Gaze-Enhanced Multimodal Interaction for Accessibility

Md. Akil Raihan Iftee
Department of Computer Science and Engineering
Khulna University of Engineering & Technology, Bangladesh
iftee1807002@stud.kuet.ac.bd

I. Overview

We may propose an integrated system that enables hands-free, efficient GUI interaction by fusing gaze tracking, physiological triggers (EMG/breath), and voice commands. The system is aimed at users with limited motor function, using gaze to point, subtle biosignals to click, and voice to execute tasks. It is designed to reduce selection time and errors for accessible GUIs.

Here's how it works:

- The user looks at something on the screen, like a button. (gaze-point)
- They say a simple voice command, such as "Open this." or "Click that." or "Move This UP" while pointing to a button or "Delete this ICON" etc. (gaze-point + Voice-click)
- A small movement (like an eyebrow raise) or a breath confirms the action.(e.g. gaze-point + EMG-click)

II. System Overview

Our system comprises the following modules:

- **Eye-Tracking Processing:** Predicts gaze scanpaths and identifies UI elements being attended using the EyeFormer model [1] trained on VSGUI10K [2] and UEyes [3].
- **Voice Command Processing:** Transcribes and interprets spoken instructions (e.g., "open this") using an instruction-tuned vision-language model like ILUVUI [4].
- **Physiological Signal Processing:** Detects discrete activations (e.g., via EMG sensors for eyebrow raises or breath detection modules).
- **Multimodal Fusion & Decision:** Combines gaze, voice, and physiological cues to infer user intent and resolve ambiguities.
- **UI Controller:** Executes actions such as clicking or opening files.

III. Methodology

Figure 1 illustrates the interaction pipeline. Gaze input is processed to locate attention focus, while voice input captures semantic intent. Physiological sensors detect confirmation gestures. The system fuses these modalities for robust UI control. The EyeFormer model is used to predict GUI-level gaze behavior, while UEyes aids in fine-tuning personalized gaze patterns. VSGUI10K provides large-scale real-world GUI gaze data. Voice command understanding is augmented via ILUVUI's instruction-following capabilities.

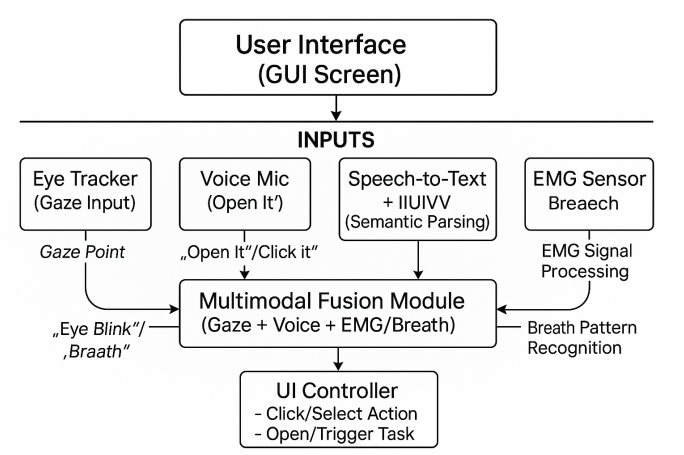

Fig. 1: The Framework of Proposed Gaze-Enhanced Multimodal Interaction

IV. Further Extension

Emotion-Aware Eye Tracking: Integrate real-time facial emotion recognition and vocal tone to optimize eye scanpath suited to the user's mood.

References

[1] Y. Jiang, Z. Guo, H. R. Tavakoli, L. A. Leiva, and A. Oulasvirta, "Eyeformer: Predicting personalized scanpaths with transformer-guided reinforcement learning," in *Proceedings of the ACM Symposium on User Interface Software and Technology (UIST)*, 2024.

[2] A. Putkonen, Y. Jiang, J. Zeng, O. Tammilehto, J. P. P. Jokinen, and A. Oulasvirta, "Understanding visual search in graphical user interfaces," 2025.

[3] Y. Jiang, L. A. Leiva, H. R. Tavakoli, P. R. B. Houssel, J. Kylmala, and A. Oulasvirta, "Ueyes: Understanding visual saliency across user interface types," in *Proceedings of the ACM CHI Conference on Human Factors in Computing Systems (CHI)*, 2023.

[4] Y. Jiang, E. Schoop, A. Swearngin, and J. Nichols, "Iluvui: Instruction-tuned language-vision modeling of uis from machine conversations," in *Proceedings of the ACM Conference on Intelligent User Interfaces (IUI)*, 2025.