rendering, and high user satisfaction, indicating that the approach is both practical and scalable. The findings highlight the potential of technology driven solutions in bridging communication gaps and supporting inclusive digital environments. Future enhancements such as expanding the vocabulary, improving facial expressions, and enabling multilingual support can further strengthen the system's effectiveness and real-world applicability.

REFRENCES

[1] K. Jain and S. Kumar, "A real-time Indian Sign Language recognition system using deep learning," IEEE Access, vol. 8, pp. 123677–123685, 2020.

[2] R. Rastogi, P. Singh, and A. Sharma, "Automatic speech recognition for low-resource languages: A comparative study," International Journal of Speech Technology, vol. 24, pp. 551–563, 2021.

[3] M. S. Hossain, "Natural Language Processing in assistive communication technologies," Journal of Intelligent Systems, vol. 30, no. 2, pp. 249–260, 2021.

[4] S. Dandapat and A. Basu, "Text-to-sign language translation using rule-based and machine learning techniques," IEEE Transactions on Human-Machine Systems, vol. 51, no. 3, pp. 215–226, 2021.

[5] N. R. Pal et al., "3D avatar-based sign language animation for communication support," Multimedia Tools and Applications, vol. 80, pp. 23149–23170, 2021.

[6] P. K. Sahoo and S. Mohanty, "Deep learning-based gesture recognition for sign languages," Neural Computing and Applications, vol. 33, pp. 10837–10852, 2021.

[7] T. H. Falk and J. Chan, "Multimodal speech processing for hearing-impaired users," IEEE Signal Processing Magazine, vol. 36, no. 1, pp. 102–111, 2019.

[8] World Health Organization, "Deafness and hearing loss," World Health Organization. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss. Accessed: 2023.

[9] A. Gupta and M. Srivastava, "A hybrid NLP approach for semantic interpretation of spoken language," Procedia Computer Science, vol. 192, pp. 3137–3146, 2021.

[10] M. Kelly, R. Liddell, and D. Leeson, Sign Language Interpreting and Translation Studies. Cambridge, U.K.: Cambridge University Press, 2020.

[11] A. Johari, R. Sharma, A. Meena, and V. Tiwari, "Advancements in Pre-Trained Language Models and Their Impact on Various NLP Tasks," International Journal of Engineering Trends and Applications, vol. 11, no. 3, pp. 201–209, 2024.

[12] K. Gautam, G. K. Soni, R. Ajmera, N. Hemrajani, J. Ahuja, and M. K. Jha, "Deep Reinforcement Learning for Stock Market Portfolio Optimization," in Proceedings of the 5th International Conference on Communication, Computing and Electronics Systems (ICCCES), pp. 1835–1839, 2026.

[13] M. K. Jha, G. K. Soni, G. Jain, S. Tiwari, K. Gupta, and B. Singhal, "Comparative Analysis of Classical Machine Learning Models for Twitter Sentiment Classification," in Proceedings of the 5th International Conference on Communication, Computing and Electronics Systems (ICCCES), pp. 1949–1954, 2026.

[14] A. Maheshwari, R. Ajmera, and D. K. Dharamdasani, "Unmasking Embedded Text: A Deep Dive into Scene Image Analysis," in Proceedings of the International Conference on Advances in Computation, Communication and Information Technology (ICAICCIT), pp. 1403–1408, 2023.

[15] N. Bhargava and A. Johari, "Enhancing deep learning approach for Tamil-English mixed text classification," in Proceedings of the International Conference on Applications of Machine Intelligence and Data Analytics (ICAMIDA 2022), Advances in Computer Science Research, pp. 829–837, 2023.

[16] A. Jangir, A. Agrawal, C. Sharma, G. K. Soni, R. Ajmera, and A. Johari, "Comparative Performance Analysis of Deep Learning and Traditional Algorithms for Facial Recognition and Image Classification," in Proceedings of the 4th International Conference on Automation, Computing and Renewable Systems (ICACRS), pp. 1172–1175, 2025.

[17] D. Saxena, J. Sharma, G. K. Soni, Y. Rao, S. Sharma, and S. Lavania, "Sentimental Analysis and Forecasting using Machine Learning Algorithms," in Proceedings of the 4th International Conference on Automation, Computing and Renewable Systems (ICACRS), pp. 917–921, 2025.

[18] M. Dahiya, N. Hemrajani, A. Kumar, S. Rani, and S. Rathee, Artificial Intelligence in Medicine and Healthcare. Boca Raton, FL, USA: Taylor & Francis, 2025.

[19] A. Maheshwari and R. Ajmera, "A comprehensive guide to natural language processing in Sanskrit with named entity recognition," in Proceedings of the ACM International Conference on Information Management and Machine Intelligence (ICIMI), 2023.

[20] R. Misra and N. Sharma, "Artificial Intelligence Driven Cybersecurity Techniques: Challenges and Future Directions," International Journal of Engineering Trends and Applications, vol. 13, no. 1, pp. 11–16, 2026.

[21] S. A. Saiyed, N. Sharma, H. Kaushik, P. Jain, G. K. Soni, and R. Joshi, "Transforming Portfolio Management with AI and ML: Shaping Investor Perceptions and the Future of the Indian Investment Sector," in Proceedings of the Parul University International Conference on Engineering and Technology (PiCET), pp. 1108–1114, 2025.

[22] S. Thapar, G. K. Soni, H. Kaushik, R. Singh, S. Bisht, and S. K. Bansal, "A Comparative Machine Learning Framework for Detecting Fake Accounts on Facebook," in Proceedings of the 4th International Conference on Innovative Mechanisms for Industry Applications (ICIMIA), pp. 1567–1571, 2025.

[23] N. Sharma, "A Study on Artificial Intelligence Applications in Autonomous Vehicle Systems," International Journal of Engineering Trends and Applications, vol. 13, no. 2, pp. 11–14, 2026.

[24] R. Ajmera, A. Johari, A. Goyal, A. Purohit, A. Kumar, and J. A. Ashok, "Multilingual Sentiment Analysis Based on Fine-Tuned Transformer Architectures," in Proceedings of the 5th International Conference on Communication, Computing and Electronics Systems (ICCCES), pp. 1589–1592, 2026.

[25] M. K. Jha, K. Kumar, N. Hemrajani, D. S. Rao, A. Goyal, and R. Ajmera, "AI Powered Student Performance Prediction using Explainable ML," in Proceedings of the 4th International Conference on Automation, Computing and Renewable Systems (ICACRS), pp. 1140–1144, 2025.

[26] I. Yadav, V. Shekhawat, K. Gautam, G. K. Soni, and R. Yadav, "Artificial Intelligence for Cybersecurity: Emerging Techniques, Challenges, and Future Trends," in Proceedings of the 3rd International Conference on Sustainable Computing and Data Communication Systems (ICSCDS), pp. 1176–1180, 2025.

[27] M. K. Jha, "Recent Trends and Emerging Applications of the Internet of Things: Transforming the Way We Live and Work," International Journal of Engineering Trends and Applications, vol. 12, no. 4, pp. 239–244, 2025.

[28] M. K. Jha, S. Agarwal, and V. Kabra, "Artificial Intelligence at Work: Transforming Industries and Redefining the Workforce Landscape," International Journal of Engineering Trends and Applications, vol. 12, no. 4, pp. 416–424, 2025.

[29] S. Kumari, S. Sharma, A. Johari, G. K. Soni, H. Kaushik, and S. Jain, "Twitter Sentiment Analysis using Machine Learning Techniques," in Proceedings of the 6th International Conference on Expert Clouds and Applications (ICOECA), pp. 1711–1717, 2026.

[30] Aayushi, L. Yadav, P. Paliwal, A. Johari, R. Ajmera, and G. K. Soni, "A Simulation-Based Evaluation of Machine Learning Models for Algorithmic Trading in Equity Markets," in Proc. 6th Int. Conf. Expert Clouds and Applications (ICOECA), pp. 1049–1054, 2026.

[31] R. Misra, N. Sharma, G. K. Soni, H. Arora, S. Chauhan, and A. Biswas, "A Hybrid and Ensemble Machine Learning Framework for Enhanced Brain Tumor Detection," in Proc.

Int. Conf. Electronics and Renewable Systems (ICEARS), pp. 1583–1588, 2026.

[32] H. Sharma and R. Ajmera, "Comprehensive Review and Analysis on Machine Learning Based Twitter Opinion Mining Framework," Tuijin Jishu/Journal of Propulsion Technology, vol. 44, no. 5, 2023.

[33] M. Kumar, R. Ajmera, and D. Kumar, "Statistical Analysis and Accuracy Assessment of Improved Machine Learning Based Opinion Mining Framework," Advances in Nonlinear Variational Inequalities, vol. 27, no. 1, 2024.

[34] H. Sharma and R. Ajmera, "Comprehensive Review and Analysis of Elderly Fall Detection System Using Machine Learning," Tuijin Jishu/Journal of Propulsion Technology, vol. 44, no. 5, 2023.

[35] K. Ahuja, H. Sekhawat, S. Mishra, and P. Jha, "Machine Learning in Artificial Intelligence: Towards a Common Understanding," Turkish Online Journal of Qualitative Inquiry, vol. 12, no. 8, 2021.

[36] P. Jha, D. Dembla, and W. Dubey, "Implementation of Machine Learning Classification Algorithm Based on Ensemble Learning for Detection of Vegetable Crops Disease," International Journal of Advanced Computer Science and Applications, vol. 15, no. 1, 2024.

[37] P. Jha, D. Dembla, and W. Dubey, "Implementation of Transfer Learning Based Ensemble Model Using Image Processing for Detection of Potato and Bell Pepper Leaf Diseases," International Journal of Intelligent Systems and Applications in Engineering, vol. 12, pp. 69–80, 2024.

[38] P. Jha, G. K. Soni, H. Dogra, D. Goswami, K. Choudhary, and H. Vaishnav, "Plant Disease Detection and Classification Using Convolutional Neural Network," in Proc. 4th Int. Conf. Automation, Computing and Renewable Systems (ICACRS), pp. 1442–1446, 2025.

[39] K. Paliwal, P. Jha, S. Kumari, V. Vaish, N. Vishwakarma, and A. Bansal, "Machine Learning in Electric Vehicle Consumption Modelling," in Proc. 9th Int. Conf. Intelligent Computing and Control Systems (ICICCS), pp. 727–730, 2026.

[40] P. Jha, P. Jain, A. Kumar, S. Soni, Y. Sharma, and P. Agarwal, "The Application of Markov Chains to Linguistic Predictions by Utilising Its Inherent Information Entropy," in Proc. 9th Int. Conf. Inventive Systems and Control (ICISC), pp. 1110–1114, 2025.

[41] P. Jha, T. Biswas, U. Sagar, and K. Ahuja, "Prediction with ML Paradigm in Healthcare System," in Proc. 2nd Int. Conf. Electronics and Sustainable Communication Systems (ICESC), pp. 1334–1342, 2021.

[42] P. Upadhyay, K. K. Sharma, R. Dwivedi, and P. Jha, "A Statistical Machine Learning Approach to Optimize Workload in Cloud Data Centre," in Proc. 7th Int. Conf. Computing Methodologies and Communication (ICCMC), pp. 276–280, 2023.