H.E.S.T.I.A: Household Efficient Service Task-Integrated Assistant

Gustavo F. Armênio[1] Bruno B. Silva[2] David Segalle[1] Victor B. Errera[1] João H. C. Pires[2] Enzo H. Gaio[1] Larissa B. Lima da Silva[2] Bruno E. Zenatti[1] Christoffer N. Spring[1] Gabriel Spadafora[1] Luiz G. S. Greca[2] Luiza R. R. Ocampos[2] Matheus P. Lima[1] Marlon O. Vaz[2,3] André C. Cordeiro[1] André S. Oliveira[2] João A. Fabro[1]

[1]DAINF - UTFPR [2]DAELN - UTFPR [3]IFPR

Task Orchestration Experiments and Ontology

Task orchestration in H.E.S.T.I.A. evolved to handle complex and reactive @Home challenges through a hybrid approach: using **State Machines** for deterministic or linear tasks and **Behavior Trees (BTs)** for tasks requiring reactivity, adaptability, and failure recovery — such as *Stickler for the Rules*. This dual approach allows modular construction of behavior logic that can gracefully degrade or switch priorities based on context and sensory feedback. State machines can be easily implemented with YASMIN[4] library and Behavior Trees with the behaviortree_cpp library, both compatible with ROS2.

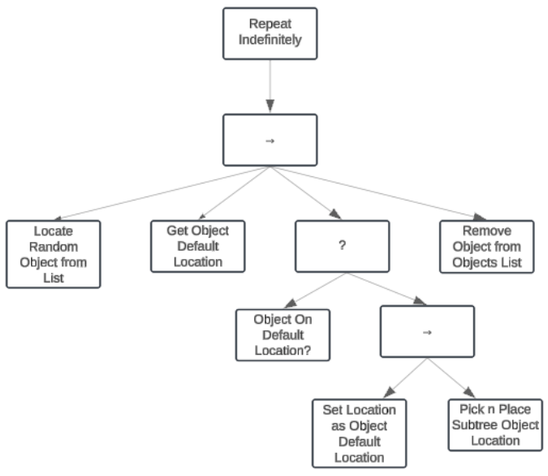

Figure 1. Orchestration of Take Out The Garbage with Behavior Trees

Moreover, an ongoing PhD research is dedicated to the development of a comprehensive **ontology for household environments**, aiming to standardize concepts used in service robotics. This ontology spans robot parts, capabilities, actions, environmental features, and symbolic goals. Aligned with this work, our orchestration framework and technical documentation are already modeled under this unified language. It also fosters collaboration with other Brazilian teams and is envisioned to inform future versions of the **CBR rulebook**, reducing ambiguity in task design and scoring.

Voice Command Interpretation and Context-Adaptable Answering

For complex dynamics and domains of Human-Robot Voice Interaction, we built an extensive pipeline to handle **multiple levels of comprehension with robustness to ambient noise**, common in the competition envirnoments. The pipeline is built for GPSR, EGPRS and common voice interaction requirements in RoboCup@Home tasks by leveraging intent detection, RAG and response generation with both a robust LLM model and a low-resource cost model, RASA[3] for Natural Language Understanding.

Speech Processing Pipeline:

- **Voice Activity Detection:** Initiated by **Silero VAD**[10] for segmenting active voice from background silence in real time.
- **Noise Suppression:** We use **RNNoise**[11], to further isolate human voice in noisy environments.
- **Speech-To-Text:** For multilingual transcription, **OpenAI's Whisper v3 turbo quantized**[7] model offers accuracy and significantly reduced latency and memory usage.
- **Text-To-Speech:** System responses are vocalized using **Coqui TTS**[9], with natural-sounding speech synthesis.

Natural Language Understanding:

- **Contextual QA and Interaction:** With a **Retrieval-Augmented Generation (RAG)** setup using **LLaMA 3.1**[2] and **k-NN** vector search — responses are optimized for a 10s deadline.
- **Command Interpretation:** **RASA[3] NLU** renders effectiveness with intent detection and short-term memory for low-resource, fast and reliable parsing of commands. **LLama 3.1**[2] can integrate rich semantics for interpreting commands as well.

Open-Source Low-Cost Hardware

Our hardware adheres to principles of modular architecture and cost-effectiveness, employing standard components and open-source designs to facilitate reproducibility and cross-team integration.

H.E.S.T.I.A Mobile Base:

- Based on commercial hoverboard motors with FOC drivers and custom ROS2 firmware
- Lightweight, modular structure using aluminum extrusions and metric hardware
- Optimized for reproducibility and adaptation

Figure 2. H.E.S.T.I.A mobile base

Theseus — High-Payload Manipulator:

- 4 DOF open-source arm with 25:1 cycloidal gearboxes and AS5048A magnetic encoders
- FOC drivers compatible with ODrive[6]
- Each actuator costs $100 (total arm $1000) — significantly cheaper than alternatives such as Dynamixel servo-motors
- Designed for 1.5kg lifting capacity
- Designed for replication: `https://github.com/UtBots-Home/theseus-project`

Figure 3. Theseus Arm

Open-Source Software Contributions

- **Navigation:** ROS2 Nav2 + AMCL, SLAM Toolbox
- **Object and Pose Recognition:** YOLOv11[5], Mediapipe Pose[1]
- **Face Recognition:** MTCNN [12], FaceNet512[8]
- **Manipulator Control:** MoveIt integrated with ODrive-based joint drivers (FOC drivers)
- **Human-Robot Interaction:** Silero VAD[10], RNNoise[11], Whisper[7], CoquiTTS[9], RASA[3], LLaMA3[2]
- **Behavior Trees and Ontology-Based Planning:** YASMIN [4], behaviortree_cpp

All source code is available at `https://github.com/UtBotsAtHome-UTFPR`

References

[1] Camillo Lugaresi et al. *MediaPipe: A Framework for Building Perception Pipelines*. 2019. arXiv: `1906.08172 [cs.DC]`. URL: `https://arxiv.org/abs/1906.08172`.

[2] Hugo Touvron et al. *LLaMA: Open and Efficient Foundation Language Models*. 2023. arXiv: `2302.13971 [cs.CL]`. URL: `https://arxiv.org/abs/2302.13971`.

[3] Tom et al. Bocklisch. *RASA NLU: Open-Source Natural Language Understanding*. 2017. URL: `https://rasa.com/docs/rasa/nlu/`.

[4] Miguel Ángel et al. González-Santamarta. *YASMIN: Yet Another State MachINe library for ROS 2*. arXiv preprint. May 2022. DOI: `10.48550/ARXIV.2205.13284`. URL: `https://arxiv.org/abs/2205.13284`.

[5] Ultralytics Jocher and contributors. *YOLOv11: You Only Look Once*. 2024. URL: `https://github.com/ultralytics/yolov5`.

[6] Alexander Jörn-Dieckmann and Madcowswe team. *ODrive: Open-Source High-Performance Motor Controller*. 2024. URL: `https://github.com/madcowswe/ODrive`.

[7] Alec et al. Radford. *Whisper: Robust Speech Recognition via Large-Scale Weak Supervision*. OpenAI Research. 2023. URL: `https://openai.com/research/whisper`.

[8] Florian Schroff, Dmitry Kalenichenko, and James Philbin. *FaceNet: A Unified Embedding for Face Recognition and Clustering*. arXiv preprint. 2015. DOI: `10.48550/arXiv.1503.03832`. URL: `https://arxiv.org/abs/1503.03832`.

[9] Coqui Team. *Coqui TTS: Open-Source Text-to-Speech*. 2024. URL: `https://github.com/coqui-ai/TTS`.

[10] Snakers4 Team. *Silero VAD: A High-Performance Voice Activity Detection Model*. 2023. URL: `https://github.com/snakers4/silero-vad`.

[11] Jean-Marc Valin and etc. Maxwell. *RNNoise: A Noise Suppression Library Based on Deep Learning*. 2018. URL: `https://github.com/xiph/rnnoise`.

[12] Kaipeng Zhang et al. *MTCNN: Joint Face Detection and Alignment Using Multi-task Cascaded Convolutional Networks*. arXiv preprint. 2016. DOI: `10.48550/arXiv.1604.02878`. URL: `https://arxiv.org/abs/1604.02878`.