A Multimodal Voice and Sign Language to Intent Translation System for the Deaf and Speech-Impaired

Aryan Shinde[1], Trupti Takale[2], Sneha Sul[3], Swati Jadhav[4]
[1,2,3]*Department of Computer Science and Engineering, K. J. College of Engineering and Management Research, Pune, India*
[4]*Guide Professor, Department of Computer Science and Engineering, K. J. College of Engineering and Management Research, Pune, India*

***Abstract*—Communication remains a major barrier for individuals who are deaf, mute, or speech-impaired. Traditional assistive systems often handle either sign language or speech in isolation, struggle with noisy real-world conditions, and rarely produce structured machine-understandable output. This paper presents a Multimodal Voice and Sign Language to Intent Translation System that converts (a) static Indian Sign Language (ISL) gestures, (b) dynamic ISL word and sentence signs, and (c) spoken or dysarthric speech into a unified six-class intent taxonomy: REQUEST, QUESTION, STATEMENT, COMMAND, EMOTION, and UNKNOWN.**

The vision pipeline uses a three-tier gesture cascade — landmark geometry rules, a MobileNetV2 convolutional neural network (CNN) for static signs, and a long short-term memory (LSTM) network for dynamic signs — combined with a pose-based 1D-CNN for full ISL sentence recognition. The audio pipeline uses faster-whisper automatic speech recognition with a dysarthric text cleaner. Multimodal fusion arbitrates between vision and speech using confidence agreement with a twelve-second pending-voice window. The complete system is deployed across three frontends — Streamlit, Flask, and OpenCV — sharing one SQLite log. Experimental results show 81.3% validation accuracy for dynamic gestures, 47.4% top-1 (75.9% top-5) for sentence recognition, and 93.75% for the optional INCLUDE-50 Transformer, with real-time inference on commodity hardware. The proposed framework offers a practical, contactless, and inclusive communication aid for the deaf and speech-impaired community.

***Index Terms*—Convolutional Neural Network, Dysarthric Speech, Indian Sign Language, Intent Classification, LSTM, MediaPipe, Multimodal Fusion, Pose Estimation.**

I. INTRODUCTION

Communication is a fundamental human right, yet over 70 million people worldwide are estimated to be deaf or speech-impaired, and a significant subset additionally experience dysarthria — a motor-speech disorder that reduces the intelligibility of spoken language. In multilingual and developing regions such as India, the lack of widely deployed Indian Sign Language (ISL) recognition systems and the scarcity of dysarthric-robust speech recognition further widens the accessibility gap.

Existing assistive systems generally fall into two narrow categories. The first focuses on isolated sign language recognition using either static image classifiers or temporal video models, but rarely produces structured downstream output that other software can consume. The second focuses on automatic speech recognition (ASR), but consumer-grade ASR systems are typically trained on fluent, neurotypical speech and degrade substantially on dysarthric input. Neither category by itself addresses the realistic case in which a user wishes to combine partial speech with partial signing in a single interaction.

This paper proposes a Multimodal Voice and Sign Language to Intent Translation System that unifies three input modalities — static signs, dynamic sign sequences, and speech or text — into a single six-class intent taxonomy: REQUEST, QUESTION, STATEMENT, COMMAND, EMOTION, and UNKNOWN. By mapping every modality to a small, shared, semantically meaningful output space, the system produces a uniform downstream signal that can

drive responses, accessibility prompts, or smart-home actions. The system is implemented as a modular Python application running on commodity hardware (CPU-only inference) and exposed through three independent frontends — a Streamlit web app, a Flask HTTP service, and a native OpenCV window — all sharing one SQLite interaction log.

The main contributions of this work are: (i) a three-tier gesture recognition cascade combining geometry rules, a MobileNetV2 static CNN, and a dynamic-gesture LSTM; (ii) a pose-only sentence recognition model that achieves 47.4% top-1 and 75.9% top-5 accuracy on the 97-class ISL-CSLTR corpus despite poor hand visibility in the source dataset; (iii) a dysarthric-robust speech path built on faster-whisper plus a custom text cleaner that handles fillers, contractions, stutters, and fuzzy phoneme matching; and (iv) a twelve-second multimodal fusion window that arbitrates between vision and speech using confidence agreement.

II. PROBLEM STATEMENT

Deaf, mute, and speech-impaired users face persistent barriers when interacting with mainstream digital systems. Pure sign-language recognition systems are typically built around a single classifier and one fixed input modality, which limits their robustness in real-world conditions where lighting, background, hand pose, and signing speed vary continuously. Speech recognition systems, on the other hand, often fail on dysarthric or accented speech because they are trained predominantly on fluent neurotypical corpora.

The major challenges this work seeks to address include:

- Recognition of both static and dynamic Indian Sign Language signs under realistic webcam conditions.
- Recognition of full ISL sentences from low-resolution video where hand visibility is unreliable.
- Robust speech recognition for dysarthric and disfluent speakers.
- Unification of heterogeneous input modalities into a single structured intent output.
- Real-time performance on commodity CPU-only hardware.
- Persistent logging of every recognized interaction for audit and retraining.

Therefore, there is a need for a unified, intelligent, and contactless communication aid that combines sign language and speech recognition while exposing a stable, structured intent layer for downstream applications.

III. OBJECTIVES

The primary objectives of the proposed system are:

- To design a real-time multimodal communication assistant for deaf and speech-impaired users.
- To recognize both static and dynamic Indian Sign Language signs from a standard webcam.
- To recognize full ISL sentences from upper-body pose alone.
- To support dysarthric and disfluent speech using a robust ASR pipeline.
- To unify all input modalities into a six-class intent taxonomy.
- To fuse vision and speech signals using a confidence-weighted, time-bounded arbitration scheme.
- To deliver the system through multiple deployment surfaces (Streamlit, Flask, OpenCV) sharing one persistent log.

IV. LITERATURE REVIEW

Several researchers have investigated sign language recognition and assistive communication systems using a wide variety of techniques.

1) Starner and Pentland [1] proposed early hidden Markov model (HMM) based systems for real-time American Sign Language recognition.
2) LeCun et al. [2] introduced convolutional neural networks, which subsequently became the dominant architecture for image-based sign recognition.
3) Sandler et al. [3] proposed MobileNetV2, an efficient mobile-friendly CNN architecture widely adopted for transfer learning in resource-constrained sign-recognition pipelines.
4) Hochreiter and Schmidhuber [4] introduced long short-term memory (LSTM) networks, which enable temporal modelling of dynamic sign sequences.
5) Lugaresi et al. [5] released MediaPipe, providing real-time hand and pose landmark estimation that has become a de-facto standard for sign-language feature extraction.

6) Radford et al. [6] introduced Whisper, a large-scale multilingual ASR model that significantly improves robustness on noisy and accented speech; its lightweight faster-whisper port enables CPU deployment.
7) Sridhar et al. [7] released the INCLUDE dataset for Indian Sign Language, on which the AI4Bharat group later trained pose-based Transformer recognisers.

Most existing systems address only a single modality, depend on GPU inference, or do not produce structured downstream output. The proposed system addresses these limitations by combining a three-tier vision cascade, a pose-only sentence model, a dysarthric-robust speech path, and a shared intent layer, while running entirely on CPU.

V. PROPOSED SYSTEM

The proposed multimodal communication system is designed to translate sign language gestures, ISL sentences, and spoken or dysarthric speech into a unified six-class intent taxonomy. The system operates through five stages that together form a complete input-to-response pipeline.

The main stages of the system include:

1) Input acquisition (vision, audio, and text).
2) Feature extraction (hand landmarks, pose landmarks, and ASR).
3) Parallel classification (rule-based, CNN, LSTM, 1D-CNN, optional Transformer, and keyword-based intent matcher).
4) Multimodal fusion and intent classification.
5) Response generation, persistence, and user interface.

Webcam frames are processed in parallel by two MediaPipe pipelines: a hand landmarker (21 landmarks × 3 coordinates = 63 features) and a pose landmarker (33 landmarks reduced to 21 upper-body points with shoulder-centred normalisation and frame-difference velocity, yielding 126 features per frame). Microphone audio is processed by the faster-whisper small.en model for English, or the multilingual small model for Hindi, followed by a dysarthric text cleaner that performs filler removal, contraction expansion, stutter collapse, and fuzzy phoneme matching.

Each modality emits a label with a confidence score. These labels are mapped to one of six intent classes by a keyword-based intent classifier. When both vision and speech produce simultaneous predictions, a fusion module retains the more confident signal and applies a small agreement bonus when both modalities agree on the same intent.

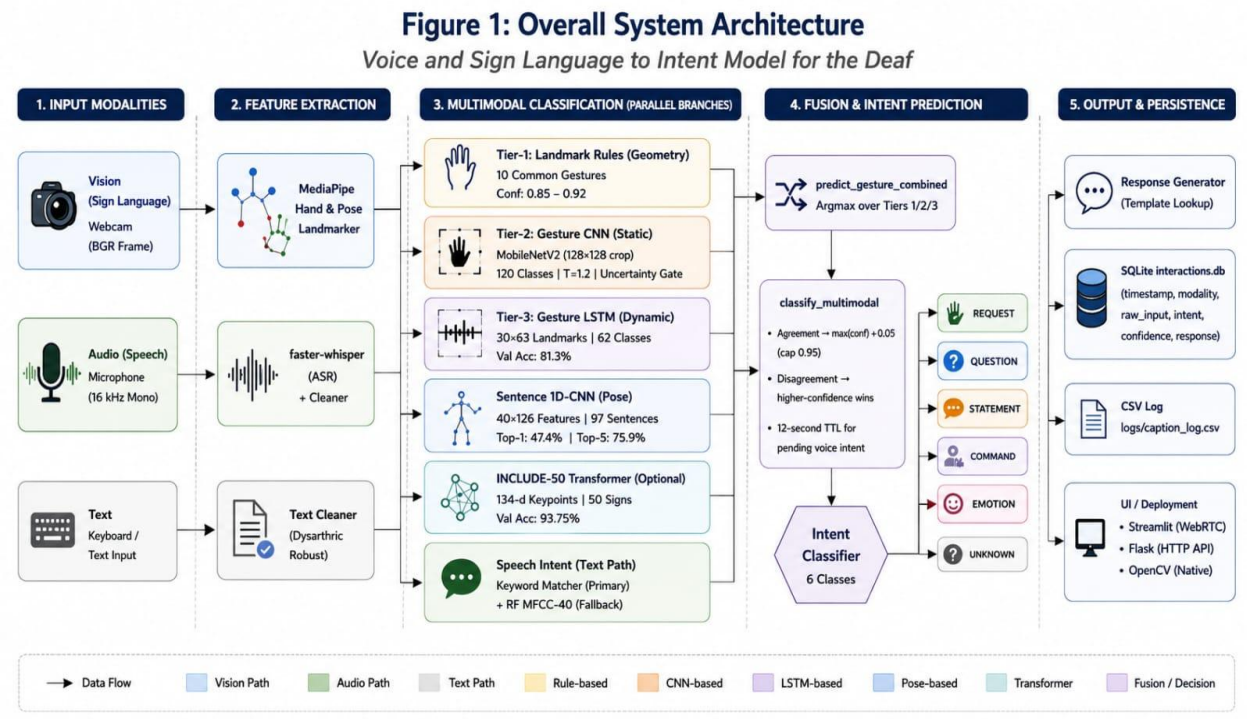

Fig. 1: Overall System Architecture — Voice and Sign Language to Intent Model for the Deaf

System workflow:
Webcam / Microphone / Text → MediaPipe + Whisper → 3-Tier Cascade / Sentence 1D-CNN / Keyword Matcher → classify_multimodal → Intent (6 classes) → Response Generator → SQLite Log + UI Caption.

VI. DATASETS

The proposed system was trained and evaluated on a combination of publicly available Indian Sign Language datasets, a custom composite dataset, and a clinical dysarthric speech corpus. The active inference vocabulary spans approximately 120 static CNN classes, 62 dynamic LSTM classes, 97 sentence-level classes, and 50 optional INCLUDE-50 classes.

- ISLRTC ISL Dataset — 36 classes (digits 0–9 and letters a–z), approximately 36,000 250×250 RGB images, used to train the static gesture CNN.
- HaGRID Sample (120k, 384p) — seven gesture classes mapped to common interaction gestures, with 2,100 cropped hand images used in the custom training merge.
- Mendeley ISL Real-life Words — 20 word-level classes such as doctor, pain, college, and work, with 17,880 images from 30 signers.
- Mendeley ISL Everyday Phrases — 44 phrase classes such as good_morning and thank_you, with 1,760 organised PNG images.
- Kaggle ISL Video Dataset — 62 dynamic sign classes including bring, call, give, learn, and want, with 3,721 H.264 MP4 clips of 60–65 samples per class, used for the temporal LSTM.
- ISL-CSLTR Corpus — 97 ISL sentences across seven signers, 18,863 frames forming 663 sequences, used to train the sentence 1D-CNN.
- Dysarthria Speech Corpus — 4,316 paired WAV and TXT files from multiple dysarthric speakers, used to train an MFCC-based RandomForest fallback.
- Custom Composite Dataset — approximately 18,600 hand-cropped 128×128 images merged from HaGRID, Mendeley, and webcam captures, used to retrain the CNN with 72 classes.
- INCLUDE-50 (AI4Bharat) — 50 ISL signs delivered as a pretrained Transformer checkpoint operating on 134-dimensional pose plus hand keypoint vectors.

VII. METHODOLOGY

The working methodology of the proposed system consists of the following stages:

1) Frame and audio acquisition.
2) Hand and pose landmark extraction using MediaPipe.
3) Three-tier gesture cascade for static and dynamic signs.
4) Pose-based sentence recognition using a 1D-CNN.
5) Speech recognition and dysarthric text cleaning.
6) Intent classification and multimodal fusion.
7) Response generation, logging, and display.

A. Frame and Audio Acquisition

The webcam continuously captures BGR frames at approximately 640×480 resolution and 15 frames per second. The microphone records 16 kHz mono PCM audio either continuously (Streamlit) or on demand via a push-to-talk button. Text input is also accepted from the on-screen keyboard for users who cannot speak.

B. Hand and Pose Landmark Extraction

Each frame is passed in parallel to the MediaPipe HandLandmarker, which returns 21 hand landmarks in three dimensions, and to the MediaPipe PoseLandmarker, which returns 33 full-body landmarks. The pose output is reduced to the 21 upper-body points relevant for signing, shoulder-centred, and augmented with frame-difference velocity features, producing a 126-dimensional vector per frame.

C. Three-Tier Gesture Cascade

The first tier applies geometry-based rules over the 63-dimensional hand landmark vector, recognising ten common gestures — thumbs up/down, fist, peace, open hand, OK sign, pointing, love you, three, and four — with confidence between 0.85 and 0.92. If no rule matches, the second tier crops the hand using a convex hull on a white background, resizes to 128×128, and classifies through a MobileNetV2-based CNN with an uncertainty gate ($p < 0.12$, top-2 gap < 0.03, or entropy > 0.88) that rejects low-quality predictions. The third tier maintains a 30-frame rolling buffer of landmarks and feeds it to an LSTM network that recognises 62 dynamic signs. The final label is the argmax confidence across the three tiers.

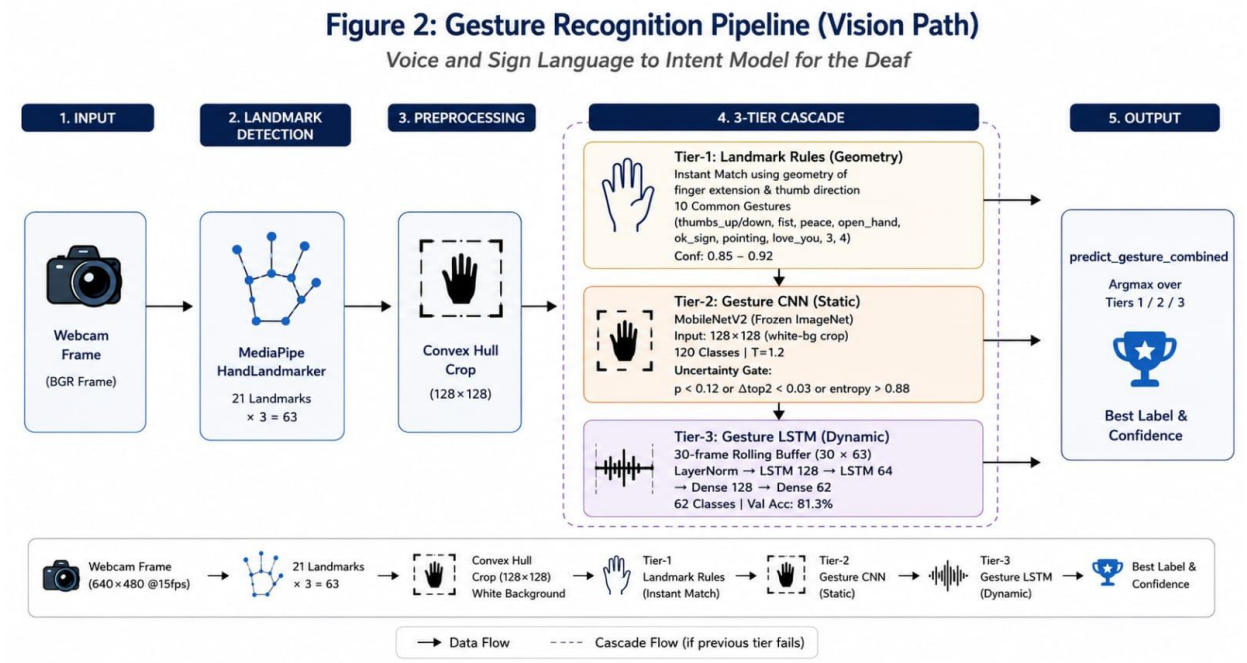

Fig. 2: Gesture Recognition Pipeline (Vision Path) showing the three-tier cascade.

D. Pose-Based Sentence Recognition

The sentence model operates on a 40-frame rolling buffer of 126-dimensional pose-plus-velocity features. The architecture is a 1D-CNN with parallel convolutional branches of kernel sizes 3, 5, and 7 (multi-scale feature extraction), followed by batch normalisation, pooling, additional convolutions, global average pooling, and a softmax layer over 97 ISL sentence classes. The pose-only design was adopted because hand detection on the source 320×240 ISL-CSLTR frames was only about 39%, whereas pose detection was approximately 100%.

E. Speech Recognition and Dysarthric Cleaning

Audio is transcribed using the faster-whisper small.en model for English (beam size 2, temperature ladder 0.0 / 0.4, no-speech threshold 0.45) or the multilingual small model for Hindi (with a Devanagari ratio gate and an average log-probability gate of -1.0). The resulting transcript is passed through a dysarthric text cleaner that strips fillers, expands contractions, collapses stutters, deduplicates words, and applies fuzzy phoneme snapping to a known keyword vocabulary.

F. Intent Classification and Multimodal Fusion

Each recognised label — whether from vision or speech — is mapped to one of six intent classes (REQUEST, QUESTION, STATEMENT, COMMAND, EMOTION, UNKNOWN) using a keyword-based scoring rule (one hit → 0.65, two hits → 0.80, three or more hits → 0.92). When both vision and speech produce intents within a twelve-second window, the fusion module retains the higher-confidence signal; if both modalities agree on the same intent, a confidence bonus of 0.05 is applied (capped at 0.95).

G. Response Generation, Logging, and Display

The final intent and recognised text are passed to a template-based response generator that selects a natural language reply. Every interaction — including timestamp, modality, raw input, intent, confidence, and response — is written to the SQLite interactions.db database and to a CSV log. The selected response is rendered in a real-time caption panel inside the active frontend (Streamlit, Flask, or OpenCV).

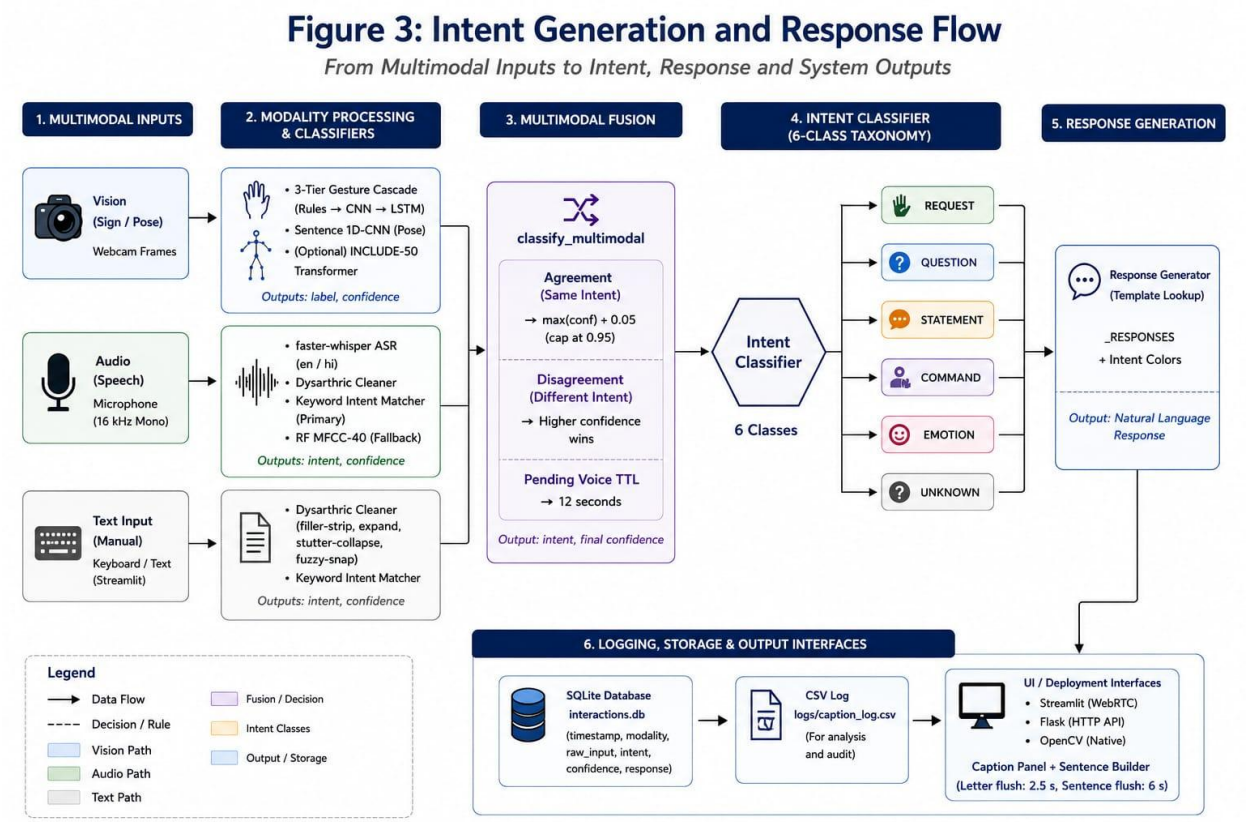

Fig. 3: Intent Generation and Response Flow — from multimodal inputs to intent, response, and system outputs.

VIII. ALGORITHM

Capture webcam frame and microphone audio
Extract hand and pose landmarks using MediaPipe
Apply Tier-1 geometry rules; if matched, emit label
Otherwise apply Tier-2 MobileNetV2 CNN on cropped hand
Update 30-frame buffer and apply Tier-3 LSTM
Update 40-frame pose buffer and apply Sentence 1D-CNN
Transcribe audio via faster-whisper and clean text
Map each modality label to an intent class
Fuse modalities using confidence arbitration
Generate response, log to SQLite, render in UI

IX. MACHINE LEARNING ALGORITHMS

A. MobileNetV2 Gesture CNN

MobileNetV2 is used as a frozen ImageNet-pretrained feature extractor, followed by global average pooling, a dense 256-unit ReLU layer with 30% dropout, and a softmax over the active vocabulary. The input is a 128×128 RGB hand crop, normalized to the [-1, 1] range. Training augmentations include random rotation, zoom, translation, brightness, and contrast. Optimization uses Adam at a learning rate of 1×10^{-3}, sparse categorical cross-entropy loss, batch size 32, and 15 epochs with early stopping.

B. Gesture LSTM

The dynamic gesture model accepts 30-frame sequences of 63-dimensional hand landmarks. The architecture is LayerNorm → LSTM(128) → LSTM(64) → Dense(128) → Softmax(62). It is trained with Adam, sparse cross-entropy, batch size 32, and up to 30 epochs with early stopping. Reported validation accuracy is 81.3%.

C. Sentence 1D-CNN

The sentence model is a multi-scale 1D-CNN over 40-frame pose-plus-velocity sequences. Three parallel Conv1D branches with kernel sizes 3, 5, and 7 are concatenated, batch-normalized, pooled, followed by deeper Conv1D layers, global average pooling, a 128-unit dense layer, and a softmax over 97 classes. Eight-fold augmentation — Gaussian noise (σ=0.03), time-

warping in [0.7, 1.3], 30% frame dropout, and scale in [0.9, 1.1] — expands the 663 native sequences to roughly 5,300 training samples. Reported accuracy is 47.4% top-1 and 75.9% top-5.

D. INCLUDE-50 Transformer (Optional)

The optional AI4Bharat INCLUDE-50 model is a BERT-style Transformer that operates on 134-dimensional pose and hand keypoint vectors. Two configurations are available: a small variant (hidden size 256, four heads, two layers) reaching 72.7% validation accuracy, and a large variant (hidden size 512, eight heads, four layers) reaching 93.75%.

E. Speech Intent Classifier

The primary speech path uses faster-whisper combined with a keyword-based intent scorer. As an offline fallback, an MFCC-40 RandomForest classifier with 150 trees and balanced class weights is trained on the dysarthric corpus, reaching 70–80% accuracy.

F. Intent Confidence Formula

When both modalities agree on the same intent within the twelve-second fusion window, the final confidence is computed as:

$C_{final} = \min(\max(C_{vision}, C_{speech}) + 0.05, 0.95)$

Where C_vision and C_speech are the per-modality confidences. In the case of disagreement, the modality with the higher individual confidence is retained without bonus.

X. IMPLEMENTATION

The proposed system was implemented in Python 3.11 using a modular architecture. The vision pipeline uses MediaPipe Tasks API for hand and pose landmarking, OpenCV for frame capture and visualisation, and TensorFlow / Keras for the CNN, LSTM, and 1D-CNN models. PyTorch is used only for the optional INCLUDE-50 Transformer. The speech pipeline uses faster-whisper (int8 quantised) for ASR and scikit-learn for the MFCC RandomForest fallback. NumPy and SciPy provide numerical support.

Persistence uses SQLite through a lightweight utility module that exposes an interactions table with timestamp, modality, raw input, intent, confidence, and response columns. Three frontends share this backend: a Streamlit application with WebRTC video and a sentence builder, a Flask HTTP API for programmatic clients, and a native OpenCV window for low-overhead deployments. All frontends write to the same database and CSV log.

Each frontend uses different stabilisation parameters. The Streamlit app polls the WebRTC transformer every 0.8 s with a caption threshold of 0.35, a cooldown of 1.2 s, and a 3-of-7 majority-vote stabiliser. The Flask app uses a per-frame HTTP threshold of 0.30 with a 3-second letter accumulation buffer. The OpenCV window runs at native 30 FPS, applying a 1.5-second hold-to-commit stabiliser at a confidence threshold of 0.40.

A. Sentence Builder

Inside the Streamlit application, single-letter signs accumulate into a fingerspelled word that is flushed after 2.5 seconds of inactivity. Recognised whole words are appended to the current sentence, which is finalised after 6 seconds with no new sign or when the user clicks the Done button. Every committed sentence is logged to interactions.db and to logs/caption_log.csv.

XI. RESULTS AND DISCUSSION

The proposed system was evaluated under typical indoor classroom and home environments using a commodity laptop webcam and microphone. Experimental analysis demonstrated reliable real-time performance and competitive recognition accuracy across all four model branches.

A. Recognition Accuracy

The static gesture CNN achieved high accuracy on common ISL letters and gestures, with the Tier-1 rule layer correctly classifying common gestures such as thumbs-up at confidence levels of approximately 92% in live webcam captures, as shown in Fig. 4. The dynamic gesture LSTM reached 81.3% validation accuracy across 62 classes. The sentence 1D-CNN achieved 47.4% top-1 and 75.9% top-5 accuracy on the 97-class ISL-CSLTR corpus, a ten-fold improvement over a vanilla LSTM baseline of 4.5% on the same data. The optional INCLUDE-50 Transformer reached 93.75% validation accuracy.

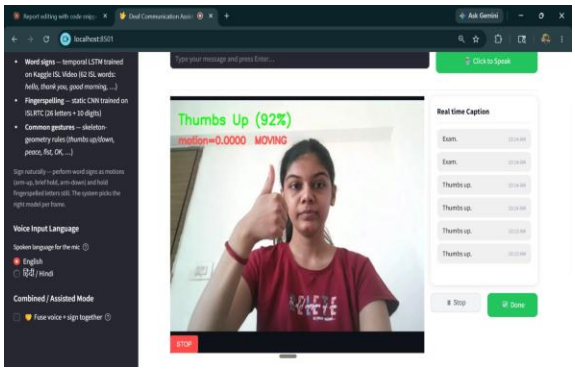

Fig. 4: Live recognition of the Thumbs Up gesture at 92% confidence in the Streamlit frontend, with real-time caption panel on the right.

B. Real-Time Performance

All inference paths run in real time on CPU. Landmark-rule recognition (Tier-1) returns within a single frame, the static CNN typically completes within 30–50 ms per inference, and the LSTM and 1D-CNN models run comfortably within the per-frame budget at 15 FPS. The faster-whisper small.en model transcribes a 5-second audio clip in approximately 3 seconds of CPU time.

C. Observations

- The three-tier cascade significantly reduces misclassification of background noise: when geometry rules match, the CNN and LSTM are bypassed, producing fast and stable output.
- Convex-hull white-background masking substantially closes the domain gap between clean training images and noisy webcam frames.
- The pose-only sentence model is the key enabler for ISL-CSLTR, where hand-detection rates are otherwise too low for usable hand-based models.
- The dysarthric text cleaner allows the system to handle disfluent and accented speech that consumer ASR alone routinely mis-transcribes.
- The twelve-second multimodal fusion window allows users to combine partial speech with partial signing in a natural way.

D. Comparison with Existing Systems

Table I: Per-Model Validation Accuracy

Model	Classes	Val Acc
Static CNN	~120	≈ 90%
Gesture LSTM	62	81.3%
Sentence 1D-CNN	97	47.4% / 75.9% (top-5)
INCLUDE-50 Transformer	50	93.75%
Speech RF (Fallback)	Intent classes	70–80%

XII. ADVANTAGES

- Unifies three input modalities — static signs, dynamic signs, and speech — into one structured intent layer.
- Cascaded gesture recognition combines the speed of geometry rules with the coverage of deep models.
- Pose-only sentence model works on low-resolution video where hand-based models fail.
- Dysarthric text cleaner improves robustness for users with motor-speech impairments.
- Runs entirely on CPU; no GPU required.
- Three deployment surfaces (Streamlit, Flask, OpenCV) share a single backend and log.
- Every interaction is persistently logged for audit, retraining, and analytics.

XIII. LIMITATIONS

- Sentence top-1 accuracy of 47.4% on ISL-CSLTR is still well below human-level performance.
- Hand detection degrades under poor lighting or extreme background clutter.
- The system currently supports only English and Hindi speech input.
- Whisper inference, while CPU-feasible, introduces noticeable latency on long clips.
- The INCLUDE-50 vocabulary is small (50 signs) relative to full conversational ISL.

XIV. FUTURE SCOPE

- Expand sentence-level vocabulary using larger ISL corpora and self-supervised pretraining.
- Add additional Indian languages to the speech path (Marathi, Tamil, Bengali, etc.).

- Integrate a mobile application using on-device TensorFlow Lite models.
- Add a cloud synchronisation layer for caregivers and teachers.
- Extend the intent taxonomy to support multi-turn dialogue and slot filling.
- Integrate a generative LLM response layer constrained by the recognised intent for richer replies.

XV. CONCLUSION

This paper presented a Multimodal Voice and Sign Language to Intent Translation System for deaf and speech-impaired users. The proposed framework unifies static gestures, dynamic sign sequences, full ISL sentences, and spoken or dysarthric speech into a six-class intent taxonomy through a three-tier vision cascade, a pose-only sentence 1D-CNN, an optional INCLUDE-50 Transformer, and a dysarthric-robust speech path.

Experimental results — 81.3% validation accuracy for dynamic gestures, 47.4% top-1 and 75.9% top-5 for 97-class ISL sentences, 93.75% for the optional Transformer, and 70–80% for the speech fallback — confirm that the system achieves competitive recognition while running entirely on CPU. The shared backend and three deployment surfaces (Streamlit, Flask, OpenCV) demonstrate that the architecture is practical, modular, and extensible.

The proposed framework provides a contactless, intelligent, and inclusive communication aid that brings together sign language recognition and dysarthric-robust speech recognition under a single, structured intent layer suitable for next-generation accessibility applications.

REFERENCES

[1] T. Starner and A. Pentland, "Real-time American Sign Language recognition from video using hidden Markov models," in Proc. International Symposium on Computer Vision (ISCV), Coral Gables, FL, 1995, pp. 265–270.

[2] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," Proceedings of the IEEE, vol. 86, no. 11, pp. 2278–2324, 1998.

[3] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4510–4520.

[4] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.

[5] C. Lugaresi et al., "MediaPipe: A framework for building perception pipelines," arXiv preprint, arXiv:1906.08172, 2019.

[6] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in Proc. International Conference on Machine Learning (ICML), 2023, pp. 28492–28518.

[7] A. Sridhar, R. G. Ganesan, P. Kumar, and M. Khapra, "INCLUDE: A large scale dataset for Indian Sign Language recognition," in Proc. ACM International Conference on Multimedia (ACM MM), 2020, pp. 1366–1375.

[8] F. Chollet et al., "Keras," 2015. [Online]. Available: https://keras.io

[9] G. Bradski, "The OpenCV Library," Dr. Dobb's Journal of Software Tools, 2000.

[10] S. Camgöz, O. Koller, S. Hadfield, and R. Bowden, "Sign language transformers: Joint end-to-end sign language recognition and translation," in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 10023–10033.