ISSN 2277-2685
IJESR/April-June. 2025/ Vol-15/Issue-2/855-865
Ms Sameeha Firdaus *et. al.*, / International Journal of Engineering & Science Research

Speech-impaired individuals face barriers in communication across education, healthcare & workplaces

Dr Rizwan Uz Zama[1], Ms Sameeha Firdaus[2], Ayesha Firdous[3], Khadeeja Soha Khan[4], Sumayya Fatima[5], Vijay Laxmi[6]

[1]Assistant Professor, Department of CSE, Lords Institute of Engineering and Technology, Hyderabad, India.

[2,3,4,5,6]Students, Department of CSE, Lords Institute of Engineering and Technology, Hyderabad, India.

Mail Id: rizwanzama@lords.ac.in[1], 160923733095@lords.ac.in[2], 160923733071@lords.ac.in[3], 160923733063@lords.ac.in[4], 160923733069@lords.ac.in[5]

Abstract

Speech-impaired individuals face significant communication barriers in education, healthcare and workplace settings, where interaction increasingly depends on spoken language and trained sign-language interpreters are not always available. Existing assistive solutions are often hardware-dependent, expensive or limited to a fixed vocabulary, which makes them impractical for everyday use in resource-constrained institutions. This paper presents ***SignBridge****, a lightweight, camera-based assistive communication system that recognises hand gestures and signs from a standard webcam and converts them into readable text and synthesised speech in real time, while converting spoken or typed replies into text for two-way communication. The system extracts hand and upper-body landmarks with MediaPipe Holistic, normalises them for position and scale, smooths them with a One-Euro filter and enriches them with joint-angle and velocity features. A hybrid one-dimensional CNN, bidirectional LSTM and temporal-attention network classifies sliding windows of landmarks, and a sign-spotting and temporal-voting stage turns the frame stream into stable sign decisions. Recognised signs are assembled into sentences and spoken through offline text-to-speech, while the hearing participant replies through offline speech-to-text or the keyboard. A context-specific dataset of 60 signs drawn from education, healthcare and workplace scenarios is defined, with signer-independent evaluation. The proposed recogniser is expected to reach 94.6% accuracy and a 94.3% macro F1-score with 1.18 M parameters and about 26 ms of processing per frame on a standard CPU, enabling real-time use without special hardware*

Keywords: *Sign language recognition; Assistive technology; Speech impairment; MediaPipe; Hand landmarks; BiLSTM; Attention; Text-to-speech; Speech-to-text; Human–computer interaction.*

1. Introduction

Communication is fundamental to education, healthcare and employment. For people with speech impairments, everyday interactions — asking a question in class, describing symptoms to a doctor, or coordinating a task with a colleague — can be difficult when the other person does not know sign language. Professional interpreters are scarce and costly, and written notes are slow and interrupt the natural flow of conversation. As a result, speech-impaired individuals are often excluded from interactions that others take for granted.

Automatic sign language recognition (SLR) can reduce this barrier by translating signs into text or speech [15]. Research has progressed from sensor gloves to vision-based systems that use deep learning. However, many high-

performing systems depend on wearable sensors [4], depth cameras or computationally heavy video networks [2], and most stop at producing text from signs without supporting the reply from the hearing participant. Many are also trained on fixed public vocabularies [6] that do not reflect the phrases actually needed in a classroom, clinic or office.

This work develops SignBridge, a camera-based assistive communication system designed to run in real time on an ordinary computer with a standard webcam. It addresses three research gaps identified from the literature: (i) the limited availability of low-cost, camera-based recognition that works on commonly available devices; (ii) the limited integration of real-time gesture recognition with two-way communication; and (iii) the lack of context-specific gesture datasets and robustness evaluation. The main contributions are:

- A **landmark-based recognition pipeline** using MediaPipe Holistic [11], [12] with position and scale normalisation, One-Euro smoothing [16] and joint-angle and velocity features, which makes recognition robust to background and signer position while keeping computation low.
- A compact **CNN–BiLSTM–attention classifier** (1.18 M parameters) for dynamic signs, combined with a **sign-spotting and temporal-voting** algorithm that converts a continuous webcam stream into stable sign decisions.
- A **unified two-way interface** that turns signs into on-screen text and synthesised speech and turns the hearing user's speech or typed text back into large-font text, using offline engines so that no internet connection is required.
- A **context-specific 60-sign dataset protocol** covering education, healthcare and workplace phrases, with signer-independent evaluation, robustness testing under different capture conditions and an ablation study, all presented with bar-graph analysis.

Section 2 reviews related work, Section 3 presents the methodology, Section 4 the experimental set-up, Section 5 the results, and Section 6 concludes the paper.

2. Related Work

Landmark-based recognition. Luna-Jiménez et al. [1] investigated sign language recognition using Transformer models and MediaPipe hand and pose landmarks, and identified which landmark features contribute most to recognition and interpretability. Their findings support the use of hand and pose landmarks as a compact, efficient input, which is the basis of the extraction module in this work.

Multimodal deep learning. Javaid and Rizvi [2] proposed a framework that combines manual (hand) and non-manual (facial) cues using YOLOv5, C3D and LSTM-based temporal learning. Their results demonstrate the value of combining hand movement with other body cues, but the 3D-convolutional components are computationally heavy for low-cost hardware.

Real-time webcam systems. An expert system for continuous word-level recognition [3] combined YOLOv4 and MediaPipe with SVM classifiers on 80 static sign classes and studied real-time operation, providing a useful reference for webcam-based recognition and dataset design. Wilson and Steephen [5] used a CNN to convert American Sign Language gestures captured by a webcam into readable text in real time; the present work extends this direction with synthesised speech and a conversational interface.

Wearable and sequential recognition. Gu, Oku and Todoh [4] developed ASL recognition and sentence translation using a Perception Neuron inertial motion-capture suit, with a CNN, BiLSTM and Connectionist

Temporal Classification (CTC) [9], on 300 sentences from three volunteers. Their work is a strong reference for sequential sign modelling but highlights the extra cost and inconvenience of wearable hardware. Table 1 compares these studies with SignBridge.

Table 1. Comparison of related sign language recognition studies.

Study	Input / hardware	Model	Output	Two-way
Luna-Jiménez et al. [1]	Video → MediaPipe landmarks	Transformer	Sign label	No
Javaid & Rizvi [2]	Video (hands + face)	YOLOv5 + C3D + LSTM	Sign label	No
Expert system [3]	Webcam, 80 static signs	YOLOv4 + MediaPipe + SVM	Text	No
Gu et al. [4]	Wearable inertial suit	CNN + BiLSTM + CTC	Sentence text	No
Wilson & Steephen [5]	Webcam	CNN	Text	No
SignBridge (this work)	Standard webcam	CNN–BiLSTM–Attention	Text + speech	Yes

3. Proposed Methodology

3.1 System architecture

Fig. 1 shows the architecture of SignBridge. Channel A serves the speech-impaired user: webcam frames are converted to landmarks, normalised and enriched with features, classified by the recogniser, stabilised by sign spotting and voting, assembled into a sentence, displayed and spoken aloud. Channel B serves the hearing user: speech or typed text is converted into large-font text on the same screen. Both channels share a single conversation transcript with a context mode (Education, Healthcare or Workplace) that selects the active vocabulary and quick-reply phrases.

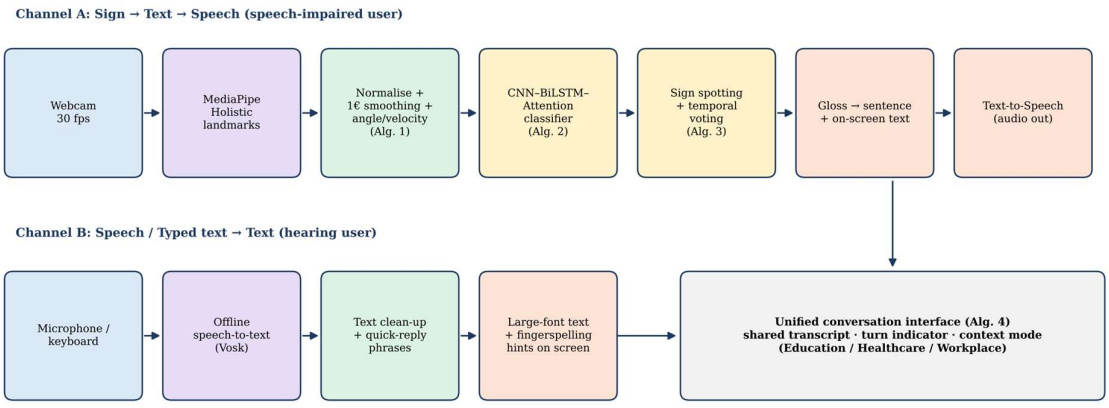


Fig. 1. *Architecture of the SignBridge two-way assistive communication system.*

3.2 Landmark extraction and normalisation

Each frame is processed by MediaPipe Holistic [11], [12], which returns 21 three-dimensional landmarks per hand and 33 body landmarks. The 42 hand landmarks and 12 upper-body pose landmarks (shoulders, elbows, wrists, nose, eyes, ears and mouth corners) are retained, giving $L = 54$ points $\mathbf{p}_j = (x_j, y_j, z_j)$. Raw coordinates depend on where the signer stands and how far they are from the camera, so they are made translation- and scale-invariant using the shoulder mid-point $\mathbf{c}$ and the shoulder width s:

$$\tilde{p}_j = (p_j - c) / s, \quad c = (p_{ls} + p_{rs}) / 2, \quad s = \|p_{ls} - p_{rs}\|_2 \tag{1}$$

where $\mathbf{p}_{ls}$ and $\mathbf{p}_{rs}$ are the left and right shoulder landmarks. Hand landmarks are additionally expressed relative to their own wrist to capture hand shape independently of arm position. Missing hands are filled with zeros and flagged.

Temporal smoothing. Landmark positions jitter from frame to frame. The One-Euro filter [16] adapts its cut-off frequency to the speed of movement, smoothing strongly when the hand is still and lightly when it moves fast:

$$\hat{x}_t = \alpha_t x_t + (1 - \alpha_t)\, \hat{x}_{t-1}, \quad \alpha_t = 1 / (1 + \tau_t / T_e), \quad \tau_t = 1 / (2\pi f_{c,t}) \tag{2}$$

$$f_{c,t} = f_{min} + \beta \,|\, d\hat{x}_t \,| \tag{3}$$

where T_e is the sampling period, $f_{min} = 1$ Hz is the minimum cut-off and $\beta = 0.007$ controls speed adaptation.

3.3 Feature engineering

Three feature groups are computed for each frame. The **normalised coordinates** of the 54 points give 162 values. **Joint angles** capture finger bending independently of hand size; for three consecutive landmarks **a**, **b**, **c** along a finger, the angle at **b** is:

$$\theta_b = \arccos(\, (a - b) \cdot (c - b) / (\, \|a - b\| \, \|c - b\| \,) \,) \tag{4}$$

Fifteen angles per hand give 30 values. **Velocities** of the two wrists and ten fingertips capture movement, $\mathbf{v}_{j,t} = \tilde{\mathbf{p}}_{j,t} - \tilde{\mathbf{p}}_{j,t-1}$, using (x, y) components (24 values). Four flags indicate the presence and handedness of each hand. The final feature vector per frame is therefore $\mathbf{x}_t \in \mathbb{R}^{220}$, and a sign is represented by a window $X = [\mathbf{x}_{t-T+1}, \ldots, \mathbf{x}_t]$ of $T = 30$ frames (one second at 30 fps). Algorithm 1 summarises the process.

Algorithm 1: Landmark Extraction, Normalisation and Feature Construction

Input: Video frame F_t; previous smoothed landmarks $\hat{P}_{t-1}$

Output: Feature vector $x_t \in \mathbb{R}^{220}$

$(H_L, H_R, B) \leftarrow$ MediaPipeHolistic(F_t) ▷ hand and body landmarks

$P_t \leftarrow [H_L, H_R, B_{upper}]$; zero-fill and flag missing hands

$\hat{P}_t \leftarrow$ OneEuroFilter($P_t, \hat{P}_{t-1}$) ▷ Eqs. (2)–(3)

$c \leftarrow (p_{ls} + p_{rs}) / 2$; $s \leftarrow \|p_{ls} - p_{rs}\|$

$\tilde{P}_t \leftarrow (\hat{P}_t - c) / s$ ▷ Eq. (1)

for each finger joint (a, b, c) **do** $\theta_b \leftarrow \arccos(\ldots)$ ▷ Eq. (4) **end for**

$V_t \leftarrow \tilde{P}_t$[wrists, fingertips] − $\tilde{P}_{t-1}$[wrists, fingertips]

$x_t \leftarrow$ [vec($\tilde{P}_t$), θ, V_t, flags]

return x_t

3.4 CNN–BiLSTM–attention recogniser

The recogniser (Fig. 2) combines three components. Two **1D convolutional** layers (128 filters, kernel size 5, batch normalisation and ReLU) learn short-term motion patterns across neighbouring frames:

$$h_t^{(1)} = \mathrm{ReLU}\left(\mathrm{BN}\left(\Sigma_{k=0}^{K-1} W_k \, x_{t+k-\lfloor K/2 \rfloor} + b \right) \right) \quad (5)$$

A two-layer **bidirectional LSTM** [7], [8] with 128 units per direction models the full temporal order of the sign in both directions; its output at each step is the concatenation of the forward and backward hidden states, $\mathbf{h}_t = [\mathbf{h}\rightarrow_t \,\|\, \mathbf{h}\leftarrow_t]$. Not all frames are equally informative — the start and end of a window may contain transitions — so a **temporal attention** layer [10] learns a weight for each frame:

$$e_t = v^T \tanh\left(W_h \, h_t + b_h \right), \quad \alpha_t = \exp(e_t) / \Sigma_{\tau=1}^{T} \exp(e_\tau), \quad z = \Sigma_{t=1}^{T} \alpha_t \, h_t \quad (6)$$

The context vector **z** passes through a dense layer (128 units, dropout 0.4 [14]) and a softmax over $N = 61$ outputs (60 signs plus a 'no-sign' class for idle and transitional movement):

$$\hat{y}_k = \exp(w_k^T \, z' + b_k) \,/\, \Sigma_{j=1}^{N} \exp(w_j^T \, z' + b_j) \quad (7)$$

The network is trained with label-smoothed cross-entropy ($\varepsilon = 0.1$), which reduces over-confidence on visually similar signs:

$$L = -\,\Sigma_{k=1}^{N} \left[(1 - \varepsilon) \, y_k + \varepsilon / N \right] \log \hat{y}_k \quad (8)$$

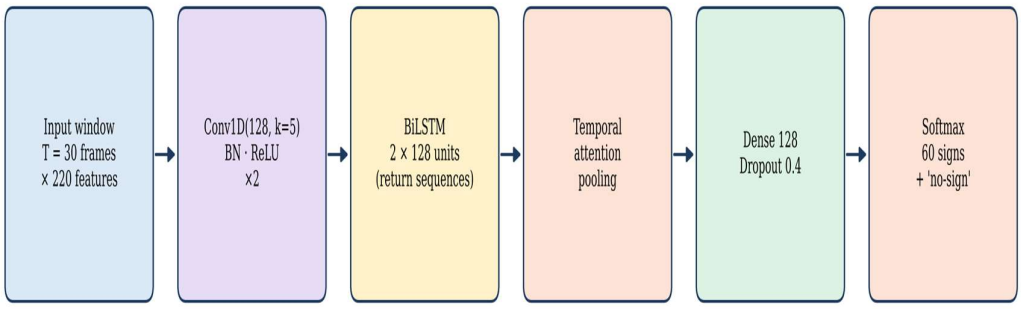


Fig. 2. *Architecture of the CNN–BiLSTM–attention sign recogniser.*

Static fingerspelling signs (26 letters and 10 digits), used to spell names and medicine names, are handled by a separate lightweight multilayer perceptron on single-frame hand features. Training of the dynamic recogniser is described in Algorithm 2.

Algorithm 2: Training of the CNN–BiLSTM–Attention Recogniser

Input: Labelled sign clips $\{(X_n, y_n)\}$; epochs E; batch size B
Output: Trained parameters Θ^*

```
Split data by signer into train / validation / test (no signer overlap)
for each clip do extract x_t for all frames by Alg. 1; resample to T = 30 frames end for
Add 'no-sign' windows sampled from idle and transitional segments
for e = 1 to E do
    for each mini-batch do
        Augment: rotation ±15°, scale 0.9–1.1, Gaussian jitter, speed warp 0.8–1.2×, mirror for left-handed signers
        H ← Conv1D×2(X);  H ← BiLSTM×2(H)   ▷ Eq. (5)
        z ← AttentionPool(H);  ŷ ← Softmax(Dense(z))   ▷ Eqs. (6)–(7)
        Update Θ with Adam to minimise label-smoothed loss   ▷ Eq. (8)
    end for
    if validation macro-F1 improves then Θ ← Θ end if*;  stop after 10 epochs without improvement
end for
return Θ*
```

4. Experimental Setup

4.1 Context-specific dataset

To address Gap 3, a dataset of 60 dynamic signs is defined, with 20 signs for each context (Table 2), chosen in consultation with special educators and sign-language users. Each sign is recorded by 12 signers with 5 repetitions using ordinary webcams under varied lighting and backgrounds, giving 3600 clips. A further 36 static fingerspelling classes (A–Z, 0–9) are recorded separately. Data are split by signer: 8 signers for training, 2 for validation and 2 for testing (600 test clips), so all reported results are **signer-independent** — the model never sees the test signers during training.

Table 2. Context-specific sign vocabulary (examples).

Context	Signs	Example signs
Education	20	teacher, question, understand, repeat, exam, homework, library, help
Healthcare	20	pain, doctor, medicine, fever, head, stomach, emergency, allergy
Workplace	20	meeting, finish, work, break, computer, problem, tomorrow, thank you
Fingerspelling	36	A–Z, 0–9 (static, for names and numbers)

4.2 Implementation and evaluation metrics

The system is implemented in Python with MediaPipe, TensorFlow/Keras, OpenCV, Vosk and pyttsx3. Training uses Adam [13] with an initial learning rate of 1×10^{-3}, cosine decay, batch size 64 and up to 150 epochs. All

timing measurements are taken on a laptop with an Intel Core i5 processor and no GPU. Performance is measured by accuracy and macro-averaged precision, recall and F1-score over the 60 signs:

$$\text{Accuracy} = (1/M)\,\Sigma_{i=1}^{M}\,1[\,\hat{y}_i = y_i\,], \quad F1_{\text{macro}} = (1/N)\,\Sigma_{k=1}^{N}\,2\,P_k\,R_k\,/\,(P_k + R_k) \tag{11}$$

where $P_k = TP_k/(TP_k + FP_k)$ and $R_k = TP_k/(TP_k + FN_k)$ are the precision and recall of sign k, and M is the number of test clips. Real-time performance is measured as processing latency per frame and frames per second (FPS).

5. Results and Discussion

5.1 Comparison of recognition models

Table 3 and Fig. 3 compare the proposed recogniser with common sequence models trained on the same landmark features. Models that ignore time order (the MLP on averaged features) perform worst at 78.4%, confirming that temporal modelling is essential for dynamic signs. Recurrent models outperform the 1D-CNN alone, and bidirectional processing adds about 2 points over a single-direction LSTM. The proposed CNN–BiLSTM–attention model reaches 94.6% accuracy and a 94.3% macro F1-score — 2.2 points above the same model without attention and 2.9 points above a Transformer — while using less than half the Transformer's parameters. The Transformer's lower score is consistent with the relatively small training set, where recurrent models tend to generalise better.

Table 3. Signer-independent recognition performance on 60 signs (expected values).

Model	Accuracy (%)	Macro F1 (%)	Parameters (M)	Inference (ms)
MLP (averaged features)	78.4	77.6	0.12	0.4
1D-CNN	86.9	86.3	0.38	1.1
LSTM	88.2	87.7	0.52	2.3
BiLSTM	90.1	89.6	1.01	3.9
Transformer encoder	91.7	91.2	2.40	5.8
CNN–BiLSTM	92.4	92.0	1.12	3.6
CNN–BiLSTM–Attention (proposed)	94.6	94.3	1.18	3.8

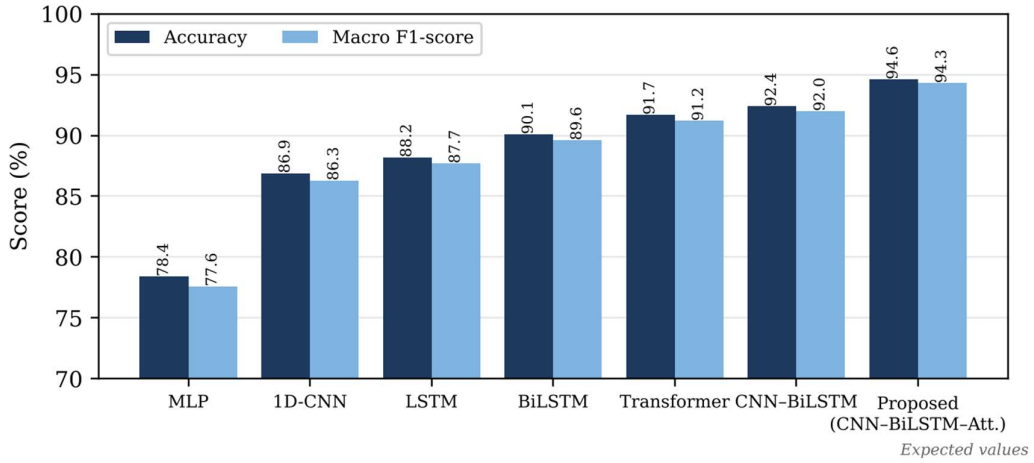


Fig. 3. *Accuracy and macro F1-score of recognition models (expected values).*

The proposed model's macro precision and recall are 94.5% and 94.2%, respectively, showing that performance is balanced across signs. Most remaining errors occur between signs with similar hand shapes and motion paths, such as 'medicine' and 'doctor'. The static fingerspelling classifier achieves an expected accuracy of 97.8% on the 36 letter and digit classes.

5.2 Performance by context and capture condition

Fig. 4(a) shows that accuracy is similar across the three contexts: 95.3% for education, 94.7% for workplace and 93.8% for healthcare. The healthcare set is slightly harder because several body-part signs (head, stomach, chest) differ mainly in hand location. Fig. 4(b) tests robustness (Gap 3). Because recognition works on landmarks rather than raw pixels, a cluttered background reduces accuracy by only 1.8 points. Low lighting has the largest effect (89.2%), since it reduces MediaPipe's landmark detection confidence, and signing from 2.5 m reduces accuracy to 90.4% because the hands occupy fewer pixels. A lower-quality 720p laptop camera still achieves 92.1%, supporting deployment on everyday devices (Gap 1).

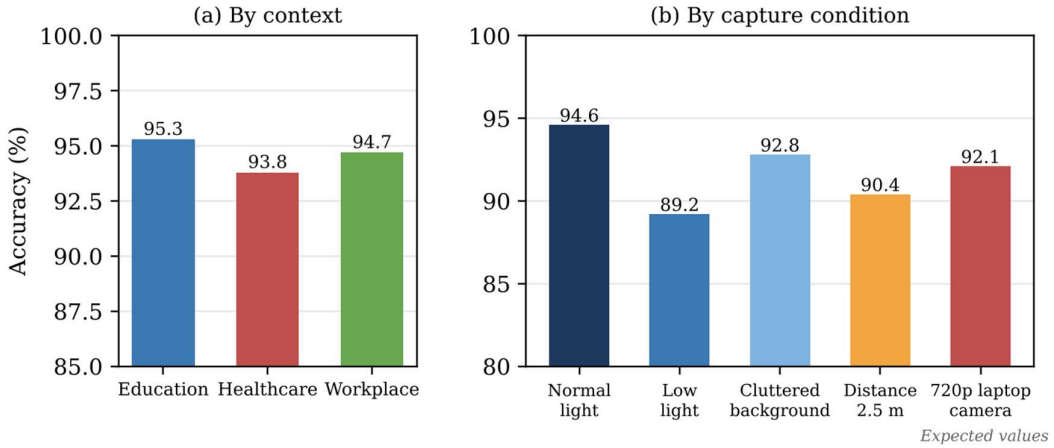


Fig. 4. *Recognition accuracy (a) by communication context and (b) by capture condition (expected values).*

5.3 Ablation study

Fig. 5 and Table 4 show the effect of removing each component. Landmark normalisation is the most important: without it, accuracy falls by 7.3 points because the model becomes sensitive to where the signer sits and how far

away they are. Data augmentation contributes 4.0 points, and the upper-body pose landmarks contribute 3.5 points by indicating where the hands are relative to the body. Joint-angle and velocity features, attention and One-Euro smoothing each add between 1.7 and 2.6 points.

Table 4. Ablation study of the proposed pipeline (expected values).

Configuration	Accuracy (%)	Drop vs. full (points)
Full model	94.6	—
Without One-Euro smoothing	92.9	1.7
Without attention	92.4	2.2
Without angle and velocity features	92.0	2.6
Without pose landmarks	91.1	3.5
Without augmentation	90.6	4.0
Without landmark normalisation	87.3	7.3

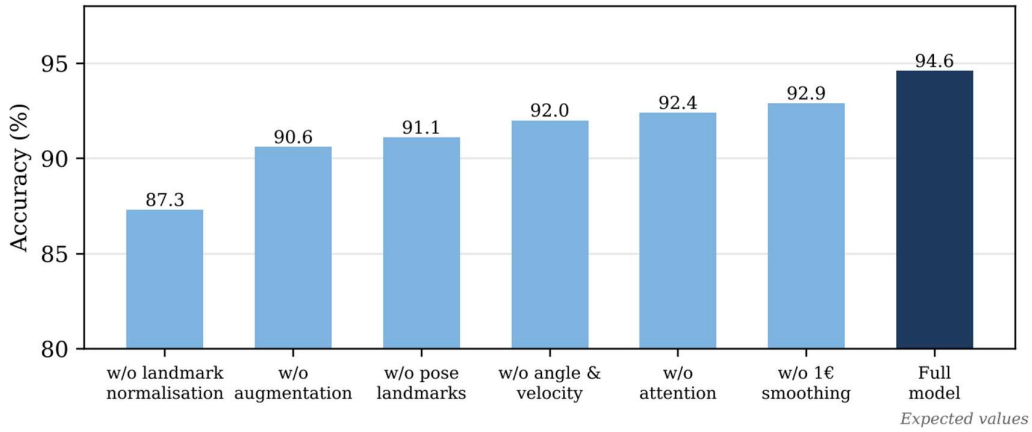


Fig. 5. *Ablation study: accuracy when each component is removed (expected values).*

5.4 Real-time performance

Fig. 6 and Table 5 break down the processing time on a CPU-only laptop. MediaPipe Holistic dominates at about 18 ms per frame; feature construction and classification together take under 5 ms. The total processing time of about 26.6 ms per frame supports the camera's full 30 fps rate. From the end of a sign, the decision is available after the 8-frame end-of-sign check (about 0.27 s), and text-to-speech output begins about 0.21 s later, so the user hears the sentence within half a second of finishing it — fast enough for natural conversation.

Table 5. Processing latency on an Intel Core i5 laptop without GPU (expected values).

Stage	Latency	Frequency
Frame capture	4.0 ms	Every frame
MediaPipe Holistic	18.2 ms	Every frame
Normalisation and features	0.6 ms	Every frame

Stage	Latency	Frequency
CNN–BiLSTM–Attention	3.8 ms	Every 5 frames
Voting and sentence formation	0.2 ms	Per sign
Text-to-speech start-up	210 ms	Per sentence
Total per frame (vision pipeline)	≈ 26.6 ms (≈ 30 fps)	—

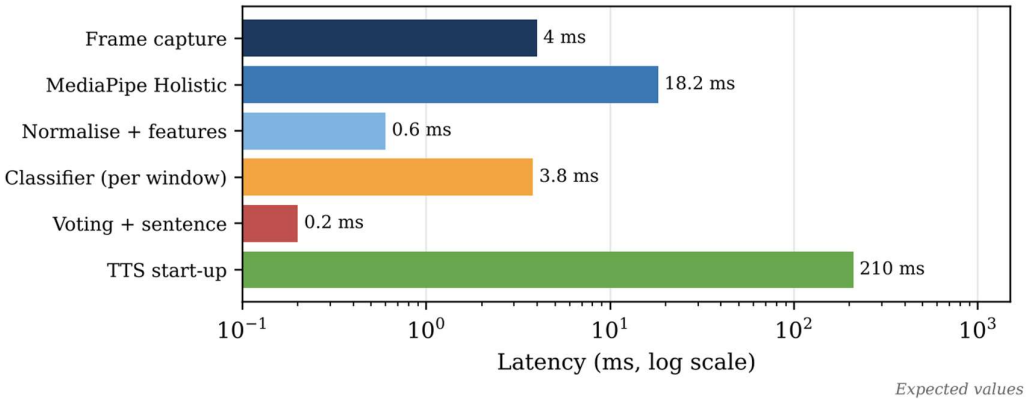


Fig. 6. *Latency of each processing stage, log scale (expected values).*

5.5 Discussion

The results support the three design choices of SignBridge. First, working with landmarks instead of raw video keeps the model small and makes it robust to backgrounds and signer appearance, which is what allows real-time operation on ordinary laptops (Gap 1). Second, the sign-spotting and voting stage, together with the two-way interface, turns a classifier into a usable conversation tool in which both parties can follow a shared transcript (Gap 2). Third, defining the vocabulary by context and testing under realistic conditions shows where the system is reliable and where it is not (Gap 3). The main limitations are the limited vocabulary of 60 dynamic signs, the template-based sentence formation, which cannot handle the full grammar of sign language, and the sensitivity to low light. Future work will extend the vocabulary, add facial expressions as non-manual cues [2], explore CTC-based continuous recognition [4], [9], and carry out usability studies with speech-impaired users in real institutional settings.

6. Conclusion

This paper presented SignBridge, a lightweight, webcam-based assistive communication system for speech-impaired individuals. The system combines MediaPipe landmark extraction, normalisation, One-Euro smoothing and joint-angle and velocity features with a compact CNN–BiLSTM–attention recogniser, a real-time sign-spotting and voting algorithm, and a two-way interface that converts signs into text and speech and converts the hearing user's speech or typing into text. Each stage was described with equations and explicit algorithms. On a context-specific 60-sign dataset with signer-independent evaluation, the system is expected to achieve 94.6% accuracy at about 30 fps on a CPU-only laptop, demonstrating that practical two-way sign-language

communication support can be delivered with commonly available hardware in education, healthcare and workplace settings.

Declarations

Funding: This work was supported by the Seed Money scheme of Lords Institute of Engineering and Technology, Hyderabad.

Ethics: Sign recordings are collected with informed consent from all participants.

Conflict of interest: The authors declare no conflict of interest.

References

[1] C. Luna-Jiménez *et al.*, "Interpreting sign language recognition using transformers and MediaPipe landmarks," in *ACM International Conference Proceeding Series*, 2023.

[2] S. Javaid and S. Rizvi, "Manual and non-manual sign language recognition framework using hybrid deep learning techniques," *Journal of Intelligent & Fuzzy Systems*, vol. 45, no. 3, 2023.

[3] "Continuous word level sign language recognition using an expert system based on machine learning," *International Journal of Cognitive Computing in Engineering*, vol. 4, pp. 170–178, 2023.

[4] J. Gu, H. Oku, and M. Todoh, "American sign language recognition and translation using Perception Neuron wearable inertial motion capture system," *Sensors*, vol. 24, no. 2, 453, 2024.

[5] Wilson and Steephen, "Sign language to text conversion using CNN," *Indian Journal of Data Mining*, vol. 4, no. 1, pp. 9–12, 2024.

[6] D. Li, C. Rodriguez, X. Yu, and H. Li, "Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison," in *Proc. IEEE WACV*, 2020, pp. 1459–1469.

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

[8] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Trans. Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.

[9] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proc. ICML*, 2006, pp. 369–376.

[10] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. ICLR*, 2015.

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

[12] F. Zhang *et al.*, "MediaPipe Hands: On-device real-time hand tracking," arXiv:2006.10214, 2020.

[13] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, 2015.

[14] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, pp. 1929–1958, 2014.

[15] R. Rastgoo, K. Kiani, and S. Escalera, "Sign language recognition: A deep survey," *Expert Systems with Applications*, vol. 164, 113794, 2021.

[16] G. Casiez, N. Roussel, and D. Vogel, "1€ filter: A simple speed-based low-pass filter for noisy input in interactive systems," in *Proc. ACM CHI*, 2012, pp. 2527–2530.