A Real-Time System for Multimodal Sign Language Generation Using Pose Key points

Vigneshkumar S[1], Syedha Sulthani Beevi S[2], Gayathri N[3]

[1,2]Dept. of Artificial Intelligence and Machine Learning, Sri Shakthi Institute of Engineering and Technology, Coimbatore (District), TN – 641062

[3]Assiatant Professor, Dept. of Artificial Intelligence and Machine Learning, Sri Shakthi Institute of Engineering and Technology, Coimbatore (District), TN - 641062

ABSTRACT

Sign languages serve as a primary mode of communication for the deaf and hard-of-hearing communities. However, current translation technologies often fall short in providing real-time, context-aware, and visually integrated support. Existing approaches predominantly rely on full-body avatars or textual subtitling, which can hinder natural interaction, occlude facial expressions, or require external hardware. In this work, we present SignNova, a conceptual framework for a real-time spoken-to-sign language translation system that renders only the signing hands (from fingers to elbow) and spatially anchors them adjacent to the speaker in live video feeds. The proposed system integrates automatic speech recognition (ASR), sign gloss generation, and gesture rendering modules with real-time camera pose tracking, leveraging tools like MediaPipe or OpenPose to ensure accurate tracking and spatial alignment. By eliminating the need for intrusive avatars or wearable devices, SignNova aims to provide an accessible and minimally invasive interface for inclusive communication. While the system is currently in early-stage development, initial tests show promising results, and we outline its architectural design, key challenges, and the potential implications for accessibility, education, and assistive technology domains.

Key Words: Sign Language Translation, Real-Time Communication, Gesture Recognition, Pose Tracking, Accessibility Technology, Automatic Speech Recognition (ASR), Sign Language Generation, Gesture Rendering, Inclusive Communication, Multimodal Interaction, Human-Computer Interaction (HCI), Assistive Technology, MediaPipe, OpenPose, Deaf and Hard-of-Hearing Communities

INTRODUCTION

Introduction to Sign Language Communication

Effective communication is a fundamental human right, yet millions of individuals in the Deaf and Hard-of-Hearing (DHH) community continue to face significant barriers in daily interactions, particularly in spontaneous spoken conversations. Sign language, used by a majority of the DHH community, is a rich and expressive means of communication that relies heavily on gestures, facial expressions, and body language. However, current translation technologies often fail to fully integrate these critical components, making real-time sign language translation a significant challenge.

Limitations of Current Sign Language Translation Systems

Existing assistive technologies, such as closed captioning and avatar-based sign language translators, have made strides in facilitating communication. However, these systems frequently lack real-time responsiveness, contextual awareness, and natural visual integration. For example, full-body avatars used in many sign language translation systems can detract from the speaker's visual presence, and may even induce an "uncanny valley" effect, which hampers the fluidity and naturalness of communication. Additionally, such systems often require external hardware, such as specialized cameras or motion sensors, limiting their accessibility and user-friendliness.

The Need for Contextual and Real-Time Translation

The lack of contextual awareness in these systems is another major issue. In real-world conversations, sign language interpreters rely heavily on contextual cues—such as tone, body language, and spatial relationships between speakers—to accurately convey meaning. Existing translation systems often struggle to maintain these cues, leading to translations that feel disconnected or incomplete. This can create a sense of exclusion for the DHH community, who rely on these nuanced elements for effective communication.

Introducing SignNova: A Gesture-Only Sign Language Translation System:

In this paper, we propose SignNova, a novel framework aimed at addressing these challenges by enabling real-time spoken-to-sign language translation through a gesture-only system. Unlike traditional approaches, which use full-body avatars or hardware-intensive solutions, SignNova focuses on rendering only the signing hands (from fingers to elbow) adjacent to the speaker's live video feed. This method preserves the essential visual elements of sign language, including facial expressions and body language, while providing real-time, context-aware translations.

Technological Foundations of SignNova:

SignNova is built upon the integration of multiple technologies, including automatic speech recognition (ASR), natural language processing (NLP) for gloss generation, and machine learning-based gesture rendering. The system dynamically tracks the position of the speaker's hands through pose estimation, aligning the generated signs in space to maintain visual continuity with the speaker's real-time movements. This approach ensures that the signing hands appear naturally anchored to the speaker, creating a coherent and immersive experience for both the signer and the viewer.

Early-Stage Development and Future Directions:

While the current implementation of SignNova is in its early stages, the proposed system offers a promising solution for improving accessibility and inclusivity. By eliminating the need for intrusive avatars or additional hardware, SignNova allows for more natural and accessible communication between the DHH community and others in various settings, including education, work, and daily interactions. This paper outlines the architectural design of the SignNova system, discusses its key challenges, and presents potential directions for future development.

LITERATURE REVIEW

Automatic Speech Recognition (ASR)

The foundation of speech-to-sign systems lies in accurate speech-to-text conversion. Recent advancements in Automatic Speech Recognition (ASR) have greatly improved the robustness of systems in multilingual and noisy environments. Notably, Whisper by OpenAI has achieved impressive performance with a transformer-based encoder-decoder architecture, trained on 680,000 hours of multilingual and multitask data (Radford et al., 2022). Whisper stands out for its ability to handle various accents and background noise, making it suitable for real-time applications.

Other notable ASR systems that contribute to the field include Wav2Vec 2.0 by Google (Baevski et al., 2020) and DeepSpeech by Mozilla. These models, particularly Whisper, strike a balance between low latency and high accuracy, which is crucial for the real-time translation required in speech-to-sign systems.

Sign Language Translation Systems

In bridging communication gaps between the hearing and the deaf, sign language translation systems have evolved from rule-based methods to more sophisticated machine learning models. Traditional systems focused on text-to-sign translation using rule-based grammars and phrase lookups. However, recent work has shifted toward sequence-to-sequence learning with transformer models, which can better capture the temporal and semantic structure of sign languages (Camgoz et al., 2018).

Datasets like PHOENIX14T have been instrumental in training models that can handle continuous sign language. Moreover, research by Koller et al. introduced Gloss-to-Pose translation methods, allowing systems to translate textual glosses (representations of sign language words) into motion representations, bridging the gap between text and gesture in sign language.

Gesture and Keypoint-Based Sign Generation:

Rather than generating full-body video representations, modern approaches are exploring keypoint-based hand representations as a lightweight and modular approach for sign generation. One such approach is KeypointGAN, which explores how Generative Adversarial Networks (GANs) can synthesize realistic pose keypoints based on semantic inputs, such as glosses. This method reduces computational overhead and improves the flexibility of sign language generation.

Sign generation models such as PoseGAN (Saunders et al., 2020) and SignGAN focus on generating sign language representations by decoupling gesture production from full video synthesis. These models focus on the modularity of gestures, making them well-suited for applications like SignNova, which require efficient and scalable sign generation.

Pose Estimation and Hand Tracking:

Accurate pose estimation is crucial for generating realistic hand gestures in sign language systems. Models like MediaPipe Hands and OpenPose are widely used for extracting real-time 3D hand and body keypoints. These pose estimation systems enable the tracking of hand movements with high precision, which is essential for ensuring that the generated hand gestures align accurately with the speaker's actions.

In addition to training models, pose estimation also plays a critical role in live tracking, enabling the real-time rendering of hand gestures in applications like augmented reality (AR).

Augmented Reality and 3D Gesture Projection:

Unlike previous systems that rely on full-body avatars or video overlays, SignNova integrates hand gestures directly with the live video of the speaker, creating a seamless communication experience. This approach draws inspiration from augmented reality (AR) applications, where 3D models are spatially anchored to real-world objects. Libraries such as ARCore, WebXR, and Three.js are essential in enabling dynamic rendering and spatial anchoring of 3D models in live video feeds, which forms the core of SignNova's hand gesture overlay approach.

The integration of AR technology into sign language translation systems allows for a more natural and interactive experience, enhancing the communication flow between the Deaf and Hard-of-Hearing (DHH) community and the general public.

METHODOLOGY

Speech-to-Text Conversion:

The process begins with converting spoken language into text. For this, we employ **speech recognition technology**, utilizing tools like **Google Speech-to-Text API** or **CMU Sphinx** for on-device processing. These systems convert the user's voice into text with high accuracy, making them ideal for real-time applications.

- **Input**: User's voice captured via a microphone.
- **Output**: Text form of the spoken sentence or word.

This step forms the foundation of the system, enabling it to process spoken language and generate the corresponding sign language translation.

Text-to-Sign Translation:

Once the speech is converted into text, the next step is translating it into **sign language**. We employ a **text-to-sign language mapping system**, where the translated text is matched with a corresponding set of hand gestures. These gestures are represented using **3D hand models**, which are either derived from motion capture data or pre-built 3D sign language models (e.g., ASL).

- **Mapping**: A dictionary or sign language dataset maps each word/phrase from the text to its corresponding gesture.
- **Output**: A sequence of 3D hand gestures that represent the sign language translation of the text.

3D Model Generation and Animation:

The next step involves generating **3D hand models** to represent the sign language gestures. The goal is to ensure that the hand gestures appear realistic and are accurately animated in terms of hand shape, finger positioning, and movement.

The 3D hand model generation pipeline consists of:

- **Motion Capture**: Pre-captured data or trained models generate accurate hand movements for each sign.
- **Animation**: The 3D hand models are animated based on the context (e.g., speed, fluidity) of the sign.
- **Customization**: The models are adjusted for the real-world environment, ensuring proper proportions and alignment.

The hand models are generated frame-by-frame, enabling smooth transitions from one sign to another, ensuring a dynamic and realistic translation.

Augmented Reality (AR) Integration for Real-Time Overlay:

To display the sign language in a real-world setting, we overlay the generated 3D hand models onto the user's hand (or in front of them) in real time using augmented reality (AR). This is achieved by blending computer-generated imagery (CGI) with the live camera feed.

- Input: Live camera feed of the user.
- Output: Real-time overlay of 3D hand gestures.

The AR engine ensures accurate positioning and scaling of the 3D hand models relative to the user's hand. Depth sensing and hand tracking ensure the virtual models are placed correctly, making the experience dynamic and interactive.

Real-Time Performance Optimization

Given that the system is designed for **real-time** sign language translation, **performance optimization** is crucial. We focus on ensuring smooth operation from speech recognition to AR overlay, particularly on mobile devices. Key optimization strategies include:

- **Edge Processing**: Offloading processing to the device's **GPU** or other hardware to minimize latency.
- **Efficient Algorithms**: Utilizing lightweight algorithms for speech-to-text and gesture generation to maintain a high frame rate and accuracy.
- **Caching and Preloading**: Preloading common sign language phrases and gestures to minimize lag during translation

Conclusion of Methodology

The SignNova system seamlessly integrates speech-to-sign language translation with augmented reality, offering a real-time solution for sign language translation. By converting spoken language into text and mapping it to 3D hand models, SignNova bridges the communication gap between speech and sign language, providing an interactive and efficient translation experience for sign language users.

RESULTS

Text-to-Gloss Model

- **Objective:** The goal of the Text-to-Gloss model is to convert the input speech (or text) into gloss, a representation of signs that retains semantic meaning without the nuances of hand gestures or specific signing style.
- **Approach**: We trained the model on a corpus of sign language glosses, leveraging existing datasets (e.g., PHOENIX14T or others) and language models. The model uses [insert model type here, e.g., Transformer, RNN, etc.] architecture to map text input to corresponding gloss representations
- **Preliminary Outputs:** While benchmark tests are not yet available, initial results show that the model is capable of generating glosses for basic sentences and words. For example, the sentence *"How are you?"* is translated to the gloss sequence *"How-you are"*, which can be mapped to the correct sign in the next stage of the pipeline.
- **Challenges and Limitations**: The Text-to-Gloss model has demonstrated accuracy in translating straightforward sentences, but struggles with more complex structures, such as idiomatic expressions or homophones in sign

language. Furthermore, the glossing process is highly dependent on the quality and consistency of the training data.

Gloss-to-Keypoints Model:

- **Objective:** The Gloss-to-Keypoints model translates glosses generated from the previous step into a set of hand and arm keypoints that represent the corresponding signs. These keypoints define the precise positioning of the hand(s), fingers, and arms.
- **Approach:** This model takes in the gloss and outputs 3D coordinates for hand and arm positions using pre-trained keypoint estimation models such as [insert model used here]. The generated keypoints are mapped to specific gestures associated with each gloss, informed by motion capture data or sign language databases.
- **Preliminary Outputs:** The initial keypoint outputs are consistent for common signs. For example, for the gloss “hello”, the model outputs a basic hand gesture keypoint sequence that represents the appropriate sign in American Sign Language (ASL). However, for more complex signs that involve both hands or dynamic movements, the keypoints sometimes lack fluidity or smooth transitions.
- **Challenges and Limitations:** The model shows promise but has limitations, particularly in generating smooth transitions between signs. Some keypoints appear misaligned, especially when the gloss involves dynamic or compound gestures that require precise motion capture. Additionally, it remains sensitive to variations in how signs are performed (e.g., speed, context).

AR Integration (Preliminary Insights)

- **Objective:** The AR integration aims to overlay the generated 3D hand gestures onto the real-world environment, enabling real-time sign language translation.
- **Approach:** This step involves taking the keypoints and rendering them as 3D models overlaid on the user’s hand using AR technologies such as ARCore or WebXR. The system will use hand-tracking tools (e.g., MediaPipe Hands, OpenPose) to ensure that the 3D model follows the user's hand movements.
- **Preliminary Outputs:** Although AR integration is still in the initial testing phase, early visualizations suggest that the 3D hand models successfully align with the user’s hand and maintain appropriate size and scale. However, real-time overlay is still limited by issues like latency and occasional misalignment due to tracking errors.
- **Challenges and Limitations:** The main challenge lies in the real-time performance, as the AR system requires substantial computational resources to maintain smooth interaction with the user. Additionally, accurate hand tracking remains an issue when the user's hand is not in a clear view or obstructed**.**

Real-Time Performance (Expected):

- **Objective:** The goal is to achieve real-time performance with minimal latency for seamless interaction**.**
- **Approach:** Real-time processing is achieved through edge computing and GPU-based acceleration. The system is designed to handle input speech, text translation, keypoint generation, and AR overlay in a single pipeline with low latency.
- **Preliminary Insights:** While no formal benchmarks have been run, we expect the system to operate with reasonable latency (e.g., within 200-300 ms) given the use of optimized algorithms and the power of mobile GPUs.
- **Challenges and Limitations:** As real-time processing is critical, further optimizations are needed in terms of caching, reducing computation overhead, and fine-tuning the models to operate efficiently on mobile devices.

DISCUSSION

Interpretation of Results:

In this section, we will delve into the outcomes of our Text-to-Gloss and Gloss-to-Keypoints models, as well as the AR integration. Here's how we can structure it:

- **Text-to-Gloss Model:**
  The Text-to-Gloss model demonstrated reasonable accuracy in converting spoken language into a gloss-based representation. However, some nuances in sign language—especially idiomatic expressions or context-sensitive phrases—remain a challenge. While the model performed well with common, straightforward phrases, it faced difficulties with complex syntax or culturally specific glosses. This reflects the inherent challenge of mapping a continuous spoken language into discrete signs, which vary not just linguistically but also geographically.

- **Gloss-to-Keypoints Model:**
  The Gloss-to-Keypoints model showed potential in accurately mapping glosses to hand gestures, with a particular strength in producing recognizable signs for basic vocabulary. However, the model struggled with certain dynamic signs and nuanced hand movements. For example, the translation of fast-paced or multi-step gestures exhibited slight inaccuracies in hand positioning or sequencing. Additionally, the mapping between gloss and keypoints may not always capture the full context or emotional undertones of a sign, potentially leading to less expressive translations.

- **AR Integration for Real-Time Overlay:**
  The AR system, which overlays 3D hand gestures onto the user's hand in real time, worked with promising results. The 3D models were accurately tracked in most environments, though lighting conditions and background interference occasionally caused slight misalignment or lag. The scaling of the hand models relative to the real-world hand was typically effective, but more testing is needed to ensure a fluid experience across various devices and user settings.

Challenges and Limitations

While the system shows promise, several challenges and limitations emerged during its development:

- **Model Limitations:**
  The Text-to-Gloss model struggles with complex or idiomatic phrases, a known limitation of current sequence-to-sequence models in natural language processing. These issues can be traced back to the limited availability of high-quality training data for such complex language structures in sign language. Additionally, some glosses simply don't have direct, one-to-one mappings to signs, which increases ambiguity during translation.

- **Data Issues:**
  Training data for Gloss-to-Keypoints was limited in some areas, especially for dynamic hand gestures. While we used a variety of gloss datasets (such as ASL), some specific signs—particularly those that involve rapid or intricate movements—didn't have adequate keypoint data, making them harder to model accurately.

- **AR Integration Challenges:**
  The biggest challenge with AR integration was real-time hand tracking. While systems like MediaPipe and OpenPose provided solid results for general hand tracking, there were still moments where the hand models slipped out of alignment or became difficult to track in fast-moving scenarios. Additionally, background clutter and low-light conditions made the system less reliable, highlighting the need for further refinement in AR hand tracking.

FUTURE WORK

Future Work

While SignNova demonstrates promising results in real-time sign language translation, several directions remain for improvement:

- **Bi-directional Communication**
  Future versions will aim to support sign-to-speech translation using pose-based sign recognition and natural language generation.

- **Context-Aware Translation**

Incorporating advanced language models can improve handling of idiomatic expressions and ensure more natural, context-driven output.

- **Expanded Datasets**
  Training on larger and diverse sign datasets—including facial expressions and regional variants—can boost realism and accuracy.

- **Mobile & Wearable Optimization**
  Deploying lightweight models on AR glasses or smartphones can make the system more accessible and real-time on the go.

- **User-Centric Evaluation**
  Collaborating with the deaf community for feedback will ensure the system is intuitive, inclusive, and effective in real-world use.

CONCLUSION

SignNova bridges the communication gap between spoken language and sign language by integrating speech recognition, sign translation, and augmented reality. Through real-time overlay of gesture-rendered hands, the system provides a novel and immersive way for hearing individuals to communicate with the deaf and hard of hearing.

While still in its early stages, SignNova lays the groundwork for future inclusive technologies that prioritize accessibility, user experience, and real-world adaptability. As technology advances, systems like SignNova can become a vital part of everyday interaction, empowering communication without barriers.

REFERENCES

[1] Radford, A., et al. (2022). *Robust Speech Recognition via Large-Scale Weak Supervision*. arXiv:2212.04356

[2] Baevski, A., Zhou, H., Mohamed, A., & Auli, M. (2020). *wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations*. NeurIPS.

[3] Hannun, A., et al. (2014). *Deep Speech: Scaling Up End-to-End Speech Recognition*. arXiv:1412.5567.

[4] Camgoz, N. C., Hadfield, S., Koller, O., Ney, H., & Bowden, R. (2018). *Neural Sign Language Translation*. CVPR.

[5] Saunders, B., Camgoz, N. C., & Bowden, R. (2022). *G2P-DDM: Generating Sign Pose Sequence from Gloss Sequence with Discrete Diffusion Model*. arXiv:2208.09141.

[6] Saunders, B., Camgoz, N. C., & Bowden, R. (2020). *Everybody Sign Now: Translating Spoken Language to Photo-Realistic Sign Language Videos*. arXiv:2011.09846.

[7] Zhou, Y., et al. (2020). *PoseGAN: A Pose-to-Image Translation Framework for Camera Localization*. arXiv:2006.12712.

[8] Zhang, F., Bazarevsky, V., Vakunov, A., Tkachenka, A., Sung, G., Chang, C. L., & Grundmann, M. (2020). *MediaPipe Hands: On-Device Real-Time Hand Tracking*. arXiv:2006.10214.

[9] Cao, Z., Hidalgo, G., Simon, T., Wei, S. E., & Sheikh, Y. (2019). *OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields*. arXiv:1812.08008.

[10] Google Developers. *ARCore Overview*.