Towards a proactive cooking companion for the elderly

Katarina Esteve[1], Morgan Fredriksson[2], Joakim Gustafson[3], Dimosthenis Kontogiorgos[3] and Timo Mashiyi-Veikkola[1]

[1]Electrolux, Sweden
[2]Nagoon, Sweden
[3]KTH Royal Institute of Technology, Sweden

Corresponding author: jocke@speech.kth.se

Abstract

We present a voice assistant designed as a cooking companion, addressing both nutritional and social needs through intelligent interaction. Through WoZ experiments, we validated: social dialogue serves functional purposes, where "*chatty*" assistants transform cooking pauses into engaging interactions while instructional-only versions create frustrating dead air, despite identical timing.

1 Introduction

Global demographic transitions challenge welfare sustainability, prompting countries to adopt "*aging-in-place*" policies supporting elderly independent living. Older adults living alone face deteriorating quality of life through declining nutrition and escalating social isolation. Cooking is a critical intervention point for promoting active aging. We present a voice-based cooking assistant designed as a companion rather than tool, empowering user agency through dual objectives. First, it motivates older adults in preparing healthy meals through clear, paced, context-aware instructions that lower cognitive barriers. It also serves as a conversational partner mitigating loneliness through meaningful dialogue beyond task commands, fostering presence and shared experience. Preliminary findings suggest senior home cooks benefit from contextualized guidance and perceive socially adaptive dialogue during cooking pauses as key to maintaining engagement.

2 Previous research

Spoken dialogue systems in the culinary domain leverage hands-free, eyes-free interfaces. Early work explored AI-generated healthier recipe alternatives aligned with user preferences (Pecune et al., 2020), while advanced systems incorporated multimodality (Hannon et al., 2024). Weber et al. (2023) moved beyond recipe selection, proposing a framework classifying cooks by competence and autonomy needs, enabling nuanced, adaptive interaction styles for intelligent kitchen partners. User-centric design is particularly crucial for older adults. Kuoppamäki et al. (2023) identified benefits (e.g. cognitive support and nutrition advice) and age-specific challenges through participatory workshops. Notably, older adults perceived agents as task collaborators but not conversational partners, highlighting needs for socially adaptive designs. Jaber et al. (2024) demonstrated that commercial voice assistants often broke down because they lacked contextual awareness. Their Wizard of Oz (WoZ) study showed that context-aware assistants with shared understanding of task states enable far more effective collaboration. Early cooking assistants like *Cooking Coach* (Laroche et al., 2013) established the hands-free value proposition allowing for recipe search, ingredient verification, and step navigation. However, these systems relied on predefined dialogue logic and transactional styles, struggling with natural conversation nuances and unexpected user behavior (Chan et al., 2025). Human Activity Recognition leverages multi-sensor fusion to classify kitchen actions like "*chopping*" (Aguileta et al., 2019). However, recipe guidance requires higher-level inference about step completion shifting from discrete action classification to goal-oriented state inference based on user actions and intents. While LLMs have advanced in reasoning and language generation, their application to situated interaction has limitations, as they lack situational awareness and are unable to perceive non-verbal cues, interpret physical actions, or understand environmental states with the latency required for conversation. There has also been some attempts at addressing this by pairing powerful LLMs for high-level dialogue with real-time human-in-the-loop operators for detection of user state and actions (Marcinek et al 2024).

3 Pilot studies

Two pilots informed the design of the main WoZ study. The first assessed state-of-the-art LLMs' ability to detect cooking step completion. 45 cooking scenes were recorded from five camera angles, and compared the ability of LLaVA (Liu et al., 2023) and Gemini 2.5 Flash to detect when a cooking step had been completed. Gemini's reasoning-based architecture outperformed LLaVA's traditional vision approach, though both models generated excessive false positives that would frustrate users. Additionally, Gemini's cloud-based processing requires stable internet connectivity and raises privacy concerns, limiting practical deployment. The second pilot investigated personalizing instruction content and detail based on self-reported cooking experience and autonomy preferences, yielding four user profiles:

The Beginner (Low/Low):

- Requires detailed step-by-step instructions

The Creative Beginner (Low/High):

- Experiments but needs error correction

The Precise Chef (High/Low):

- Follows recipes, prefers concise instructions

The Creative Expert (High/High):

- Improvises, needs flexible support

Four subjects (one of each personality) cooked in their home while getting instruction from a WoZ-controlled voice assistant via a zoom call, where only the human operator had access to the video feed. Subjective evaluations showed high satisfaction scores, and post-experiment interviews revealed distinct preferences: *the beginner* appreciated clear steps that reduced uncertainty and taught new techniques (e.g., why to save pasta water); *the creative beginner* valued the balance between freedom and guidance (e.g., cooking times for added ingredients); *the precise cook* favored well-paced step-by-step instructions; and *the creative expert* viewed the AI as inspiring rather than essential. Based on the experiences from these pilots a WoZ study was designed to evaluate the benefits of a cooking assistant with human-level action detection, that apart from cooking instruction would provide encouraging comments and interesting food trivia during cooking. The human operator decided when to give the next cooking instruction and if the user needed further instruction to complete a task, also deciding when to provide social encouragements and trivia.

4 Wizard of Oz experiments

The project conducted WoZ experiments where users interacted with a Swedish voice-controlled cooking assistant while preparing omelets in a smart kitchen lab. Two versions of the system were evaluated: an *Instructional AI Chef* that only gave cooking instructions and a *Chatty AI Chef*, that provides encouraging comments and fun facts.

4.1 System description

The WoZ system was implemented as a JavaScript/HTML client-server architecture that streams rendered audio prompts from the operator's interface to the participant's client via WebSockets. Instructional content was structured in a JSON configuration file containing cooking steps, where each step had cooking instructions, additional instructions, and trivia prompts. At the bottom of the interface were prompts for encouragements "*You are doing fine!*", meta utterances like "*yes I can hear you!*" and deflections like "*That's an interesting question, but let's go back to cooking!*". All prompts were pre-rendered using ElevenLabs' Swedish TTS engine. To mitigate TTS platform rendering inconsistencies, the system implemented dynamic re-rendering with caching mechanisms to minimize latency. The operator interface offered two modes: edit mode for modifying, adding, and repositioning prompts, and runtime mode featuring an ad-hoc text input for generating unforeseen instructions in real-time. All experimental sessions were recorded from multiple angles using video cameras and microphones. Side-view and top-view videos were combined via OBS (Open Broadcaster Software) and streamed to an 80-inch screen, enabling monitoring participants' progress in real-time while following system instructions (Fig. 1).

4.2 User study

The user study took place in the KTH Intelligence Augmentation lab, a smart home lab equipped with synchronized sensors for multimodal activity detection and behavior recording. The facility includes a fully functioning kitchen and control room with GPU servers and real-time video feeds for human wizards in human-in-the-loop recordings. Six Swedish senior participants (ages 63-66, two men and four women) regarded themselves as average to skilled chefs, see Table A1. Five lived alone; two had occasionally used voice assistants like Siri or Alexa. Participants were

initially interviewed about their cooking habits and introduced to the experiment, then interacted with the kitchen voice assistant to cook a mushroom omelet. Post-interaction, they discussed the experience, compared it to following written recipes, and suggested improvements.

5 Results

All interactions were transcribed with Whisper-KB and annotated with Qwen3-Omni [1] to capture moment-by-moment analysis of user actions, certainty levels, and emotional engagement. All interaction went smooth and all participants were able to follow the instructions and make similarly looking omelets. The users typically did not speak to the assistant, except for one user that spoke at length when experiencing frustrating breakdowns. This is in line with our previous study on a cooking assistant, where users produced "more complex utterances" and higher word counts in attempts to resolve failures compared to normal instruction flow (Kontogiorgos et al 2020). The wizard never used of the ability to create new responses or handle out-of-domain requests, see Table A3.

5.1 From Partnership to Problem-Solving

Most users experienced positive sessions characterized by high enjoyment and confidence, demonstrating the potential for truly collaborative cooking partnerships. Peak enjoyment moments consistently linked to three interaction types:

Social Connection: Warm greetings and rapport-building questions (e.g., "*Do you like omelets with mushroom filling*?") established a collaborative atmosphere, eliciting smiles and warm responses.

Positive Reinforcement: Simple compliments like "*That turned out well*" created clear moments of satisfaction and encouragement.

Successful Completion: Executing final recipe steps of folding and plating the omelet consistently generated high satisfaction.

5.2 When the Partnership Breaks Down

Some sessions revealed the fragility of user experience facing system/environmental failures:

Hardware and Environmental Mismatch: One user unfamiliar with the induction hob experienced significant frustration. He exclaimed "*This stove has gone crazy!*" when the child lock was activated, requiring researcher intervention.

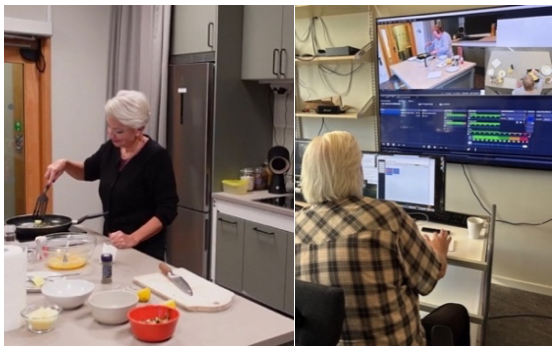

Figure 1: User in IA-lab (left) and Wizard (right)

Perceptual Failure and User Mistakes: When the same user mistakenly chopped oyster mushrooms together with the champignons led to problems as the next step was to grate the oyster mushrooms. Despite use repeatedly stating "*I don't see any oyster mushrooms*"), the assistant continued with trivia about the non-existent ingredient, shattering the illusion of a situation-aware partner.

5.3 Social Chatter vs. Adaptive Guidance

The Chatty AI-chef's non-instructional dialogue were of two types.

Social Chatter: Quick, interactive elements like greetings, personal questions, praise effectively built rapport. Longer fun facts and trivia received neutral responses during manual tasks but were appreciated in post-interaction interviews.

Adaptive Guidance: Context-aware instructions proved very valuable, for example by providing concrete guidance on how to set the temperature: "*Low means you set the stove to three.*" This targeted adaptation, directly addressing earlier uncertainty, was perceived as highly intelligent and helpful in post-experiment interviews.

6 Post-experiment User Experience

Post-experiment interviews revealed clear experiential divergence between *Instructional AI Chef* and *Chatty AI Chef participants*. Both groups comprised experienced home cooks comfortable adapting recipes, with varying voice assistant experience—yet their reactions differed markedly based on interaction style rather than individual cooking habits or tech-savviness.

[1] https://github.com/QwenLM/Qwen3-Omni

6.1 The Instructional Group

Post-experiment feedback was mixed. While one participant called it "*Wonderful... Better than a cookbook*" praising clear adaptations like specifying "*medium heat*" on the unfamiliar hob, she explicitly rejected more conversation: "*I wouldn't like him to chat more while I'm cooking.*" Two of three participants found the session "*long and slow*" perceiving the assistant as a simple timer unaware of their actions. Their frustration stemmed not from lacking social chatter but from poor responsiveness and dead air during cooking pauses.

6.2 The Chatty Group

Post-experiment reactions were uniformly positive. None experienced slowness; instead perceiving the assistant as smart and action-aware. One participant praised the non-instructional content: "*I thought that was really fun, while you were still waiting for something to be ready.*" The social chatter and trivia transformed temporal gaps that were experienced as slowness by the instructional group into continuous, engaging companionship.

6.3 Comparative Analysis

Chattiness served functional purposes beyond entertainment. The instructional group interpreted cooking pauses as slowness and system unawareness, while the chatty group experienced these same pauses as engaging, perceiving the assistant as more intelligent and responsive. The instructional group's issue wasn't lacking trivia but perceived poor adaptive pacing. The chatty version integrated waiting time into collaborative experience rather than speeding up. Successful kitchen assistants must exceed recipe reading: user experience depends on resilience, social grace, and demonstrated awareness through adaptation.

7 Conclusions

This study evaluated the usefulness of a proactive voice-based cooking assistant for older users. Our initial pilot investigated whether todays LLMs can detect cooking step completion from video feeds. Finding the technology insufficiently mature for real-time deployment, we employed a WoZ approach where a human operator monitored video feeds to determine optimal instruction timing, based both on the user's cooking actions and additional instructions based on their displayed level of uncertainty. The main WoZ study targeted older users with moderate to high cooking skills, comparing two assistant versions: instruction-only vs instructions supplemented with encouragement and food-related trivia. Both groups found the system useful. Notably, the conversational elements transformed the assistant's perceived responsiveness and intelligence. Participants receiving encouragement and trivia during cooking activities experienced the system as more aware and engaged, suggesting that social dialogue serves functional purposes beyond entertainment by filling natural cooking pauses with meaningful interaction. When analyzing the duration of each cooking step across participants (see Table A2), we observed that some steps show very consistent timing, such as cutting a lemon, while others vary considerably, such as peeling and cutting garlic. The timing variations for different kinds of cooking actions will be used in the design of the first fully automated version of our AI chef.

8 Future Work

Future work will enhance perceptual capabilities through multimodal human action detection, leveraging visual real-time object detection via YOLO world (Cheng et al., 2024) and 3D mesh recovery through SMPL-X for classifying cooking motions (Pavlakos et al., 2019). The kitchen's soundscape provides distinct auditory signatures, where models like BEATS (Chen et al., 2022) can identify cooking sounds that could contribute to human action detection. Combining visual "chopping" motions with knife-on-board sounds would create robust recognition that surpasses isolated modalities. We plan to use our conversational speech synthesis with controllable prosody (Lameris et al., 2023), enabling the assistant to adapt its speaking style to different dialogue situations: delivering task instructions in a clear, read-speech style; engaging in social side-conversations with spontaneous speech; marking urgency through prosodic emphasis when interruptions require immediate user response; and increasing vocal effort when loud appliances are operating (Marcinek et al., 2025).

Acknowledgements

This work was funded through the Vinnova-funded project FoodTalk. We would like to thank Axel Sundelin, Allan Inma, Amanda Herbe and Elise Cars for their invaluable work in the two pilots, that served as input to the design of the main study.

References

Aguileta, A.A., Brena, R.F., Mayora, O., Molino-Minero-Re, E. and Trejo, L.A., 2019. Multi-sensor fusion for activity recognition—A survey. Sensors, 19(17), p.3808.

Chan, S., Li, J., Yao, B., Mahmood, A., Huang, C.M., Jimison, H., Mynatt, E.D. and Wang, D., 2025. " Mango Mango, How to Let The Lettuce Dry Without A Spinner?": Exploring User Perceptions of Using An LLM-Based Conversational Assistant Toward Cooking Partner. Proceedings of the ACM on Human-Computer Interaction, 9(7), pp.1-35.

Chen, S., Wu, Y., Wang, C., Liu, S., Tompkins, D., Chen, Z., & Wei, F. (2022). Beats: Audio pre-training with acoustic tokenizers. arXiv preprint arXiv:2212.09058.

Cheng, T., Song, L., Ge, Y., Liu, W., Wang, X. and Shan, Y., 2024. Yolo-world: Real-time open-vocabulary object detection. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 16901-16911).

Hannon, B., Kumar, Y., Li, J.J. and Morreale, P., 2024. Chef Dalle: transforming cooking with multi-model multimodal AI. Computers, 13(7), p.156.

Jaber, R., Zhong, S., Kuoppamäki, S., Hosseini, A., Gessinger, I., Brumby, D.P., Cowan, B.R. and Mcmillan, D., 2024. Cooking with agents: Designing context-aware voice interaction. In Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (pp. 1-13).

Kontogiorgos, D., Pereira, A., Sahindal, B., Van Waveren, S., & Gustafson, J. 2020. Behavioural responses to robot conversational failures. In Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction (pp. 53-62).

Kuoppamäki, S., Jaberibraheem, R., Hellstrand, M. and McMillan, D., 2023. Designing multi-modal conversational agents for the kitchen with older adults: a participatory design study. International Journal of Social Robotics, 15(9), pp.1507-1523.

Lameris, H., Gustafson, J. and Székely, É., 2023. Beyond style: synthesizing speech with pragmatic functions. In Proceedings of Interspeech 2023, Dublin, Ireland, Aug 20 2023-Aug 24 2023 (pp. 3382-3386).

Laroche, R., Dziekan, J., Roussarie, L. and Baczyk, P., 2013. Cooking coach spoken/multimodal dialogue systems. In Proceedings of the IJCAI Workshop on Cooking with Computers.

Liu, H., Li, C., Wu, Q. and Lee, Y.J., 2023. Visual instruction tuning. Advances in neural information processing systems, 36, pp.34892-34916.

Marcinek, L, Beskow, J. and Gustafson, J., 2024. A Dual-Control Dialogue Framework for Human-Robot Interaction Data Collection: Integrating Human Emotional and Contextual Awareness with Conversational AI. In International Conference on Social Robotics 2024.

Marcinek, L., Beskow, J. and Gustafsson, J., 2025. Towards Adaptable and Intelligible Speech Synthesis in Noisy Environments. In 26th Interspeech Conference 2025, Rotterdam, Netherlands, August 17-21, 2025 (pp. 2165-2169).

Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A., Tzionas, D. and Black, M.J., 2019. Expressive body capture: 3d hands, face, and body from a single image. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 10975-10985).

Pecune, F., Callebert, L. and Marsella, S., 2020, September. A Recommender System for Healthy and Personalized Recipes Recommendations. In HealthRecSys@ RecSys (pp. 15-20).

Weber, J., Esau-Held, M., Schiller, M., Thaden, E.M., Manstetten, D. and Stevens, G., 2023. Designing an interaction concept for assisted cooking in smart kitchens: focus on human agency, proactivity, and multimodality. In Proceedings of the 2023 ACM Designing Interactive Systems Conference (pp. 1128-1144).

Appendix

Table A1. Users conducting three cooking actions (anonymized using Firefly Image 3 in Photoshop)

	Slice the lemon	Crush the garlic	Fry the omelet
In1			
In2			
In3			
Ch1			
Ch2			
Ch3			

Table A2. Times to perform the main cooking steps

Cooking step	chatty1	chatty2	chatty3	instruct1	instruct2	instruct3
slice mushrooms	48	43	47	54	38	59
grate oyster mushrooms	42	26	17	14	30	20
slice the lemon	12	9	19	18	15	8
heat oil	102	130	44	29	45	60
fry the mushrooms	123	122	83	112	89	45
cut the garlic, add to pan	51	48	106	62	52	60
season the mushrooms	40	93	56	40	40	26
add butter and lemon	30	37	31	51	15	28
wipe the pan	14	39	36	19	26	23
crack eggs	24	30	26	22	24	25
whisk eggs	30	53	25	39	41	20
heat butter	30	47	27	61	29	47
fry the omelet	142	280*	163	137	146	180
place mushrooms	8	16	21	25	15	22
fold the omelet	8	8	8	11	7	11
put omelet on plate	10	17	18	12	7	12

* longer time due to hardware failure

Table A3. Number of utterance types from the system

	Core Instructions	Additional Instructions	Praise / Social	Trivia
chatty1	28	28	8	14
chatty2	28	33	3	18
chatty3	28	29	5	14
instruct1	28	22	2	-
instruct2	28	26	5	-
instruct3	28	26	3	-