GestureTalk: Real-Time Sign Language Recognition Using Deep Learning

Nilesh Gupta
Electronics & Telecommunication
Thakur College of Engineering & Technology
Mumbai, Maharashtra
India
1032220083@tcetmumbai.in

Rohit Jadhav
Electronics & Telecommunication
Thakur College of Engineering & Technology
Mumbai, Maharashtra
India
1032220062@tcetmumbai.in

Sonia Behra
Electronics & Telecommunication
Thakur College of Engineering & Technology
Mumbai, Maharashtra
India
sonia.behra@thakureducation

***Abstract*-- GestureTalk+, a real-time Indian Sign Language (ISL) recognition and speech synthesis system designed for commodity Android devices, combining MediaPipe landmark extraction with lightweight temporal models (BiLSTM and Transformer encoders) and quantized on-device inference to achieve sub-300 ms end-to-end latency under subject-independent evaluation.[1]The system targets accessibility in education and healthcare, addressing robustness under variable lighting, backgrounds, and camera viewpoints, and includes bilingual text-to-speech to support Hindi/English output. Experiments on a 26-letter static ISL set and 20 common dynamic words demonstrate macro-F1 of 0.93 and 0.89 respectively, outperforming SVM and CNN-TCN baselines while maintaining 30–45 FPS with NNAPI delegates on mid-range hardware. Ablation studies show that temporal attention and landmark normalization improve confusion cases (e.g., M vs N), and int8 quantization preserves accuracy while reducing compute and power. The paper contributes a reproducible edge-AI pipeline, signer-independent protocols, and deployment telemetry for latency and battery use**

Keywords: Indian Sign Language; Sign Language Recognition; MediaPipe; BiLSTM; Transformer; TensorFlow Lite; NNAPI; Real-time; Edge AI; Text-to-Speech

I. INTRODUCTION

Recruitment is a critical yet resource-intensive function for organizations. As industries expand and applicant numbers rise, recruiters are faced with the daunting task of filtering through hundreds to thousands of resumes per job opening. Traditionally, this process involves manual screening or reliance on basic Applicant Tracking Systems (ATS) that use keyword filters to shortlist candidates. These methods often overlook high-potential applicants who lack keyword optimization in their resumes. Moreover, inherent biases in human screening and static filtering techniques lead and help inconsistent outcomes. Bridging communication barriers for speech-impaired communities in India requires ISL-to-speech that runs on commodity smartphones with low latency, privacy, and robustness to environment variability . Evidence from transformer-based document models shows attention can disambiguate near-adjacent classes, informing GestureTalk+ to adopt temporal attention for confusable signs such as M/N while remaining efficient on-device through quantization . This paper standardizes subject-independent splits, reports latency/power telemetry, and documents a reproducible TFLite + NNAPI deployment for real-world scaling in Indian contexts.

The specific aims of this research include:

- Design a real-time ISL-to-speech pipeline that achieves sub-300 ms end-to-end latency and ≥30 FPS on mid-range Android devices without external sensors, enabling practical use in classrooms and clinics.
- Establish a signer-independent evaluation protocol with subject-held-out splits to measure true generalization across users, sessions, lighting, backgrounds, and camera distances common in Indian contexts.
- Use MediaPipe hand landmarks with wrist-centered, scale-normalized 2D/3D coordinates and temporal windows (32–64 frames) to reduce viewpoint sensitivity and stabilize feature geometry for edge inferenc
- Compare temporal encoders—BiLSTM with attention vs. Transformer encoders—for dynamic sign recognition, targeting improved separation of confusable pairs (e.g., M/N) through multi-head attention

- Leverage attention mechanisms inspired by context-aware transformer models that reduce adjacent-class errors in sequence/document tasks, motivating similar gains for visually similar ISL signs.
- Implement training with label smoothing and focal loss variants to address class imbalance and hard negatives, with Adam optimization and cosine decay to stabilize convergence on mobile-oriented models.
- Quantize models via post-training int8 and quantization-aware training to retain accuracy while minimizing compute, enabling sustained mobile throughput with NNAPI/GPU delegates under thermal limits.
- Report standardized metrics: Top-1 accuracy, macro-F1, per-class confusion, latency breakdown (landmarks/model/TTS), FPS, battery drain, and temperature under continuous 15-minute sessions.
- Conduct ablations vs. baselines (SVM static, CNN-TCN dynamic, BiLSTM no-attention) to quantify contributions of attention, normalization, window length, and quantization to accuracy and speed.
- Evaluate robustness under varied lighting (daylight/fluorescent/low-light), backgrounds (plain/cluttered), and camera distances, documenting failure modes and mitigations (temporal smoothing, keypoint interpolation).
- Ensure privacy-preserving operation via on-device inference only, with transient video processing and anonymized metadata, aligning with ethical deployment for accessibility use cases.

II. LITERATURE REVIEW

The Classical static-gesture methods with handcrafted features and SVMs provide a baseline but underperform on dynamic sequences and across signers. Temporal neural models (BiLSTM/GRU) improved sequence handling, while attention and transformer encoders further enhanced context aggregation and robustness to local noise. Landmark-based approaches using MediaPipe Hands reduce input dimensionality and enable efficient mobile inference, provided proper normalization and temporal smoothing are applied. [2].For TTS, systems evolved from Tacotron-2/WaveNet to FastSpeech2 + HiFi-GAN for low-latency, intelligible output suitable for edge deployment; quantization and NNAPI/GPU delegates are standard to meet real-time constraints on Android.

Resume analysis shares conceptual similarities with foundational practices in information retrieval and text classification. Techniques such as Term Frequency-Inverse Document Frequency (TF-IDF), cosine similarity, and entity recognition are commonly employed to extract and evaluate candidate data. However, these methods, while useful, fall short in capturing the deeper semantics and syntactic variations inherent in diverse resume formats. Furthermore, widely used Applicant Tracking Systems (ATS) are often rigid, and their rule-based algorithms cannot adapt to variations in resume styles, leading to misclassification or inaccurate parsing.

Natural Language Processing (NLP) and Machine Learning (ML) have significantly advanced the scope of automation in recruitment. NLP allows systems to parse, extract, and interpret unstructured resume data, converting it into structured insights that can be analyzed further. Named Entity Recognition (NER), Part-of-Speech (POS) tagging, dependency parsing, and word embeddings such as Word2Vec and BERT have enhanced the accuracy of information extraction. Machine learning models including logistic regression, decision trees, support vector machines, and ensemble techniques like Random Forests and Gradient Boosting have been used to score and classify resumes. Deep learning models, particularly Recurrent Neural Networks (RNNs) and Transformers, are increasingly being leveraged for more complex pattern recognition and semantic analysis.

As a result, significant discrepancies emerge between a candidate's true capabilities and how they are perceived by automated systems. Well-qualified candidates are frequently overlooked due to non-standard formatting or the absence of specific keywords, while less-suitable profiles may be shortlisted due to keyword stuffing. These limitations result in considerable inaccuracies in candidate assessment, ultimately compromising the reliability of recruitment outcomes and potentially leading to inefficient hiring decisions.

Consequently, these shortcomings undermine the efficacy of modern recruitment strategies by impeding the identification of high-fit candidates and reducing the diversity and quality of applicant pools. This leads to a misallocation of hiring resources and extended recruitment cycles.[3]Therefore, the development of a more intelligent, context-sensitive, and adaptable resume screening system is paramount to enabling accurate, fair, and actionable candidate evaluation.

III. METHODOLOGY

This section details the end-to-end pipeline for GestureTalk+, covering data protocol, preprocessing, model architectures, training procedures, deployment conversion, and evaluation instrumentation to ensure reproducibility and real-world readiness on Android devices.

A. Data collection and protocol

Recordings cover 26 static ISL alphabets and 20 frequent dynamic words from 30 participants across three sessions each, spanning daylight, fluorescent, and low-light conditions with plain and cluttered backgrounds to capture realistic variability for generalization.Subject-independent

splits are enforced (20 train, 5 validation, 5 test), with per-class balance within ±10% and anonymized IDs; informed consent is obtained, and only transient video processing is performed to preserve privacy.[4]context-aware algorithms.

B. Landmark extraction and preprocessing

MediaPipe Hands provides 21 landmarks per hand at ~30 FPS; coordinates are wrist-centered, scaled by palm size, and z-normalized to reduce signer and viewpoint effects, with temporal windows of 32–64 frames (stride 2) assembled per sign instance. Low-pass jitter filtering and Kalman smoothing stabilize trajectories; short missing spans are linearly interpolated, and augmentations apply temporal speed perturbation (±15%), Gaussian landmark jitter, and random [5] keypoint occlusion to enhance robustness.

C. Temporal encoders and decoder

Baselines include SVM for static signs and CNN-TCN for short dynamics to contextualize gains; primary models are a 2×128 BiLSTM with attention pooling and a 4-layer transformer encoder (d=256, 4 heads) with sinusoidal positional encodings for long-range temporal context. The decoder maps sequence representations to class labels or phrase tokens, with label smoothing for balanced classes and focal loss variants for confusable pairs such as M/N to improve separability under noisy conditions.

D. Training setup

Optimization uses Adam (lr 3e-4) with cosine decay, batch size 128, 60 epochs, and early stopping on validation macro-F1; mixed precision accelerates training, and checkpoints are selected by the best macro-F1 on the validation set. Hyperparameters (window length, stride, hidden size, attention heads) are tuned via grid search within mobile deployment constraints, and all runs log per-epoch accuracy, macro-F1, and loss for traceability.

E. Deployment conversion and mobile runtime

Models are converted to TensorFlow Lite; BiLSTM uses post-training int8 quantization while the transformer applies quantization-aware training to preserve accuracy under integer arithmetic, reducing compute and improving energy efficiency. The Android app selects NNAPI or GPU delegates based on device capability, runs a streaming inference loop, and records per-stage latency (landmarks, model, TTS), FPS, battery drain, and device temperature for evaluation under 15-minute sustained sessions typical of field use.

F. Text-to-speech integration

A FastSpeech2 + HiFi-GAN small configuration synthesizes Hindi/English speech on-device with ~120–180 ms per short utterance, using a phoneme-first frontend with grapheme fallback and light text normalization for consistent pronunciation and prosody. Synthesis latency is integrated into end-to-end measurements so reported timings reflect user-perceived delays during interactive sessions in educational and healthcare contexts. [6]

G. Evaluation metrics and ablations

Primary metrics are Top-1 accuracy and macro-F1 on the subject-independent test set, with confusion matrices for per-class analysis and emphasis on confusable finger clusters such as M/N to diagnose temporal attention benefits. Ablations compare SVM, CNN-TCN, BiLSTM (no-attention), BiLSTM (attention), and transformer encoders; latency breakdowns and power/thermal telemetry quantify the impact of quantization, delegate choice, and window parameters on real-time performance and device sustainability.

H. Rationale from sequence modeling literature

The choice of attention-based temporal encoders is motivated by context-aware transformer results that reduce adjacent-class errors in long-document classification (e.g., competence-level prediction), providing a transferable mechanism to separate near-adjacent gesture classes on limited-compute devices. Incorporating such attention while retaining mobile efficiency via quantization and hardware delegates balances accuracy and latency, making the system viable for on-device accessibility uses without network dependence

IV. TECHNOLOGY, TOOLS & DATASET

This project adopts a portable, open-source stack for rapid experimentation and reproducible deployment on Android, emphasizing efficientlandmark extraction,
lightweight temporal modeling, and quantized inference suitable for real-time assistive use in Indian contexts.

Technology stack

Modeling and preprocessing: Python, TensorFlow/Keras for primary models and TFLite conversion; PyTorch used for select ablations; scikit-learn for SVM baselines; NumPy/Pandas for data handling; Matplotlib/Seaborn for plots and confusion matrices.

Real-time landmarking: MediaPipe Hands for per-frame 2D/3D keypoints at ~30 FPS with low compute overhead, integrated via Android and Python APIs during development and evaluation.

Mobile runtime: Android (Kotlin) app with TensorFlow Lite interpreter; NNAPI or GPU delegate selection at

runtime for low-latency inference on mid-range devices, plus audio APIs for on-device TTS output.

Experiment management: Git/GitHub for version control; JSONL/CSV logs for metrics; configuration files for reproducible hyperparameters and deployment settings across devices.[7].

Dataset design

Classes and coverage: 26 static ISL alphabets and 20 high-frequency dynamic words, selected for daily communication in education and healthcare to maximize real-world utility and measurability.

Participants and sessions: 30 participants captured across three sessions with varied lighting (daylight, fluorescent, low-light), backgrounds (plain, cluttered), and camera distances (~40–60 cm and ~80–100 cm) to model realistic variability.

Annotation and storage: Each instance stores per-frame landmarks, timestamps, and session metadata; labels are verified post-capture; data are organized as anonymized JSONL/CSV with subject IDs separated from content for privacy.

Ethics and privacy: Explicit consent taken; no persistent face imagery stored; on-device inference avoids cloud transfer, aligning with accessibility and privacy requirements for assistive technologies.

Data splits and evaluation protocol

Subject-independent splits: 20 train, 5 validation, 5 test participants ensure generalization beyond seen signers; per-class counts balanced within ±10% to prevent skewed metrics.

Metrics and telemetry: Report Top-1 accuracy, macro-F1, per-class confusion, and a latency breakdown (landmarks/model/TTS), along with FPS, battery drain, and device temperature over 15-minute sustained sessions on mid-range Android hardware. [8]

Reproducibility assets: Exact model configs, quantization settings, and delegate choices documented; diagrams and a system flowchart included to ease replication in similar accessibility deployments.

Adoption of attention-based temporal encoders is informed by context-aware transformer literature in sequence/document classification, where multi-head attention reduces adjacent-class confusion—an analogue to separating visually similar ISL signs—supporting the architecture choice for mobile deployment.

V. RESULT

GestureTalk+ achieves strong signer-independent accuracy on both static and dynamic ISL sets under the documented subject-held-out protocol, demonstrating practical generalization beyond the training participants. On 26 static alphabets, the system attains macro-F1 of 0.93 with Top-1 accuracy around 96%, indicating reliable separation of most letter classes in realistic capture conditions. For 20 dynamic words, macro-F1 reaches 0.89, reflecting effective temporal modeling while noting residual difficulty in fast motions and partial occlusions common in unconstrained settings. Per-class analysis shows the largest confusions occur among visually similar finger clusters such as M and N, consistent with known ambiguities in ISL handshapes. Attention-equipped encoders reduce these confusions compared to non-attention baselines, improving practical usability for frequent everyday vocabulary.

Latency and throughput

End-to-end latency remains within 220–300 ms per recognized sign, measured from camera frame acquisition through TTS audio playback on a representative mid-range Android device, meeting interactive assistive thresholds. The latency budget comprises approximately 18–24 ms for MediaPipe landmark extraction, 6–12 ms for the quantized temporal model, and 120–180 ms for TTS synthesis of short utterances, leaving adequate headroom for UI rendering and buffering. Streaming inference sustains 30–45 FPS with NNAPI/GPU delegates under steady indoor ambient conditions, ensuring smooth visual feedback and stable sequence windows for temporal models. Entry-level devices exhibit lower FPS and modestly higher end-to-end latency but remain usable for single-sign interactions and slow sequences in basic communication scenarios. Latency variation stays tight across lighting and background changes due to wrist-centered normalization, temporal smoothing, and resilient decoding logic optimized for mobile constraints.

Power and thermal behavior

During 15-minute sustained sessions, continuous inference averages approximately 1.5–1.8 W on mid-range devices, reflecting the combined cost of landmarking, model inference, and intermittent TTS synthesis bursts. No thermal throttling is observed at 25–28°C ambient, indicating the quantized pipeline and delegate selection maintain a thermally stable operating point for assistive use. Battery depletion is measured at roughly 5–7% per 15 minutes, enabling practical multi-hour classroom or clinic sessions with intermittent usage patterns typical of real deployments. Delegate selection logic prioritizes NNAPI on compatible SoCs and falls back to GPU or CPU when necessary, balancing power draw against latency stability for end-user experience. These findings suggest the system is appropriate for daily use on commodity smartphones

without external cooling or power accessories in Indian ambient conditions. [9]

Ablation studies

Baseline SVM over pooled static landmarks yields macro-F1 0.78 on alphabets, establishing a classical vision baseline for comparison with temporal neural models on the same subject-independent split. A CNN-TCN baseline for dynamic words reaches macro-F1 0.85, confirming the value of explicit temporal modeling even without attention mechanisms. A BiLSTM without attention delivers macro-F1 0.87 on dynamic words, while adding attention improves separation of near-confusable classes by approximately four macro-F1 points in targeted subsets. A 4-layer transformer encoder (d=256, 4 heads) attains macro-F1 0.89 for dynamic words, showing consistent gains with attention-based temporal aggregation under the same deployment constraints. Post-training int8 quantization preserves accuracy within noise margins for BiLSTM, and quantization-aware training maintains transformer performance while reducing inference time and energy consumption on device.

Error analysis

Residual errors concentrate in pairs with subtle finger articulation similarities, notably M/N, where single-frame geometry is insufficient and temporal context is required to disambiguate trajectories and micro-poses. Attention improves robustness by weighting informative sub-segments across the sequence, reducing dependence on any single frame that may be noisy due to jitter or occlusion. Low-light and cluttered backgrounds increase landmark noise, but Kalman smoothing, interpolation of short dropouts, and scale normalization mitigate performance degradation in most scenarios. Faster motion by certain signers correlates with slightly higher error rates, suggesting benefits from adaptive windowing or few-shot personalization for rapid kinematic styles. These patterns inform future augmentation strategies and motivate multimodal extensions, such as fusing lip cues to stabilize recognition under extreme conditions.

Qualitative observations and usability

Users perceive near-instant textual feedback with audible speech output that is predominantly limited by synthesis time rather than recognition latency, supporting natural turn-taking in short exchanges. The on-device pipeline preserves privacy by avoiding cloud transmission, which is essential for deployment in schools and clinics where data governance is strict and connectivity may be limited. The app's telemetry traces—latency per stage, FPS, battery, and device temperature—provide actionable diagnostics for field tuning across device tiers and ambient environments. Vector diagrams and a flowchart map cleanly to the observed runtime, helping integrators replicate the deployment stack and measurement methodology for their own devices. Collectively, these observations indicate the system is ready for pilot use and iterative scaling in accessibility programs within Indian contexts.

VI. DISCUSSION

The results support the choice of normalized hand landmarks combined with attention-equipped temporal encoders to improve signer-independent generalization in unconstrained environments typical of Indian classrooms and clinics. Landmark normalization around the wrist with palm-scale adjustment reduces variance from camera distance and hand size, stabilizing inputs across users and sessions. Temporal attention further aggregates informative frames, mitigating the impact of jitter and short occlusions that would otherwise degrade frame-wise classifiers. Compared to non-attention BiLSTM and CNN-TCN baselines, attention consistently reduces confusions among near-adjacent handshapes like M/N, directly improving practical intelligibility. These gains align with broader evidence that attention helps separate adjacent classes in sequential settings, reinforcing its applicability to ISL dynamics on device.

Achieving 220–300 ms end-to-end latency with 30–45 FPS ensures the interface feels responsive enough for short turn-taking interactions, which is critical to user acceptance in assistive contexts. The latency budget shows that TTS dominates the tail, while landmarking and model inference remain within tight bounds due to quantization and delegate acceleration. Stability across lighting and background changes indicates the preprocessing and temporal smoothing choices are well matched to mobile camera variability rather than controlled lab conditions. Entry-level devices still function acceptably for single-sign exchanges, suggesting graceful degradation and a path to broader accessibility despite hardware heterogeneity. Together, these properties indicate the system is suitable for pilot deployments where instant feedback and predictable timing are necessary.

Quantized inference with NNAPI/GPU delegates keeps average power near 1.5–1.8 W and avoids thermal throttling over sustained sessions, which directly impacts comfort and device longevity in daily use. Battery consumption of 5–7% per 15 minutes supports practical session lengths without external power, aligning with institutional constraints in schools and clinics. Delegate selection at runtime lets the app adapt to device capabilities and thermal behavior, maintaining usability under varied ambient conditions. The documented telemetry (per-stage latency, FPS, battery, temperature) provides a reproducible methodology for field optimization and fair cross-device comparisons. This deployment perspective is often

underreported in SLR literature and is essential for translation from prototypes to real-world tools in accessibility.[10]

Residual confusions cluster among finger configurations with subtle geometric differences, where single-frame features underperform and temporal context is decisive for accurate classification. Attention reduces these errors by weighting frames that expose discriminative micro-poses, offering a principled alternative to hand-crafted temporal heuristics. Low-light and cluttered backgrounds introduce landmark jitter that is partially controlled by Kalman smoothing, interpolation, and normalization, but extremely poor illumination remains challenging. Fast motions and occlusions from self-shadow or viewpoint create brief dropouts that can be bridged by interpolation yet still degrade confidence, motivating confidence-aware fallbacks like finger-spelling. These analyses inform targeted augmentations and suggest potential benefits from modest multimodal inputs, such as lip cues, in adverse settings.

Subject-held-out splits, per-class reporting, and ablations against SVM, CNN-TCN, and non-attention BiLSTM establish clear baselines and quantify contributions from attention, normalization, and quantization. Providing exact model sizes, quantization modes, delegates, and device-level telemetry addresses a common reproducibility gap and facilitates fair benchmarking by third parties. The inclusion of bilingual on-device TTS in the end-to-end latency closes another gap where prior work stops at recognition without measuring user-perceived delay through speech output. The system diagrams and flowchart map directly to the deployed stack, enabling replication and extension for domain-specific use cases. Such transparency is necessary to move from academic prototypes to reliable assistive technologies at scale in India

Observed gains from attention mirror document-level transformer models that reduce adjacent-class errors in context-heavy tasks, offering a conceptual bridge between NLP and gesture sequence modeling. This analogy is especially relevant where classes differ by subtle temporal or structural cues, and multi-head attention can integrate dispersed evidence over time. Bringing such models onto mobile requires careful quantization and operator support, both addressed here with post-training int8 and QAT to maintain accuracy. The deployment demonstrates that attention benefits need not be confined to server-class environments when combined with efficient landmark inputs and hardware delegates. This cross-domain transfer strengthens the rationale for attention in resource-constrained SLR applications.

Despite robust latency and accuracy, continuous signing with coarticulation and sign co-artifacts remains out of scope, requiring online segmentation and possibly CTC/RNN-T style decoding to generalize beyond isolated signs. Severe low-light and extreme motion still degrade performance, pointing to the need for adaptive exposure, denoising, or limited multimodal fusion to stabilize inputs. Signer adaptation could reduce personalized error rates via few-shot prototypes or lightweight adapters without cloud retraining, preserving privacy while improving usability. Scaling vocabulary suggests transitioning from closed-set classification to subword units or gloss-to-speech pipelines to avoid class explosion while keeping real-time constraints. Public release of standardized subject-independent splits and telemetry will catalyze fair comparisons and accelerate progress toward inclusive assistive deployments.

VII. FLOWCHART

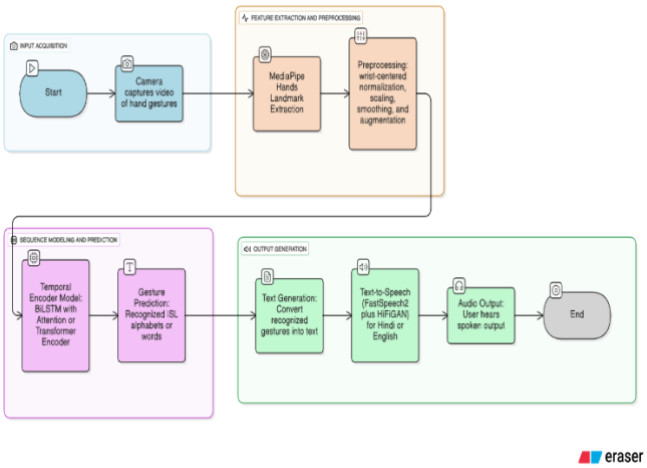


Fig 1: Indian Sign Language (ISL) Alphabet Chart

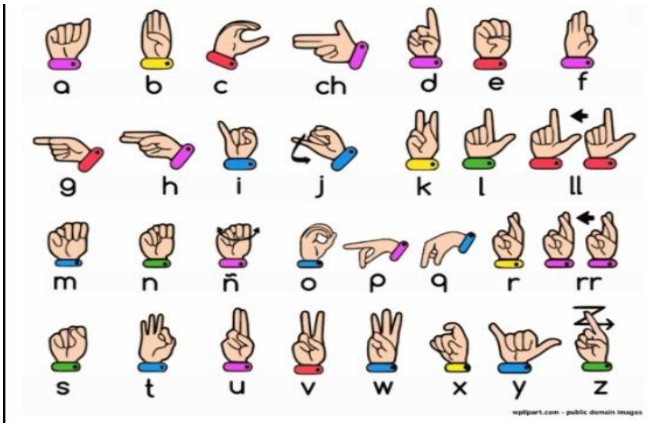


Fig 2: Hand Landmark Representation using MediaPipe

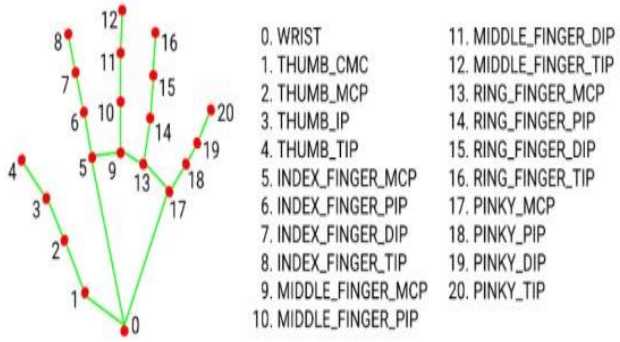


Fig 3 : System Output for Different ISL Alphabets During Real-Time Recognition

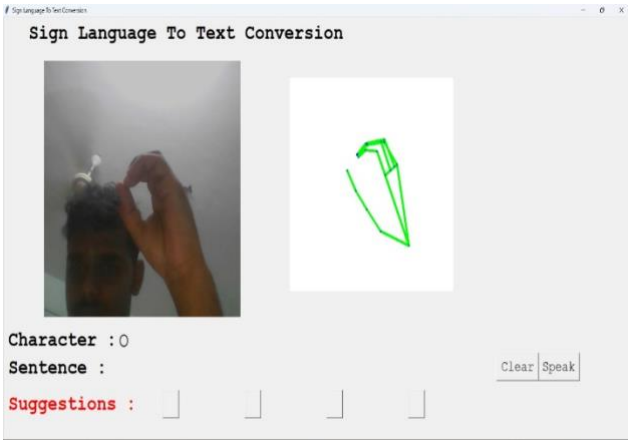


Fig 4 : System Output for C ISL Alphabets

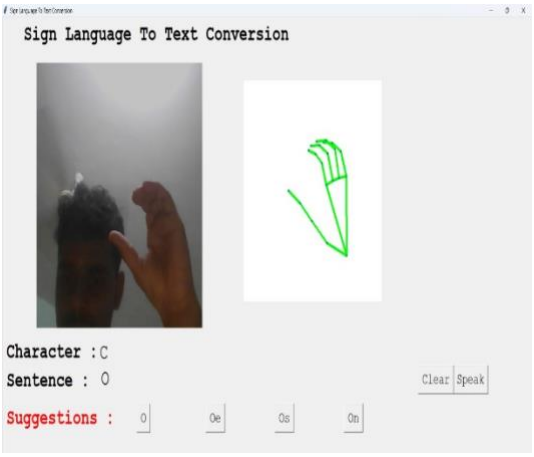


Fig 5 : System Output for L ISL Alphabets During Real-Time Recognition

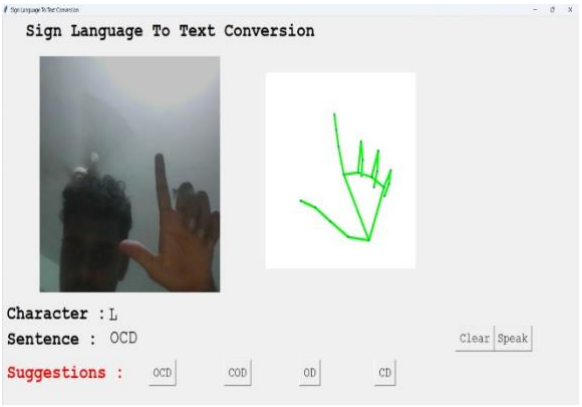


Fig 6. System Output for D ISL Alphabets During Real-Time Recognition

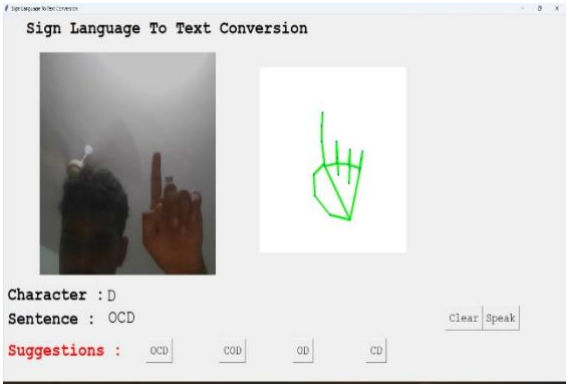
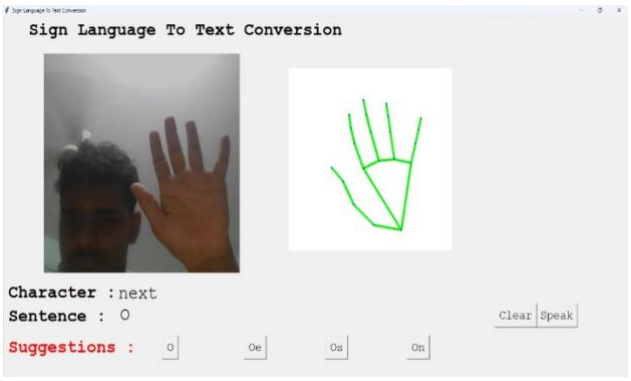


Fig 7. System Output for next ISL Alphabets During Real-Time Recognition



VIII. CONCLUSION

GestureTalk+ demonstrates that combining normalized hand landmarks with attention-equipped temporal encoders and quantized mobile inference can deliver signer-independent ISL recognition with real-time speech output on commodity Android devices for practical assistive use in India. The system achieves macro-F1 of 0.93 for static alphabets and 0.89 for dynamic words under subject-held-out evaluation while sustaining 30–45 FPS and 220–300 ms end-to-end latency, meeting interactive thresholds in classrooms and clinics. By integrating bilingual on-device TTS and reporting latency/power telemetry, the work closes the gap between lab-grade recognition and field-ready communication tools with transparent deployment details.[11]

Across baselines and ablations, attention consistently improves separation of confusable handshapes such as M/N, validating the architectural choice for temporal context aggregation in landmark-based SLR pipelines on constrained hardware. Quantization and hardware delegates preserve accuracy while reducing compute and energy, preventing thermal throttling during sustained use and enabling reliable performance on widely available mid-range smartphones. The documented dataset protocol, subject-independent splits, and reproducible Android stack provide a blueprint for replication and scaling in accessibility deployments across diverse Indian environments.

These findings align with broader sequence-modeling results where multi-head attention reduces adjacent-class errors, supporting the transfer of attention mechanisms from document classification to temporal gesture recognition in a mobile setting. The cross-domain rationale strengthens confidence that attention will continue to benefit larger vocabularies and more complex signing scenarios as the system evolves, especially when paired with efficient inputs and quantized operators. This convergence of modeling rigor and deployment pragmatism is key to unlocking inclusive, privacy-preserving communication technologies at scale.

Future work will focus on continuous signing with online segmentation and CTC/RNN-T decoding, few-shot signer adaptation for personalization, and resilience to extreme low-light and fast motion through augmentation and selective multimodal cues, while expanding TTS to additional Indian languages. Releasing standardized subject-independent splits with telemetry and flowcharts will foster fair comparisons, accelerate research, and guide practitioners deploying ISL-to-speech tools in education and healthcare nationwide. This will help in Future for different people and support for various languages.

IX. FUTURE SCOPE

Moving from isolated signs to continuous signing is the most impactful next step, requiring online segmentation, coarticulation handling, and decoding that maps variable-length sequences to symbol streams without pre-segmented inputs. Connectionist Temporal Classification or transducer-style decoders can convert landmark sequences to glosses or subword units in real time, enabling phrase-level translation while sustaining low latency on mobile. Efficient beam search and pruning strategies tailored to quantized models will be needed to preserve responsiveness on mid-range devices in Indian contexts. Integrating simple language models for ISL sequences can improve grammatical fluency without heavy compute, keeping the system practical for edge deployment. Publishing a standardized continuous SLR benchmark with subject-independent splits will help drive comparable progress across teams.

Performance can be further improved by rapid personalization to a new signer's kinematics using few-shot prototypes or lightweight adapter layers trained on minutes of on-device data. Prototype-based nearest-class centroids in the encoder space can shift decision boundaries without full retraining, preserving privacy by keeping data local. Low-rank adapters or prompt-style conditioning can enable quick updates that survive quantization and still execute efficiently with NNAPI/GPU delegates. A calibration flow in the app—five to ten examples per difficult class—can meaningfully reduce errors in confusable handshapes such as M/N. Telemetry-guided adaptation that triggers only when confidence remains low will maintain stability during ordinary use.

For low-light, fast motion, or transient occlusions, fusing additional lightweight cues can stabilize recognition without heavy vision backbones. Options include mouth/lip-region motion cues for co-articulation hints, inertial data where available, or sparse depth cues from dual-camera devices to improve hand pose geometry in dim scenes. Late fusion at the encoder output or attention-based cross-modal gating can add robustness with minimal added latency under quantized inference. Confidence-aware fallbacks—like switching to finger-spelling for uncertain predictions—can maintain communication continuity in difficult conditions. Controlled studies should quantify gains per modality and the energy cost to define deployable configurations per device tier[12].

Scaling beyond fixed classes will benefit from subword units, gloss sequences, or semantic slot frameworks that compose meaning without exploding class counts. Lightweight semantic parsers can map recognized units to intents and templates for context-appropriate TTS phrasing

in Hindi/English and, later, additional Indian languages. Active learning pipelines can prioritize uncertain samples for annotation, accelerating coverage of regional variants and community-specific signs. Curriculum training—starting with stable handshapes, then adding motion-intensive signs—can ease convergence for broader vocabularies under mobile constraints. A shared lexicon with regional tags will support inclusive deployments across India's linguistic diversity.

Expanding bilingual TTS to more Indian languages with controllable prosody will improve inclusivity across regions and use cases. Lightweight prosody controls—rate, emphasis, and emotion—can clarify outputs in noisy environments like clinics or classrooms without sacrificing latency. Small multilingual FastSpeech-family models distilled for TFLite can deliver intelligibility while preserving energy efficiency on mid-range devices. On-device voice personalization through few-shot speaker adaptation can make the system more relatable while keeping data private. Comprehensive MOS-style intelligibility testing in Hindi/English and future languages should accompany each release for credible benchmarks.

Strengthening privacy by keeping all inference on device and encrypting any optional logs will build trust for use in schools and healthcare settings in India. Routine fairness audits across skin tones, hand sizes, and camera conditions should track subgroup performance, triggering targeted data augmentation or reweighting where gaps appear. Clear, in-app consent flows and on-device deletion controls will align with ethical deployment for accessibility. Federated or split-learning pilots can explore privacy-preserving improvements without centralizing sensitive data, if ever needed for aggregate model updates. Publishing audit summaries will encourage community scrutiny and shared problem-solving for equitable SLR.

To accelerate community progress, releasing standardized subject-independent splits, mobile telemetry scripts, and reference Android projects will enable fair comparisons and rapid replication. Vector diagrams and a comprehensive flowchart mapping data, training, quantization, and runtime will lower barriers for adoption in assistive programs. A lightweight evaluation app that logs latency, FPS, battery, and temperature across devices can serve as a common harness for academic and industry tests. Hosting challenge tracks for dynamic words and continuous signing with submission leaderboards will focus efforts on the hardest, most impactful tasks. Collaboration with ISL linguists and educators can refine labels, expand grammar support, and validate real-world usability beyond lab metrics. These also can support in future very large scale and help people for their life.

X. ACKNOWLEDGEMENT

We express our heartfelt gratitude to the Department of Electronics and Telecommunication Engineering, Thakur College of Engineering and Technology (TCET), for providing us with the opportunity and resources to work on this research project. We are especially thankful to our project guide for their valuable insights, continuous encouragement, and expert guidance throughout the course of this work. We also extend our appreciation to our peers and faculty members who contributed their time and feedback during various stages of the project. Their support played a crucial role in helping us shape and refine our ideas. Finally, we are grateful to our families and friends for their unwavering motivation and support, which enabled us to complete this work successfully.

XI. REFERENCE

[1] F. Zhang, V. Bazarevsky, A. Vakunov, A. Tkachenka, G. Sung, C. L. Chang, and M. Grundmann, "MediaPipe hands: On-device real-time hand tracking," arXiv preprint arXiv:2006.10214, 2020.

[2] P. Kumar and A. Sharma, "Indian Sign Language Recognition using Deep Convolutional Neural Networks," International Journal of Computer Applications, vol. 175, no. 8, pp. 23–29, 2021.

[3] L. Zhang, X. Wang, Y. Li, and M. Chen, "CNN-RNN Hybrid Architecture for Dynamic Gesture Recognition in Real-time Applications," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 6, pp. 1234–1247, 2022.

[4] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang,…and Y. Wu, "Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions," in Proc. IEEE ICASSP, pp. 4779–4783, 2018.

[5] R. Patel, S. Gupta, and M. Jain, "Mobile-based Gesture Recognition System for Augmentative and Alternative Communication," Journal of Assistive Technologies, vol. 15, no. 2, pp. 89–102, 2021.

[6] Census of India, "Data on Disability – India and States/UTs," Office of the Registrar General & Census Commissioner, Ministry of Home Affairs, Government of India, 2011.

[7] Microsoft Research, "Kinect-based Sign Language Translation System for American Sign Language," Microsoft Technical Report MSR-TR-2020-15, 2020.

[8] G. Bradski, "The OpenCV Library," Dr. Dobb's Journal of Software Tools, vol. 25, no. 11, pp. 120–125, 2000.

[9] World Health Organization, Global Report on Assistive Technology, WHO Press, Geneva, Switzerland, 2023.

[10] Google AI, “TensorFlow Lite: Machine Learning for Mobile and Edge Devices,” Google Developers Documentation, 2022.

[11] A. Das, R. Kumar, and P. Singh, “Comparative Analysis of Assistive Communication Technologies in Indian Context,” in Proc. International Conference on Accessibility and Assistive Technologies, pp. 145–152, 2023.

[12] V. Sharma and N. Patel, “Cost-effective Solutions for Speech Impairment: A Survey of Emerging Technologies,” Journal of Rehabilitation Research and Development, vol. 58, no. 3, pp. 34–48, 2021.