ISI Voice Image: Voice-Controlled Clipboard Image Transformation

Frank Dierolf[1] Nadja Pirchheim[1]

[1]Ulm Institute of Spoken Intelligence (ISI)

Abstract — Voice control remains conspicuously absent from image editing despite decades of multimodal interaction research and the maturity of modern AI. We present ISI Voice Image, a desktop application that enables voice-controlled clipboard image transformation: users copy an image, speak a command (e.g., "remove the background"), and paste the transformed result. Built in 48 hours using Deepgram for speech recognition and Google Gemini 3 Pro Image for image transformation, the system demonstrates that voice-controlled image editing is technically feasible with current cloud AI infrastructure. We position this work within the Research Through Design tradition, demonstrating that the specific combination of voice input, clipboard-based I/O, and desktop-wide availability constitutes an unexplored design space. This prototype validates the technical feasibility of voice-controlled image editing and opens new research directions in multimodal creative tools.

CCS Concepts: Human-centered computing → Natural language interfaces; Interaction paradigms

Keywords: voice interface, image editing, multimodal interaction, clipboard, speech-to-text

Introduction

Copy. Speak. Paste. This simple interaction pattern—copying an image to the clipboard, speaking a transformation command, and pasting the result—represents an unexplored intersection of mature technologies. Despite 45 years of multimodal interaction research since Bolt's "Put-That-There" [1] and the recent explosion of AI-powered image editing tools, voice control remains absent from mainstream image editing software.

We present ISI Voice Image, a desktop application that fills this gap. Users activate a global hotkey (Cmd+Shift+I), speak natural language commands such as "make it black and white" or "remove the background," and receive transformed images directly to their clipboard. The system demonstrates that voice-controlled image transformation is technically feasible with current AI infrastructure.

Our contribution is threefold: (1) we identify and validate a research gap at the intersection of voice interfaces, clipboard workflows, and desktop image editing; (2) we present a working prototype using commodity cloud APIs; and (3) we provide design insights from constructing a voice-controlled creative tool. We position this work within the Research Through Design tradition [18], where the prototype itself serves as the primary research contribution.

Related Work

Multimodal Interaction Foundations

The paradigm of combining voice with visual manipulation traces to Bolt's seminal "Put-That-There" system [1], which enabled users to manipulate graphics through speech augmented by gesture. Oviatt's "Ten Myths of Multimodal Interaction" [14] established that well-designed multimodal systems achieve mutual disambiguation, reducing errors compared to unimodal approaches. The QuickSet system [3] demonstrated 3–9× speed improvements when combining pen and voice for spatial tasks.

Wickens' Multiple Resource Theory [16] provides cognitive justification for voice-controlled image editing: voice commands engage the auditory-verbal channel while visual output occupies the visual-spatial channel, predicting minimal interference. Oviatt et al. [13] found that users spontaneously shift to multimodal interaction as cognitive load increases—from 59% during low-difficulty tasks to 75% at high difficulty.

Voice and Image Systems

The closest prior work is PixelTone [10], which combined speech and touch for mobile photo editing. However, PixelTone differs fundamen-

tally from our approach: it requires a dedicated app context (not system-wide availability), uses touch-based localization alongside voice, and targets mobile rather than desktop platforms.

Recent advances in instruction-following image models enable natural language editing without masks or fine-tuning. InstructPix2Pix [2] learns to edit images from text instructions using a diffusion model trained on synthetic instruction pairs. MGIE [5] uses multimodal LLMs to derive "expressive instructions" from brief commands—directly applicable to expanding terse voice commands into detailed editing guidance.

Despite these advances, no prior work combines voice input, clipboard-based I/O, and desktop-wide availability for image transformation. Adobe demonstrated a voice-controlled photo editing prototype in 2016–2017 that never shipped. Canva offers voice dictation for text prompts but not voice-controlled editing commands. This gap validates our research positioning.

System Design

Interaction Design

ISI Voice Image follows a command-based paradigm rather than conversational interaction. Users press Cmd+Shift+I to begin recording, speak their command, and press the hotkey again to stop (toggle mode). The clipboard serves as both input and output—users copy an image before invoking the system and paste the result afterward.

This "invisible interface" design offers several advantages: it requires no dedicated window or context switch, integrates into existing workflows, and works with any application that supports clipboard images. The push-to-talk activation prevents false triggers while avoiding the privacy concerns of always-on listening [12].

Technical Architecture

Figure 1: System pipeline showing data flow from clipboard through speech recognition to image transformation.

The system pipeline (Figure 1) consists of five stages:

1. **Clipboard capture**: On hotkey activation, the application reads the current clipboard image.
2. **Audio recording**: The system captures audio until the user presses the hotkey again (toggle mode).
3. **Speech-to-text**: Audio is sent to Deepgram [4] for transcription.
4. **Image transformation**: The transcribed command and clipboard image are sent to Google Gemini 3 Pro Image [6], which generates the transformed image.
5. **Clipboard output**: The transformed image replaces the clipboard contents, ready for pasting.

Implementation

We built ISI Voice Image using Tauri 2.0 [15] with a Rust backend and Vue.js 3 frontend. Tauri provides native clipboard access and global hotkey registration while maintaining a small footprint—our application bundle is under 20MB compared to 150MB+ for Electron alternatives.

The technology choices prioritize simplicity and rapid development. Deepgram provides reliable speech-to-text transcription. Gemini 3 Pro Image provides native image editing capabilities through natural language instructions, accepting an input image and text prompt to generate a transformed output.

Demonstration

The core workflow proceeds as follows: A user working in a presentation copies a photograph, presses Cmd+Shift+I, says "remove the background," presses the hotkey again, and pastes the result—all without leaving their current application.

Example transformations we have tested include: background removal ("remove the background"), style transfer ("make this look like a watercolor painting"), color adjustment ("increase the brightness" or "make it black and white"), object removal ("remove the water-

mark"), and compositing ("add a sunset sky behind the mountains").

In our usage, end-to-end latency—from finishing the voice command to receiving the transformed image—typically ranges from 10 to 20 seconds. The image transformation step dominates this latency, as the multimodal model must process both the input image and the natural language instruction to generate a new image. This latency is comparable to other AI image generation tools such as DALL-E or Midjourney, though notably longer than traditional non-AI image filters.

Discussion

Latency Considerations

Nielsen's foundational response time thresholds [11, 12] establish that delays under 1 second preserve user flow, under 10 seconds maintain attention, and beyond 10 seconds users may disengage. ISI Voice Image's typical 10–20 second latency exceeds the attention threshold, placing it in territory where users benefit from clear progress feedback.

However, this latency should be contextualized against comparable AI image tools. DALL-E and Midjourney typically require similar or longer waiting times for image generation. Users of AI creative tools have developed expectations that generative operations take time—unlike traditional filters that apply instantaneously.

Research on progress indicators [8] demonstrates that visual feedback during waiting reduces perceived duration. Our implementation addresses this through a dual feedback mechanism: (1) native OS notifications at each workflow stage ("Recording," "Processing," "Transforming," "Done!"), and (2) a system tray icon that changes between idle, recording, and processing states. The tray tooltip also updates to reflect current status. This provides persistent visual state while allowing users to continue other work during the 10–20 second transformation. Future work should explore more sophisticated progress indication such as streaming previews or percentage-based progress bars.

Limitations

As a 48-hour prototype developed within the Research Through Design tradition [9, 18], ISI Voice Image prioritizes demonstrating feasibility over production readiness. We acknowledge several limitations:

No formal user study: Following Greenberg and Buxton's [7] guidance on premature evaluation, we defer comprehensive user studies to future work. This prototype establishes technical feasibility; user acceptance studies would require the stability of a production system.

No systematic latency measurement: The 10–20 second latency range reflects our informal usage observations, not controlled benchmarking. Actual latency varies with network conditions, image size, and API server load.

Clipboard mental model: Users unfamiliar with clipboard-centric workflows may find the interaction non-intuitive. Power users who regularly use clipboard managers will adapt quickly; novice users require onboarding.

Accessibility: While voice control can benefit users with motor impairments, speech recognition accuracy degrades significantly for users with dysarthria or strong accents. Alternative input methods (text commands, GUI) should accompany voice in production systems.

API dependency: The system relies on third-party cloud services, introducing latency variability and potential availability concerns.

Contribution Framing

Following Wobbrock and Kientz's [17] framework for artifact contributions, we position ISI Voice Image as a prototype that "reveals new possibilities" in multimodal creative tools. The specific combination of voice input, clipboard-based I/O, and desktop-wide availability has not been previously explored, validating a novel design space worthy of further investigation.

The 48-hour development timeline, rather than being a limitation, demonstrates the maturity of current AI infrastructure. That a working voice-to-image transformation system can be built in a weekend using commodity APIs suggests this interaction paradigm is ready for broader exploration.

Conclusion

ISI Voice Image demonstrates that voice-controlled clipboard image transformation is tech-

nically feasible with current AI infrastructure. Our prototype validates a novel intersection of voice interfaces, clipboard workflows, and desktop-wide availability that prior work has not explored.

Current latency (typically 10–20 seconds) places the system in the realm of AI creative tools rather than instant filters, requiring appropriate user expectations and feedback mechanisms. Future work should pursue formal user studies, systematic latency optimization, accessibility improvements, and expanded command vocabularies for professional editing workflows.

The prototype is available as open source at https://github.com/frankdierolf/isi-research.

References

[1] Richard A. Bolt. 1980. "Put-That-There": Voice and Gesture at the Graphics Interface. In Proceedings of the 7th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '80), 1980. ACM, 262–270. https://doi.org/10.1145/800250.807503

[2] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. 2023. InstructPix2Pix: Learning to Follow Image Editing Instructions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023. 18392–18402.

[3] Philip R. Cohen, Michael Johnston, David McGee, Sharon Oviatt, Jay Pittman, Ira Smith, Liang Chen, and Josh Clow. 1997. QuickSet: Multimodal Interaction for Distributed Applications. In Proceedings of the Fifth ACM International Conference on Multimedia, 1997. ACM, 31–40. https://doi.org/10.1145/266180.266328

[4] Deepgram. 2023. Introducing Nova-2: The Fastest, Most Accurate Speech-to-Text API.

[5] Tsu-Jui Fu, Wenze Hu, Xianzhi Du, William Yang Wang, Yinfei Yang, and Zhe Gan. 2024. Guiding Instruction-based Image Editing via Multimodal Large Language Models. In Proceedings of the International Conference on Learning Representations (ICLR), 2024.

[6] Gemini Team, Google. 2023. Gemini: A Family of Highly Capable Multimodal Models. arXiv preprint arXiv:2312.11805 (2023).

[7] Saul Greenberg and Bill Buxton. 2008. Usability Evaluation Considered Harmful (Some of the Time). In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '08), 2008. ACM, 111–120. https://doi.org/10.1145/1357054.1357074

[8] Chris Harrison, Zhiquan Yeo, and Scott E. Hudson. 2010. Faster Progress Bars: Manipulating Perceived Duration with Visual Augmentations. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '10), 2010. ACM, 1545–1548. https://doi.org/10.1145/1753326.1753556

[9] Ilpo Koskinen, John Zimmerman, Thomas Binder, Johan Redström, and Stephan Wensveen. 2011. Design Research Through Practice: From the Lab, Field, and Showroom. Morgan Kaufmann.

[10] Gierad Paul Laput, Mira Dontcheva, Gregg Wilensky, Walter Chang, Aseem Agarwala, Jason Linder, and Eytan Adar. 2013. PixelTone: A Multimodal Interface for Image Editing. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13), 2013. ACM, 2185–2194. https://doi.org/10.1145/2470654.2481301

[11] Robert B. Miller. 1968. Response Time in Man-Computer Conversational Transactions. In Proceedings of the AFIPS Fall Joint Computer Conference, 1968. 267–277.

[12] Jakob Nielsen. 1993. Usability Engineering. Morgan Kaufmann Publishers.

[13] Sharon Oviatt, Rachel Coulston, and Rebecca Lunsford. 2004. When Do We Interact Multimodally? Cognitive Load and Multimodal Communication Patterns. In Proceedings of the 6th International Conference on Multimodal Interfaces (ICMI '04), 2004. ACM, 129–136. https://doi.org/10.1145/1027933.1027957

[14] Sharon Oviatt. 1999. Ten Myths of Multimodal Interaction. Communications of the ACM 42, 11 (1999), 74–81. https://doi.org/10.1145/319382.319398

[15] Tauri Contributors. 2024. Tauri 2.0: Build Smaller, Faster, and More Secure Desktop Applications.

[16] Christopher D. Wickens. 2002. Multiple Resources and Performance Prediction. Theoretical Issues in Ergonomics Science 3, 2 (2002), 159–177. https://doi.org/10.1080/14639220210123806

[17] Jacob O. Wobbrock and Julie A. Kientz. 2016. Research Contributions in Human-Computer Interaction. 2016. 38–44. https://doi.org/10.1145/2907069

[18] John Zimmerman, Jodi Forlizzi, and Shelley Evenson. 2007. Research Through Design as a Method for Interaction Design Research in HCI. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '07), 2007. ACM, 493–502. https://doi.org/10.1145/1240624.1240704