Audio-Visual Recipe Guidance for Smart Kitchen Devices

Caroline Kendrick*
Technische Hochschule Ingolstadt
Research Institute AImotion Bavaria
Ingolstadt, Germany
cg_kendrick@outlook.com

Mariano Frohnmaier
Technische Hochschule Ingolstadt
Research Institute AImotion Bavaria
Ingolstadt, Germany
mariano.frohnmaier@thi.de

Munir Georges
Technische Hochschule Ingolstadt
Research Institute AImotion Bavaria
Ingolstadt, Germany
munir.georges@thi.de

Abstract

An important degree of accessibility, novelty, and ease of use is added to smart kitchen devices with the integration of multimodal interactions. We present the design and prototype implementation for one such interaction: guided cooking with a smart food processor, utilizing both voice and touch interface. The prototype's design is based on user research. A new speech corpus consisting of 2,793 user queries related to the guided cooking scenario was created. This annotated data set was used to train and test the neural-network-based natural language understanding (NLU) component. Our evaluation of this new in-domain NLU data set resulted in an intent detection accuracy of 97% with high reliability when tested. Our data and prototype (VoiceCookingAssistant, 2021) are open-sourced to enable further research in audio-visual interaction within the smart kitchen context.

1 Introduction

The importance of cooking in daily life makes the kitchen a natural focus for emerging technologies, as shown by Khot and Mueller in their analysis of human-food interaction (Khot and Mueller, 2019). Smart kitchen gadgets are part of this growing market (Research Private Ltd and Markets, 2020), including the all-in-one food processor, a countertop device which combines the expected blending utility with additional functionalities, such as weighing and cooking (Fries et al., 2018). This paper proposes a multimodal guided cooking experience for such a device, incorporating voice input and output, and touch interface.

The kitchen is a complex environment, and certain modes of interaction may be unavailable to a chef - for example, if their hands are oily, a touchscreen will be difficult to use. Offering multiple modes of interaction allows the user to interact with the device via the modality most natural to them in a context. Essential to natural interaction with any voice assistant is the Natural Language Understanding (NLU) component of the system, which greatly depends on both the quality of the model used and the amount of training data for the underlying domain. Although speech corpora for smart home and kitchen devices exist, to the best of our knowledge there is currently no public corpus with annotated voice commands for step-by-step guided cooking.

Our contributions are the design and implementation of a prototype for multimodal guided cooking with an all-in-one food processor. Moreover, we built an English-language NLU text corpus annotated with 39 intents and 19 entities for the guided cooking domain, and evaluated the quantity of training data. The prototype and data set are open source (VoiceCookingAssistant, 2021), and can be extended for further research in the area of smart cooking voice assistants.

2 Related Work

Existing guided recipe solutions require many sensors and 'smart' accessories to monitor a chef's progress. For example, the *Smart Kitchen* requires accessories such as radio frequency identification (RFID) tags to identify ingredients (Hashimoto et al., 2008), and *Shadow Cooking* requires a depth camera, projector, and digital scale (Sato et al., 2014). Both *KogniChef* (Neumann et al., 2017) and *Kochbot* (Alexandersson et al., 2015) make use of an entire smart kitchen system to monitor the user's actions and communicate recipe instructions, although the *Kochbot* may also be used alone as a mobile app providing recipe guidance via voice and screen (Schäfer et al., 2013). The research of Bouchard et. al (Bouchard et al., 2020) focuses

* See Acknowledgements on the very last page.

on the elderly and cognitively impaired, and their safety in the kitchen. They propose and prototype a smart range that monitors and guides cooking via a touch interface, but only monitors the stove, with limited recipes. These solutions and others successfully provide a multimodal approach to recipe guidance, and can also provide support for users with unique needs.

Blasco et al. (Blasco et al., 2014) discuss the importance of a voice interface in their smart kitchen proposal, which focuses on Assisted Living for the elderly and incorporates multiple kitchen and home appliances. They emphasize that voice commands are not only a natural interaction, but they also provide accessibility for users with visual or cognitive impairments. Although their research focuses on nutritional guidance, Angara et. al (Angara et al., 2017) outline another important benefit for conversational interaction in the kitchen: users prefer voice interaction when multi-tasking, and cooking is a complex and hands-on activity.

Our approach differs in that, for the majority of recipes, no other devices are needed to prepare or monitor a dish. Recipe guidance and cooking implements are contained in the same product, and the user must only interact with one device. It is also significant that, as recipe guidance is both visual and auditory, the chef does not *need* to look at the visual interface while cooking, and can focus on the meal being prepared.

3 Behind the Prototype and NLU Corpus

A detailed interaction flow (Section 3.1) supported by user interviews (Section 3.2) provided the basis for the prototype. The speech corpus (Section 3.3) was developed after the design & interaction flow were established. Section 3.4 briefly discusses the architecture and implementation of our prototype.

3.1 Interaction Flow for Guided Cooking

We began by evaluating guided cooking experiences, charting the potential interactions between the user and the system in an interaction flow. Examples of guided cooking experiences - such as instructional videos, product reviews, and recipe tutorials - were analyzed, from which we created an adaptive recipe framework, similar to that of Hashimoto et al. (Hashimoto et al., 2014). We found that recipe steps consist of the same or similar sub-steps, for which we created accompanying voice/touch interactions. A simplified example of a recipe step may be seen in Figure 1. The diagram is based on Bollen's understanding of flow charts (Bollen, 2010).

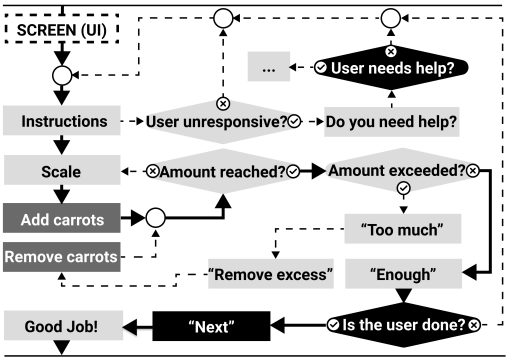

Figure 1: Excerpt from detailed interaction flowchart.

This flow provides the basis of the prototype logic, affordances for the visual interface, and provides a general framework to adapt recipes for use with the food processor.

3.2 Visual & Voice Design based on User Studies

Two studies were conducted with potential users, which formed the basis of the NLU corpus and provided requirements for the visual design. Both qualitative and quantitative methods were employed, and the studies focused on identifying user needs and potential device interactions. The *first study* was a usability test which simulated a guided cooking scenario. While a participant used an all-in-one food processor to cook a recipe, the researcher dictated the recipe steps. The *second study* was a semi-structured interview, without using the food processor, in which 12 participants imagined a specific guided cooking scenario and were asked how they would give commands. For example, how would the participant increase the number of people a recipe would feed if they were preparing for a dinner party later that day?

Using affinity diagramming (Lucero, 2015) to analyze the interviews provided design insights, which informed the user expectations for the visual interface and user experience of the device. Figure 2 displays a sample of the visual interface, showing the overview of a selected recipe.

3.3 Generating the NLU Data

In order for our prototype to understand the user's intentions, we had to create a sufficiently large basis of annotated user queries (Allen, 1988). The language of the queries should sound as *natural* as

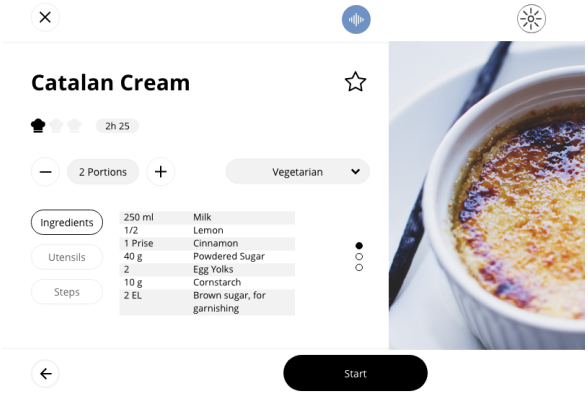

Figure 2: Visual Screen-Design: Recipe Overview.

possible. Therefore, we noted the specific language used by the participants from the second study in Section 3.2. The resulting 93 sample commands served as a starting point for several iterations to generate more commands and define the intents and entities.

Another condition of the definition of our domain-specific intents and entities was to define them in such a way that every interaction possible via the visual (or touch) interface can also be realized via speech. A detailed study of all of our prototype screens, like the one in Figure 2, and the interaction flow (see Figure 1) allowed us to fine-tune the definition of intents and entities. Following this strategy, we generated a set of 2803 text-based user queries with an in-domain vocabulary size of 436 and 15231 running words. Each sentence of this query base, denoted by $\mathcal{C}$, carries an intent label, but must not necessarily be labeled with an entity.

The resulting in-domain corpus $\mathcal{C}$ enables a user to navigate through step-by-step recipes, search for specific recipes, and add recipes to favourites. Also, it is possible to set device parameters like temperature, process duration, or blender speed. Overall there are 39 intents and 19 entities (VoiceCookingAssistant, 2021).

3.4 Architecture of the Prototype

As depicted in Figure 3, the prototype architecture is comprised of three components:

1. The **State Machine Frontend (SMF)** controls the user interface using high-fidelity graphics and a voice & touch interaction layer. It streams audio continuously to the
2. **Middleware Backend (MB)** which connects the SMF with the
3. **Logical Backend (LB)** using the MQTT protocol (Light, 2017). The LB processes the voice signal and provides the classified user intent in a JSON structure. (Pezoa et al., 2016).

Any component can be executed locally on the potential device without an internet connection as motivated by (Stemmer et al., 2017), or with partial internet access as proposed by (Georges et al., 2014). The heart of the LB uses *Rhasspy* (Hansen, 2021), an open-source collection of offline voice assistant services. It contains all subsystems necessary to processing a spoken query uttered by a user in a guided cooking scenario.

The prototype waits for a query that starts with a wake word. As soon as the wake word is recognized, the query is transcribed using the automatic speech recognition (ASR) system Kaldi (Povey et al., 2011). The recognized text is then forwarded to Rhasspy's intent recognition system in order to determine the user's intention. Here, the developer can choose the state-of-the-art NLU capabilities (Bocklisch et al., 2017).

4 Evaluation of the NLU Corpus

Rasa (Rasa Technologies, 2021) was used to evaluate our NLU dataset $\mathcal{C}$, as it can be selected in the Rhasspy pipeline. In addition, Rasa allowed us to easily use the **D**ual **I**ntent and **E**ntity **T**ransformer (DIET) architecture, a powerful state-of-the-art system for joint intent classification and entity recognition (Bunk et al., 2020).

4.1 First Analysis of the Speech Corpus

From the original NLU corpus $\mathcal{C}$ from Section 3.3, we carefully selected 839 sentences as our global test data set, denoted by $\mathcal{T}$, with 4507 running words. The remaining 1964 queries formed our training data set, denoted by $\mathcal{D} := \mathcal{C} \setminus \mathcal{T}$, with a total of 10724 running words. We repeatedly trained[1] the DIET model using the training data $\mathcal{D}$ and tested each model using the test set $\mathcal{T}$ ($n = 10$). The resulting averaged evaluation metrics (see Table 1) promise a good NLU performance. However, it is more interesting to know if we have collected *enough* user queries. The absolute numbers above do not allow us to make a statement about this.

[1]The configuration file which we used to specify Rasa's NLU training pipeline can be found on our GitHub repository.

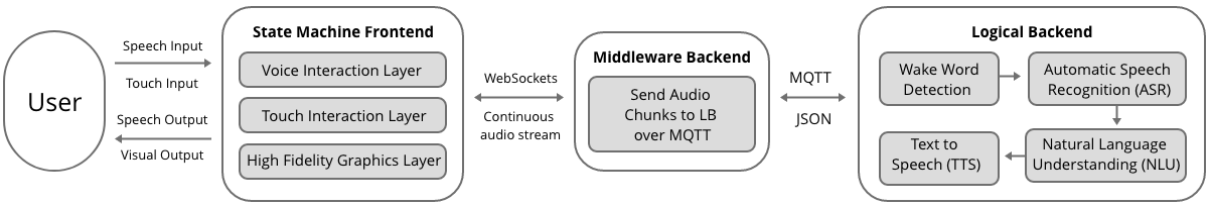

Figure 3: Architecture diagram for the high fidelity prototype. Touch input is processed by the same pipeline.

Table 1: Evaluating using 1964 user queries for training & 839 for testing, respectively.

Recognition	Precision	Recall	f1-score
Intent	0,975	0,973	0,971
Entity	0,948	0,958	0,944

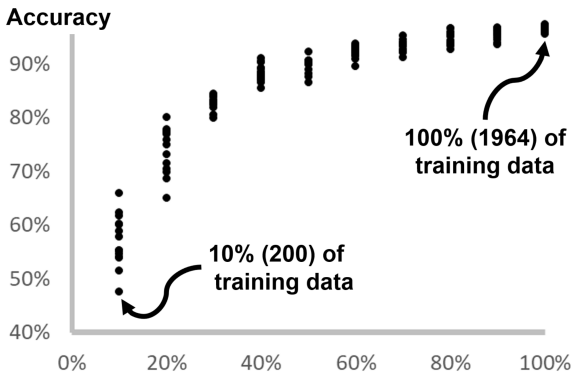

Figure 4: Evaluation using different amounts of training data $\mathcal{D}$.

4.2 Amount of Training Data vs. Intent Recognition Accuracy

We addressed this problem by running 200 experiments with different numbers of user queries $\mathcal{D}^{p_i}$ in each training phase. The process of successive enrichment of training data was simulated by starting with one-tenth of the original training data, denoted by $\mathcal{D}^{1/10}$, and repeatedly sampling about 200 new training examples until we collected all queries $\mathcal{D}$. This process yielded ten reduced training data sets $\mathcal{D}^{p_1} \subset \ldots \subset \mathcal{D}^{p_{10}} = \mathcal{D}$, where $p_i = i/10$, $i = 1, \ldots, 10$, denotes the ratio of the original amount of training data.

We simulated the above collection process 20 times, each time starting with a different randomly sampled set $\mathcal{D}^{1/10}$ resulting in different reduced training sets in each iteration. We used the same training parameters for the DIET models and the same test queries $\mathcal{T}$ as in Section 4.1. Evaluation started with ratio $p_1 = 1/10$ of the available training data and ended with the complete training data $\mathcal{D}$. Figure 4 shows the intent accuracy evaluation in more detail. When using 10% of the training data, the accuracy varied between 47% and 62% depending on the query selection. This means one may be lucky getting suitable queries, but a small number of user queries to train a NLU model is not reliable. The more training data is used, the higher the accuracy with decreasing dispersion.

5 Discussion and Future Work

This prototype is still a work in progress and was developed to test the recipe guidance, therefore it is not yet integrated into a device. The dataset is generic to multimodal assistants in the kitchen context, and may provide a basis for further research. To prove that the approach is generic to the kitchen context, similar work with emphasis on representative studies with larger data sets is needed. At the time of writing this paper, we recorded a subset of the NLU corpus in an anechoic chamber together with kitchen and device noises. The recordings will be published to enable academic and industrial research in this challenging speech domain.

6 Conclusion

The need for creating audio-visual interfaces is growing with the increasing availability of voice assistants in the home environment. Any new device is expected to provide a natural way of interaction. This short paper proposes a novel guided cooking experience, consisting of an audio-visual interface for a multi-functional food processor, in addition to an accompanying prototype with limited functionality.

Moreover, we built a speech corpus based on the proposed prototype design and user studies. Both the prototype and NLU dataset are freely accessible (VoiceCookingAssistant, 2021). This work provides a starting point for further research in the area of Natural and Spoken Language Understanding in the smart cooking domain.

Acknowledgements

This research was conducted as part of the User Experience Design Master Program at Technische Hochschule Ingolstadt. We would like to thank additional contributors **Malik Ali, Sadia Butt, Laura Forster, Nadine Kupitza, Viktoria Langeder, Alina Megos, Liia Mytareva, Subha Nair, Niklas Pachaly, Eunji Park, Daniel Peters, Andreas Riedel, Gülsüm Sanverdi**, and **Christian Sutter** for their work in designing and executing both research and prototype, as well as earlier versions of this paper. In addition, we thank **Manuel Kirschner** and **Tobias Hauser** for their support and domain expertise.

References

Jan Alexandersson, Ulrich Schäfer, Jochen Britz, Maurice Rekrut, Frederik Arnold, and Saskia Reifers. 2015. Kochbot in the intelligent kitchen–speech-enabled assistance and cooking control in a smart home. In *8. Deutscher AAL-Kongress (AAL)*, volume 8, pages 396–405.

James Allen. 1988. *Natural language understanding*. Benjamin-Cummings Publishing Co., Inc.

Prashanti Angara, Miguel Jiménez, Kirti Agarwal, Harshit Jain, Roshni Jain, Ulrike Stege, Sudhakar Ganti, Hausi A. Müller, and Joanna W. Ng. 2017. Foodie fooderson a conversational agent for the smart kitchen. In *Proceedings of the 27th Annual International Conference on Computer Science and Software Engineering*, page 247–253, USA.

Rubén Blasco, Álvaro Marco, Roberto Casas, Diego Cirujano, and Richard Picking. 2014. A smart kitchen for ambient assisted living. *Sensors*, 14(1):1629–1653.

Tom Bocklisch, Joey Faulkner, Nick Pawlowski, and Alan Nichol. 2017. Rasa: Open source language understanding and dialogue management.

P.W.L. Bollen. 2010. BPMN: A meta model for the happy path.

Bruno Bouchard, Kevin Bouchard, and Abdenour Bouzouane. 2020. A smart cooking device for assisting cognitively impaired users.

Tanja Bunk, Daksh Varshneya, Vladimir Vlasov, and Alan Nichol. 2020. Diet: Lightweight language understanding for dialogue systems.

Andreas Fries, Anne-Wiebke Bergmeister, and Marie Spindler. 2018. *Thermomix by Vorwerk – A New Way of Cooking*, pages 73–90. Springer Fachmedien Wiesbaden, Wiesbaden.

Munir Georges, Stephan Kanthak, and Dietrich Klakow. 2014. Accurate client-server based speech recognition keeping personal data on the client. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3271–3275, Florence, Italy. IEEE.

Michael Hansen. 2021. Rhasspy the docs. Accessed: 2021-08-12.

Atsushi Hashimoto, Jin Inoue, Takuya Funatomi, and Michihiko Minoh. 2014. How does user's access to object make hci smooth in recipe guidance? In *Cross-Cultural Design*, pages 150–161, Cham. Springer International Publishing.

Atsushi Hashimoto, Naoyuki Mori, Takuya Funatomi, Yoko Yamakata, Koh Kakusho, and Michihiko Minoh. 2008. Smart kitchen: A user centric cooking support system.

Rohit Ashok Khot and Florian Mueller. 2019. Human-food interaction. *Foundations and Trends® in Human–Computer Interaction*, 12(4):238–415.

Roger A Light. 2017. Mosquitto: server and client implementation of the mqtt protocol. *Journal of Open Source Software*, 2(13):265.

Andrés Lucero. 2015. Using affinity diagrams to evaluate interactive prototypes. In *Human-Computer Interaction – INTERACT 2015*, pages 231–248. Springer International Publishing, Cham.

Alexander Neumann, Christof Elbrechter, and Nadine et al. Pfeiffer-Leßmann. 2017. "kognichef": A cognitive cooking assistant. *KI - Künstliche Intelligenz*, 31(3):273–281.

Felipe Pezoa, Juan L. Reutter, Fernando Suarez, Martín Ugarte, and Domagoj Vrgoč. 2016. Foundations of json schema. In *Proceedings of the 25th International Conference on World Wide Web*.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al. 2011. The kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*.

Inc Rasa Technologies. 2021. Open source conversational ai. Accessed: 2021-08-11.

Markets Research Private Ltd and Markets. 2020. Smart home market. Accessed: 2021-06-21.

Ayaka Sato, Keita Watanabe, and Jun Rekimoto. 2014. Shadow cooking: situated guidance for a fluid cooking experience.

Ulrich Schäfer, Frederik Arnold, Simon Ostermann, and Saskia Reifers. 2013. Ingredients and recipe for a robust mobile speech-enabled cooking assistant for german. In *KI 2013: Advances in Artificial Intelligence*, volume 8077 of *Lecture Notes in Artificial Intelligence (LNAI)*, pages 212–223. Springer.

G. Stemmer, Munir Georges, and J. Hofer et al. 2017. Speech recognition and understanding on hardware-accelerated dsp. In *INTERSPEECH*, pages 2036–2037, Stockholm, Sweden. ISCA.

VoiceCookingAssistant. 2021. Audio-visual-cooking-assistant.