Towards an AI-Based Multi-Agent Sign Language System Using Smart Rings and Multimodal Dialogue Detection

Jeong Woo LEE

March 2025

Abstract

This paper proposes a conceptual AI-based sign language system that integrates multimodal speaker identification with wearable smart rings. The proposed system addresses key limitations in existing sign language technologies, such as reliance on vision-based input, limited support for multi-speaker scenarios, and mediated interaction. By assigning unique sign language avatars to each speaker and allowing direct sign input via smart rings, this system offers a natural and inclusive communication experience for Deaf users in real-world environments such as dramas, education, and public speaking. Furthermore, the wearable nature of smart rings enables long-term, real-world sign language data collection, supporting personalization and future model adaptation.

1 Introduction

Sign language is an essential means of communication for the Deaf and hard-of-hearing community. Despite the growing interest in AI-driven sign language translation, most existing systems rely heavily on camera-based vision and support only single-speaker scenarios. This creates multiple problems in real-world applications:

- **Device dependency:** Users must face a camera, limiting mobility and freedom.
- **Limited speaker handling:** Systems struggle with multi-speaker situations common in dramas, sermons, and education.
- **Environmental constraints:** Lighting, occlusions, and background clutter degrade vision-based performance.
- **Mediated interaction:** Communication often feels indirect, with devices acting as intermediaries.

To overcome these limitations, this paper proposes a conceptual AI-based sign language system that integrates multimodal speaker identification and wearable smart rings. The key contributions are: (1) speaker-aware multi-agent sign language output, and (2) sensor-based sign input that enhances real-world usability, data availability, and user autonomy.

2 Related Work

Sign language recognition has traditionally relied on computer vision techniques such as OpenPose [2] and MediaPipe [3]. While effective in controlled environments, these methods are sensitive to noise and lighting. Sensor-based systems, such as smart gloves [1], provide more robust input, but are bulky and not socially acceptable for daily use.

For speaker identification, multimodal diarization tools like PyAnnote [5] and Whisper [6] enable robust speaker tracking by combining voice and visual cues [4]. However, prior work has not explored the combination of speaker-aware sign output with wearable sign input.

Recent studies on IMU and EMG integration for hand gesture recognition show promising potential in capturing dynamic hand movements accurately [7]. This strengthens the case for smart ring integration in real-world sign input systems.

3 Proposed System Design

3.1 System Overview

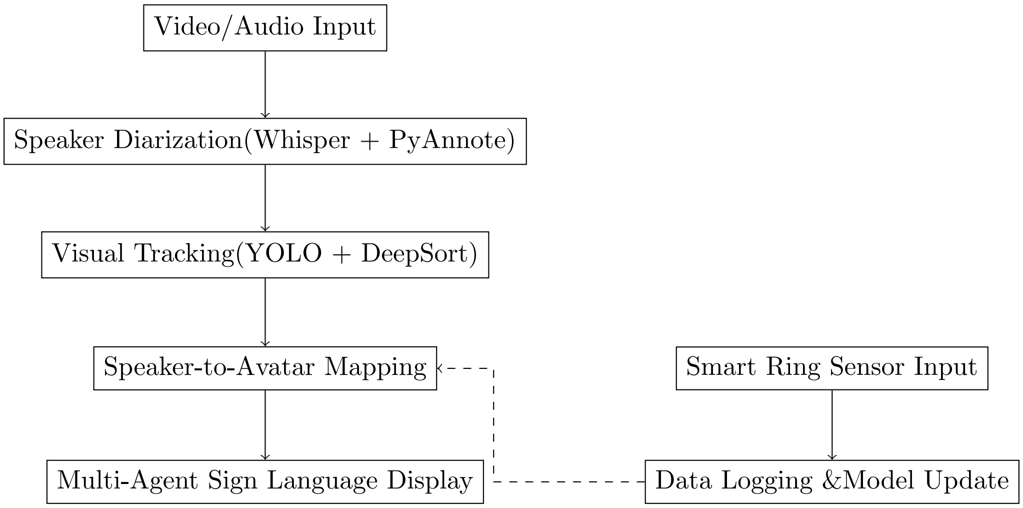
The proposed system addresses multi-speaker sign language translation through three components:

- **Multimodal Speaker Detection:** Uses audio diarization and face/body tracking to identify who is speaking.
- **Avatar Assignment:** Each speaker is matched with a distinct sign language avatar or animation style.
- **Smart Ring Input:** Allows Deaf users to input signs naturally using wearable IMU/EMG sensors [7]. The system can continuously learn from these inputs to improve over time.

3.2 System Architecture

1. Input: Multi-speaker video/audio stream
2. Apply Whisper [6] and PyAnnote [5] for speaker diarization
3. Use YOLO + DeepSort to track visual speaker identity
4. Map each speaker to an avatar with a unique signing style
5. Display synchronized multi-avatar sign output
6. Log smart ring input for model retraining and personalized adaptation

System Architecture Diagram


Video/Audio Input
Speaker Diarization(Whisper + PyAnnote)
Visual Tracking(YOLO + DeepSort)
Speaker-to-Avatar Mapping
Smart Ring Sensor Input
Multi-Agent Sign Language Display
Data Logging &Model Update

4 Use Cases and Scenarios

- **Drama and multimedia translation:** Each character signs with a unique avatar.
- **Sermons and public speaking:** Automatically assign avatars to multiple speakers.
- **One-on-one conversation:** Smart ring allows direct interaction without needing a camera.
- **Education:** Differentiates teacher and student sign roles.
- **Everyday interaction and learning:** Smart rings worn daily can unobtrusively collect sign input data in natural settings, supporting the development of personalized or regional sign models over time.
- **Accessibility in public environments:** In noisy places where voice input is difficult, sign input via smart ring can provide a silent, non-intrusive alternative.

5 Challenges and Future Work

- **Real-time performance:** Managing avatar synchronization and latency.
- **Sensor training:** Building a robust dataset for ring-based sign recognition.
- **Scalable dataset collection:** Continuous sign data input requires privacy-preserving and user-consented logging methods.
- **Sign expressiveness:** Capturing emotion, emphasis, and regional variation.
- **Personalized learning:** Daily smart ring usage opens opportunities for adapting models to individual users' styles.

6 Conclusion

This paper presents a novel conceptual system that combines multimodal speaker detection and wearable input to improve the usability of sign language AI systems. By matching avatars to speakers and enabling wearable sign input, the system promises more inclusive and natural communication experiences.

Furthermore, smart rings offer not only a convenient and discreet input method but also unlock the potential for continuous sign data collection in everyday life. This opens new avenues for training richer sign language models, supporting user-specific adaptation, and ultimately fostering a more accessible communication landscape.

This research contributes not only technically, but also socially, by enhancing communication equity and supporting the independence of Deaf users through unobtrusive, wearable AI technologies.

One limitation is the absence of a large-scale open dataset combining speaker diarization and sign input, which may affect generalizability. Future work includes building a benchmark dataset combining synchronized multi-speaker audio, video, and ring-based sign input. Additionally, user studies with Deaf participants will help validate usability and comfort of ring-based sign input.

References

[1] N. Singh, R. Sinha, and R. Singh, ”Sign language conversation interpretation using wearable sensors and machine learning,” *arXiv preprint arXiv:2312.11903*, 2023.

[2] I. Kim and I. Jeong, ”A Study on Korean Sign Language Motion Recognition Using Deep Learning Based OpenPose,” *Journal of the Korea Multimedia Society*, vol. 24, no. 5, pp. 593–600, 2021.

[3] Anonymous, ”Application of sign language gesture recognition using Mediapipe,” *Journal of Digital Contents Society*, vol. 22, no. 6, pp. 1087–1094, 2021.

[4] K. Kumatani, M. Omologo, and C. Fuegen, ”Multimodal integration for speaker diarization,” in *Proc. INTERSPEECH*, 2019.

[5] H. Bredin, ”pyannotate-audio: Neural speaker diarization toolkit,” [Online]. Available: `https://github.com/pyannote/pyannote-audio`

[6] OpenAI, ”Whisper: Robust Speech Recognition via Large-Scale Weak Supervision,” [Online]. Available: `https://openai.com/research/whisper`

[7] A. D. Roche, S. W. Tsang, C. Castellini, and B. M. Hill, ”Differences in Perspective on Inertial Measurement Unit Sensor Integration in Myoelectric Control,” *arXiv preprint arXiv:2003.03424*, 2020. Available: `https://arxiv.org/abs/2003.03424`