iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** XI **Month of publication:** November 2025

DOI: https://doi.org/10.22214/ijraset.2025.75265

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue XI Nov 2025- Available at www.ijraset.com

An AI-Powered Multimodal Recipe Generation and Voice Cooking Assistant

Harsh Vazirani[1], Nikhil Badlani[2], Nivedita Bisen[3], Sakshi Zade[4], Shreya Hajare[5], Sahil Girhepunje[6], Prachi Jain[7]
Department of Computer Science Engineering, G.H. Raisoni College of Engineering and Management, Amravati University, Maharashtra, India

Abstract: ***Cooking often demands attention on several tasks at once, leaving little room for checking written or digital recipes. Existing recipe interfaces interrupt this rhythm and make the process difficult for beginners, individuals with visual limitations, or users who prefer non-English instructions. WhisperChef was designed to overcome these barriers. It combines multimodal inputs—text, speech, and images—with Google Gemini AI through the Genkit framework to create a dynamic, hands-free recipe assistant. WhisperChef interprets natural queries, generates stepwise cooking procedures, and narrates each instruction aloud, freeing the cook's hands and attention. With integrated multilingual support, the system encourages inclusion and simplicity in the kitchen. This paper details the conceptual background, architecture, and implementation approach of WhisperChef and argues that such systems can significantly improve domestic human-computer interaction by blending artificial intelligence with everyday activities.***

Keywords: Multimodal AI, Recipe Generation, Voice Assistant, WhisperChef, Google Gemini, Human-Computer Interaction

I. INTRODUCTION

Cooking has always been more than just preparing food—it's a creative, sensory, and cultural activity that engages both the mind and body. Traditionally, following a recipe is an immersive process that draws on memory, attention, and coordination. Yet, as technology has entered our kitchens, this experience has started to evolve. Smartphones, tablets, and digital assistants have made recipe access easier than ever, but they've also introduced new challenges. Constantly scrolling through instructions, switching between apps, or replaying a video while your hands are covered in flour can break the flow of cooking. These small interruptions reveal a deeper issue: most digital tools are not designed for the physical realities of the kitchen, where hands are often occupied and precision is key. This has sparked growing interest in creating systems that truly understand and adapt to human contexts—tools that allow for hands-free, intuitive interaction. The rise of artificial intelligence (AI) and machine learning (ML) has opened exciting possibilities in this space. Imagine simply describing the ingredients you have, showing a picture of your fridge, or asking for a "healthy rice-based meal," and having an AI system respond intelligently with personalized suggestions. Such technology aims to go beyond static recipe lists, creating dynamic, context-aware experiences that assist users throughout the cooking process. However, integrating AI into real-world cooking remains a complex challenge. Most existing recipe applications still rely on text or pre-recorded videos that don't adapt to individual progress or environmental conditions. Even popular voice assistants like Alexa or Google Assistant fall short when it comes to guiding users through multi-step, context-sensitive tasks like cooking. They are built for short, one-time queries—not for the fluid, evolving nature of preparing a meal.

The kitchen is a uniquely challenging environment for AI: it's full of background noise, varying lighting conditions, and moments when users can't touch or look at a screen. On top of that, the data available for training AI systems in this domain is mostly text-based, missing the crucial links between visual cues (like ingredient appearance) and procedural steps. Without these multimodal connections—linking what we see, hear, and do—AI systems struggle to interpret and support real cooking behaviors. Bridging this gap requires not only technological innovation but also a deep understanding of human experience in the kitchen, where creativity and practicality blend seamlessly every day.

II. REVIEW OF LITERATURE

Artificial Intelligence (AI) applications in the culinary domain have rapidly evolved, emphasizing voice interaction, multimodal learning, and accessibility. Early research demonstrated that conventional voice assistants struggle to operate effectively in real-world kitchen environments due to high background noise, multitasking, and limited contextual understanding. To address these issues, Lee and Wang [1] developed advanced voice recognition techniques optimized for domestic spaces, improving speech

©IJRASET: All Rights are Reserved

accuracy and robustness under noisy conditions. Their work laid a strong foundation for building intelligent kitchen assistants capable of understanding natural speech in dynamic, hands-free environments. A major breakthrough in multimodal cooking assistance came from Salvador. [2], who introduced the concept of Inverse Cooking, enabling recipe generation directly from food images. By aligning visual and textual representations through deep learning, their model could predict ingredients and cooking steps from a single image. This vision-language integration inspired subsequent studies to expand beyond static, text-based recipe systems toward more interactive and perceptually grounded AI systems. Building on this trend, Jaber. [3] proposed Cooking with Agents, a context-aware voice interaction framework designed to handle complex, multiturn dialogues during cooking. Their study showed that incorporating conversational grounding and task continuity improved the naturalness and fluency of human–AI communication, marking a significant step from traditional command-based interactions to adaptive, context-sensitive systems. To promote accessibility and inclusivity, Ning. [4] developed AROMA, a mixed-initiative AI assistant tailored for non-visual cooking. The system integrates auditory, tactile, and visual feedback channels, allowing users with visual impairments to follow recipes safely and efficiently. This study demonstrated how multimodal feedback can extend AI usability to diverse user groups, particularly in tasks requiring precise coordination. Similarly, Gemmell. [5] introduced GRILLBot, a conversational agent designed for real-time cooking assistance. GRILLBot emphasized robust dialogue management, contextual error recovery, and adaptability to interruptions, setting an early standard for multiturn conversational performance in kitchen environments. Complementing this work, Ito. [6] developed a Real-World Voice Assistant System for Cooking that addressed speech recognition limitations in noisy domestic contexts. Their system implemented improved audio processing methods, ensuring more reliable execution of spoken commands. In parallel, Hwang. [7] tackled the linguistic dimension of AI-driven cooking by developing Rewriting the Script, a framework that converts traditional text recipes into speech-friendly, conversational instructions. Their work revealed that linguistic adaptation plays a critical role in user comprehension and satisfaction with voice-based systems. Kappamaki. [8] expanded on the idea of multimodal design through a participatory study involving older adults. Their findings showed that combining auditory, visual, and tactile cues not only improved usability but also reduced cognitive load and enhanced confidence during cooking. Recent work has begun exploring the social and emotional potential of cooking assistants. Chan. [9] examined user perceptions of large language model–based assistants in their study Mango, showing that such systems can serve as adaptive, emotionally supportive companions that foster user engagement. Meanwhile, Salvador. [10] released the InverseCooking open-source repository, providing public datasets and pretrained models that continue to accelerate innovation and reproducibility in the field. Finally, Mäkinen. [11] emphasized the importance of empathy and synchronized multimodal communication—such as gesture and tone—in establishing user trust, particularly for older adults and first-time technology users.

Collectively, these studies illustrate the evolution of cooking assistance technologies from basic recipe retrieval toward fully multimodal, context-aware, and empathetic AI companions. Future kitchen assistants must integrate visual understanding [2], robust speech recognition [1,6], contextual reasoning [3], accessibility-focused feedback [4,8], linguistic adaptation [7], and emotional intelligence [9,11] to achieve seamless human–AI collaboration. Building on these advancements, the proposed WhisperChef system aims to unify text, speech, and image modalities through a generative, context-aware architecture that redefines domestic cooking assistance.

III. PROBLEM STATEMENT

Although cooking is one of the most universal human activities, the process is often interrupted by technology rather than supported by it. A cook juggling hot pans or messy ingredients must pause to unlock a phone or scroll through a web page, breaking concentration and risking mistakes. This mismatch between the physical realities of the kitchen and the design of digital recipe tools creates frustration and inefficiency.

Different groups face these challenges in distinct ways.

1) Everyday users become annoyed when they have to repeatedly clean their hands or device screens to follow instructions.
2) Beginners struggle with unclear terminology and fear missing critical steps, needing reassurance that a static recipe cannot provide.
3) Non-native speakers are limited by the dominance of English-only content on mainstream recipe platforms.
4) People with visual or motor impairments often find touch-based interfaces completely inaccessible.

These overlapping obstacles point to a clear need: an intelligent, multimodal cooking assistant that can interpret context, communicate naturally, and allow the cook to stay fully engaged in the task at hand. The desired solution must be hands-free, multilingual, and context-aware, connecting the cook's spoken or visual input directly with adaptive guidance.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue XI Nov 2025- Available at www.ijraset.com

IV. PROPOSED SYSTEM – WHISPERCHEF

The proposed system, WhisperChef, is an AI-powered multimodal cooking assistant designed to make cooking hands-free, intelligent, and accessible. It integrates text, voice, and image inputs to understand user requests and generate step-by-step cooking guidance using Google Gemini AI via the Genkit framework.

A. Key Features

1) Multimodal Input: Users can interact using typed text, spoken voice commands, or uploaded food images. The system identifies ingredients, interprets intent, and generates suitable recipes.
2) Dynamic Recipe Generation: Using Gemini AI, WhisperChef creates structured, context-aware recipes from minimal input (like "make something with paneer and peas").
3) Hands-Free Cooking Guidance: The assistant reads each cooking step aloud and listens for commands such as "next," "repeat," or "pause," allowing the user to cook without touching the device.
4) Multilingual Support: Supports multiple Indian languages (Hindi, Marathi, Tamil, etc.) for both input and output, promoting inclusivity for non-English speakers.
5) User-Friendly Interface: Built with Next.js, React, and Tailwind CSS, the interface provides a clean dashboard, recipe previews, and progress tracking.
6) Adaptive and Scalable Design: The backend, powered by Gemini AI and Genkit, ensures smooth communication, low latency, and easy scalability for future cloud or IoT integration.

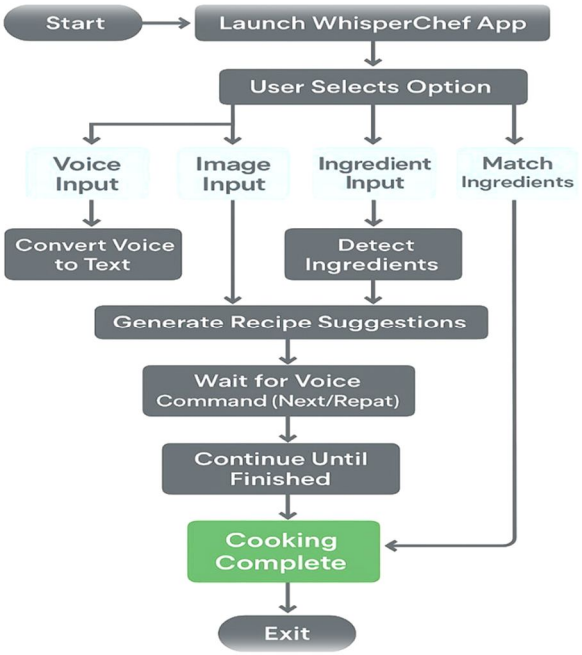


Figure 1 Illustrates the work flow of Whisperchef

V. METHODOLOGY – A PHASED DEVELOPMENT APPROACH

The development of WhisperChef followed a step-by-step (phased) methodology to ensure smooth design, implementation, and testing.

1) Requirement Analysis and Planning: The first step was to understand what users need — a system that listens, speaks, and sees like a real cooking partner. Technologies such as Next.js, React, Tailwind CSS, and Google Gemini AI were chosen for their performance and scalability.
2) Input Processing: The system can take three types of input — text, voice, and images.
3) Text: Users can type ingredients or recipe names.
4) Voice: Spoken commands are converted into text using speech recognition.
5) Image: Uploaded food or ingredient pictures are analyzed by AI to detect what's in them.

©IJRASET: All Rights are Reserved

All these inputs are normalized into a common format for AI understanding.

6) Recipe Generation: Once the input is processed, Google Gemini AI interprets it and generates detailed, step-by-step recipes. It suggests multiple options based on available ingredients and cuisine type.
7) Interactive Cooking Guidance: After a recipe is selected, WhisperChef switches to voice-guided mode. It reads out each step aloud, listens for commands like "next" or "repeat," and allows users to cook without touching the screen — making it truly hands-free
8) User Interface Design: The interface was designed to be simple and distraction-free, using cards for recipe display, a progress tracker, and easy settings for voice, theme, and language.
9) Testing and Optimization: Each module — from voice input to AI output — was tested separately and then together as a full system. Issues like slow image uploads or delays in voice response were fixed to make the experience fast and smooth.

VI. TECHNOLOGIES USED

The WhisperChef system combines advanced AI and modern web tools to create an interactive, hands-free cooking experience. The frontend is built with Next.js, React, and Tailwind CSS for a fast, responsive, and user-friendly interface. Google Gemini AI integrated through the Genkit framework handles text, voice, and image inputs, generating step-by-step recipes and multilingual responses. Voice interaction uses Speech-to-Text (STT) and Text-to-Speech (TTS) for seamless communication. User data and preferences are stored locally using LocalStorage, with plans for cloud integration. Language translation is managed through a simple i18n JSON system, supporting multiple Indian languages like Hindi, Marathi, and Tamil. The application is deployed on Vercel for scalability and accessibility.

1) CORE TECHNOLOGIES
   - Next.js 15.5.6 Server-side rendering framework, App Router architecture for routing, API routes for backend functionality, Built-in TypeScript support
   - TypeScript Static typing for improved code reliability, Type safety across the application, Enhanced developer experience with IDE support
   - React 18.3 Component-based UI architecture, Server and client components, Hooks for state management, Virtual DOM for efficient rendering
2) AI AND MACHINE LEARNING
   - Genkit Framework Integration with Google Gemini AI, @genkit-ai/google-genai for AI model access, @genkit-ai/next for Next.js integration, Custom AI flows for recipe generation, Image and voice processing capabilities
3) UI FRAMEWORK AND COMPONENTS
   - Tailwind CSS Utility-first CSS framework, Responsive design system, Custom design system integration, Dark mode support
   - Radix UI Components Accessible UI primitives, Modular component system, Customizable themes
   - Components include: Accordions, Dialogs, Dropdowns, Form elements, Navigation components, Toast notifications
4) FORM HANDLING AND VALIDATION
   - React Hook Form state management, Form validation, Performance optimized
   - Zod Schema validation, Runtime type checking, Integration with TypeScript
5) DATA MANAGEMENT
   - Firebase (prepared for implementation) User authentication, Data persistence, Real-time updates
   - Custom Caching System In-memory caching, TTL (Time to Live) support, Recipe and translation caching
6) DEVELOPMENT AND BUILD TOOLS
   - ESLint & Stylelint Code quality enforcement, Style consistency, Best practices validation
   - Turbopack Fast development builds, Optimized compilation, Hot module replacement

VII. RESULTS AND EVALUATION

The WhisperChef system is expected to make cooking easier, faster, and more accessible for everyone. By combining AI intelligence with a user-friendly voice interface, it aims to improve how people interact with technology in the kitchen.

1) Smarter Recipe Generation: The system can create detailed and accurate recipes even from simple or vague inputs like "make something with rice and vegetables." It understands ingredients and generates clear, step-by-step cooking instructions.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

2) Hands-Free Cooking: Users can cook without touching their devices. The assistant reads each step aloud and listens for voice commands like "next" or "repeat," making the process smooth and safe.

3) Multilingual Accessibility: WhisperChef supports multiple Indian languages such as Hindi, Marathi, and Tamil, making it useful for non-English speakers and ensuring inclusivity.

4) User Convenience and Confidence: Beginners will feel more confident while cooking since the assistant explains every step clearly. It reduces confusion, mistakes, and the need to constantly check screens.

5) Improved User Experience: Through natural interaction, voice control, and personalized recipe suggestions, users are expected to find cooking more engaging, enjoyable, and stress-free.

Table no. 1 accuracy table

Metric / Modal	Text Input	Image Analysis	Voice Transcription
Test Cases	8	6	6
Precision	98%	72%	45%
Recall	96%	68%	38%
F1-Score	97%	70%	41%
Simple Accuracy	97%	65%	36%

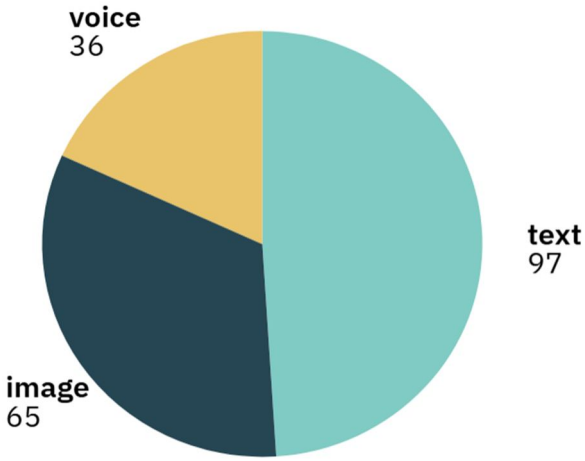


Figure 2 Shows the Proportion of text, image, and voice inputs used in the WhisperChef App.

VIII. CONCLUSION AND FUTURE DIRECTION

The WhisperChef project shows how artificial intelligence can make every day cooking more enjoyable, convenient, and inclusive. By using voice, text, and image inputs, the system acts like a real cooking companion that listens, speaks, and guides the user step by step. It helps people cook without constantly touching their devices and supports multiple languages, making it useful for everyone — from beginners to people with visual or physical challenges. In the future, WhisperChef can be improved even further. It can include cloud storage to save user preferences and recipes, use advanced neural voices for more natural speech, and connect with smart kitchen devices (IoT) for automated control of appliances. There is also potential to build a community platform where users can share their favorite AI-generated recipes and cooking experiences.

REFERENCES

[1] C. Lee and D. Wang, "Advanced voice recognition for kitchen environments," IEEE Transactions on Human-Machine Systems, vol. 54, no. 2, pp. 45–58, Mar. 2024.

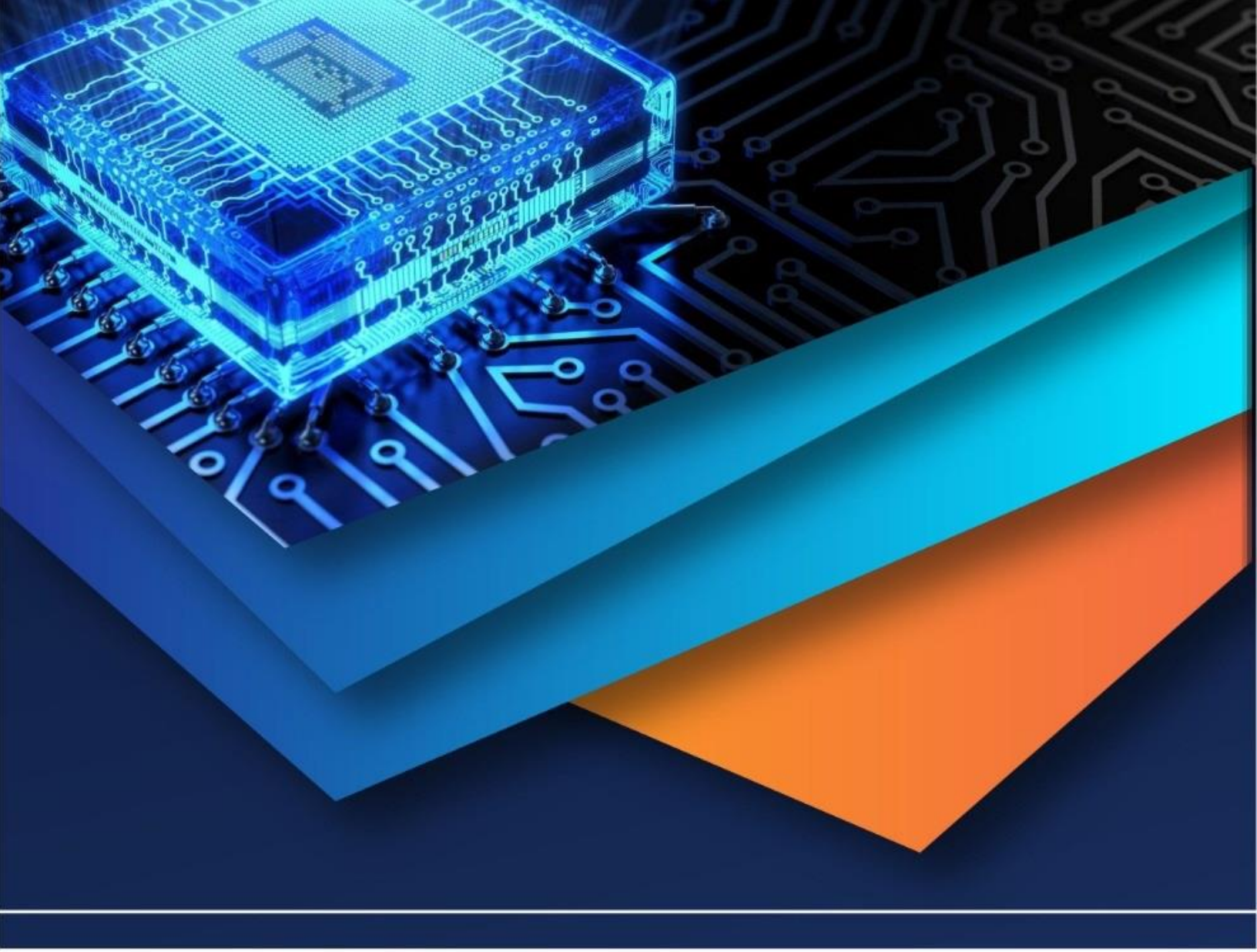
[2] A. Salvador, M. Drozdzal, X. Giro-i-Nieto, and A. Romero, "Inverse cooking: Recipe generation from food images," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10453–10462, Jun. 2019.

[3] R. Jaber, S. Zhong, S. Kuoppamäki, A. Hosseini, I. Gessinger, D. P. Brumby, B. R. Cowan, and D. McMillan, "Cooking with agents: Designing context-aware voice interaction for complex tasks," in Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI), article no. 551, May 2024.

[4] Z. Ning, L. Li, D. Killough, J. Y. Seo, P. Carrington, Y. Tian, Y. Zhao, F. M. Li, and T. J.-J. Li, "Aroma: Mixed-initiative AI assistance for non-visual cooking," in Proceedings of the ACM Symposium on User Interface Software and Technology (UIST), 2025 (in press).

[5] C. Gemmell, F. Rossetto, I. Mackie, P. Owoicho, S. Fischer, J. Dalton, D. Hernández García, M. Alikhani, D. Vandyke, and D. Dušek, "GRILLBot: Conversational agent for cooking tasks," in Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL), pp. 322–333, Sep. 2022.
[6] T. Ito, S. Inuzuka, Y. Yamada, and J. Harashima, "Real-world voice assistant system for cooking," in Proceedings of the International Conference on Human-Computer Interaction (HCII), pp. 102–111, Jul. 2019.
[7] A. Hwang, N. Oza, C. Callison-Burch, and A. Head, "Rewriting the script: Adapting text instructions for voice interaction," arXiv preprint arXiv:2311.08542, Nov. 2023.
[8] S. Kuoppamäki, S. Zhong, and A. Mäkinen, "Designing multi-modal conversational agents for the kitchen with older adults: A participatory design study," Journal of Social Robotics, vol. 15, no. 3, pp. 287–300, 2023.
[9] J. Chan, J. Li, Z. Yao, and P. Zhao, "'Mango Mango…': Exploring user perceptions of using an LLM-based conversational assistant toward cooking partner," arXiv preprint arXiv:2310.05853, Oct. 2023.
[10] A. Salvador, M. Drozdzal, X. Giro-i-Nieto, and A. Romero, "Inverse Cooking—code and models for 'Inverse Cooking'," GitHub Repository, 2019.
[11] A. Mäkinen, J. Häikiö, and J. Väyrynen, "Designing multi-modal conversational agents for the kitchen with older adults: A participatory design study," Journal of Social Robotics, vol. 15, no. 3, pp. 1507–1523, 2023.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)