Human Interaction and Emerging Technologies (IHIET-AI 2025), Vol. 161, 2025, 260–267
https://doi.org/10.54941/ahfe1005917

AI-Powered Auditory Control and Augmented Reality Interfaces for UAVs—A Contactless Control and Situation Awareness Concept

Joshua Gehlen, Alina Schmitz-Hübsch, Sebastian Thomas Handke, and Wolfgang Koch

Fraunhofer FKIE, 53343 Wachtberg, Germany

ABSTRACT

Unmanned Aerial Vehicles (UAVs) are increasingly utilized in military and civilian tasks like search and rescue, however, traditional operation methods can be risky in hazardous situations. This article presents a novel UAV control concept leveraging artificial intelligence (AI) and Augmented Reality (AR) technology, allowing operators to manage drones without handheld devices through audio-based input and output. The suggested system employs headsets and AR glasses to provide real-time visual feedback, enhancing situational awareness and decision-making by displaying critical data such as UAV position and detected hazards within the operator's field of view. The concept comprises five key components implemented within the Robot Operating System (ROS): Audio Input, Task Allocation, UAV Control, Situation Picture, and Output Units. Speech is processed using models such as Whisper, and commands are interpreted by a Large Language Model (LLM) like GPT-4, ensuring accurate recognition even in noisy environments. Initial experiments show high command recognition accuracy, indicating the concept's potential for reliable UAV control in real-world scenarios. Overall, this approach aims to improve operational efficiency and safety in UAV operations, with future work focusing on system refinement and advanced language processing.

Keywords: UAV-control, Augmented reality, Artificial intelligence, Speech-based interaction

INTRODUCTION

The use of Unmanned Arial Vehicles (UAVs) has significantly increased in recent years across both military and civilian applications, such as search, rescue, and disaster response (Luftfahrt-Bundesamt, 2023). Military special operations are also supported by UAVs for reconnaissance purposes. In most applications, one team member must focus on controlling the UAV using a remote control or a ground control station (see Figure 1), which often exposes them to acute danger. For example, in civilian scenarios, the drone pilot may be part of a rescue team and needs to continue their activities such as climbing while operating the drone. In military applications, the drone pilot might need to protect himself while managing the drone. To address this issue, we propose a concept that enables contactless UAV

control powered by AI. For this purpose, audio-based in- and output in combination with Augmented Reality (AR) in the form of a head-up display or glasses is considered. This combination enables the system not only to receive commands, but also to provide an acoustic and visual feedback, and consequently display a complex situation picture.

AR has progressively garnered attention as a powerful tool for visualization of situation pictures. Previous research and applications have demonstrated its potential to enhance the perception of and interaction with spatial information (Qiu, Ashour, Zhou, & Kalantari, 2023). For instance, AR has been employed in military and emergency response settings to overlay critical data onto real-world views, thereby aiding decision-making in high-stress environments (Livingston et al., 2011). The use of AR headsets like the Microsoft HoloLens to provide first responders with dynamic, real-time information such as maps, hazard locations, and routes were considered in Furmanski, Azuma and Daily (2002).

Figure 1: Example of 3D-model of drone ground control. Source: https://www.turbosquid.com/de/3d-models/3d-uav-mobile-ground-control-station-pbr-1811494.

Overall, existing studies and implementations consistently highlight AR's capability to provide immersive, intuitive, and efficient interfaces for situation awareness (Woodward, 2023) not only in military applications but also in medical. Situational awareness is the perception of elements in the environment within a volume of time and space, including the meaning and a projection into the near future (Endsley, 1988). In most scenarios discussed in the literature, UAV operators need to have at least one hand free for control. If this condition is met, the operator can use a traditional controller interface (as shown in Figure 1), gesture control, or brain control (Tezza & Andujar, 2019). In the case of brain control, a complex electroencephalography (EEG) headset can be used to translate the measured brain activity into commands. For gesture control, innovative methods in hand-gesture recognition are being developed, frequently integrated with augmented reality (Mathew, Westhoven, Conradi, & Alexander, 2018).

In comparison, the proposed auditory-based concept requires only a microphone, headphones, and an AR device. Since the scenario assumes the operator's hands are occupied, neither a game controller nor gesture control can be used. Unlike brain control, audio control is simpler, more cost-effective, and less error prone. It also eliminates the need for wiring, making

it less intrusive. Additionally, brain control is often hindered by high noise levels, whereas audio control proves to be more reliable. Furthermore, gesture control involves learning complex movements, while audio input offers a much faster learning curve with fewer errors.

Speech recognition systems for control purposes can be found in many applications e.g. in assistance systems for cars, home automation and UAV control (Contreras, Ayala, & Cruz, 2020; Rajapaksha, Illankoon, Halloluwa, Satharana, & Umayanganie, 2019). Controlling a UAV manually is a complex task, requiring a high level of abstraction for auditory commands to ensure effective operation. In addition, a higher level of automation allows the operator to focus on the situation picture and the decision-making. The possibility of abstract commands is supported by the autonomous sensor management in our concept, which allows an optimal flight path adaptation without user interaction.

As an enabler, our system uses a simple but effective Deep Learning pipeline. Combined with the power of large language models (LLM), it is possible to extract commands from larger context based on speech recognition. For the auditive pipeline, we use off-the-shelf products like Whisper and GPT-4 from OpenAI. Additionally, a task-allocation unit provides effective management of possible algorithms. For system communication, we apply the Robot Operating System (ROS) which is a widely used framework for robotic applications. In the subsequent sections, the system is outlined, followed by a description of the initial experiments and a summary.

THE SYSTEM

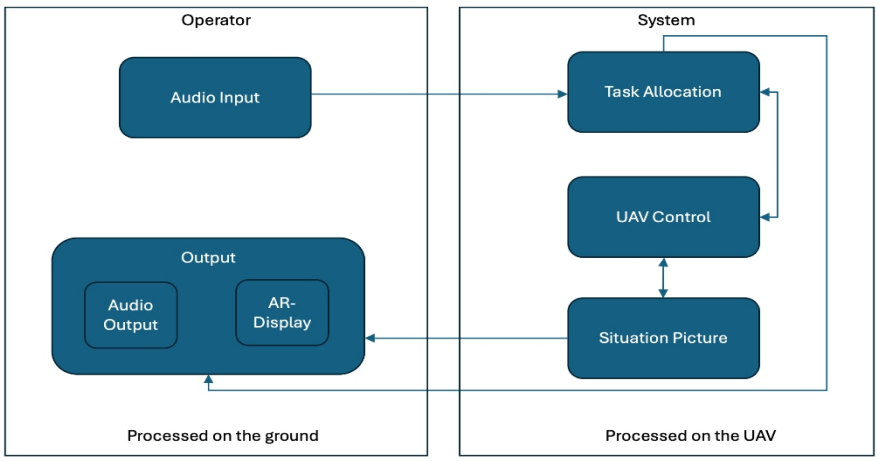

Figure 2: Overview of the system.

The proposed system consists of five units. All units are implemented as ROS nodes. Thus, a simple and effective communication is enabled. An overview of the system is shown in Figure 2.

Audio Input

The audio input unit processes the user's auditory input by recording it with a microphone and subsequently converting it into text using a speech-to-text machine learning model (see Dhanjal and Singh (2024) for a review). For the experiments, we use models like Whisper (Radford et al., 2022) or Chirp (Zhang et al., 2023). While Whisper is trained on a large-scale weak supervision dataset, which also contains machine transcribed samples, the chirp model is firstly trained on unsupervised data and secondly fine-tuned on supervised data.

Following, the transcribed audio input is processed by a LLM like Llama 3 (Meta, 2024) or GPT-4 (OpenAI, 2023). As an input, the LLM receives the transcribed text in combination with all possible commands and the instruction to determine one of the given commands from the input context.

Dividing this process into multiple steps as described has the advantage that the LLM generalizes better than the standard approaches like finding specific audio snippets in the input. As the unit's output, the command is then passed to the task allocation.

Task Allocation

The task allocation unit receives the command and processes it into multiple subtasks. Thus, it also determines the exact order of execution. As an example, we consider the command "search person". The allocation algorithm first checks the current state of the UAV. If the UAV is already busy, this is reported to the output unit for user feedback. Otherwise, the command is started. As an example, we will consider two different cases. In the first setting, the UAV is on the ground. Following the allocation algorithm split, the search command is divided in multiple subtasks:

- Check for take off
- Start the UAV and take off
- Reaching the flight height
- Begin to search using dynamic path planning.

Each of these subtasks is then executed by the UAV control unit and can be aborted due to errors, which are communicated back to the user by visual or auditive feedback.

As a second setting, the UAV is already in the air. In this setting, only the task of searching is executed.

Consequently, the task allocation is a complex challenge. One approach to solve this are algorithms based for example on finite-state machines. Depending on the current state, the UAV spatial position can change also. This information is used in the UAV control unit.

UAV Control

In the UAV control unit, the commands are translated into executable UAV commands based on the situation picture. For each command, waypoints and sensor settings are chosen to optimize battery consumption, task fulfilment and environmental conditions. Theoretically, the described problem is an optimization problem. Thus, in addition to path planning

algorithms such as A*, optimization algorithms like knowledge graph-based optimization can be applied (Hwang, Kim, Lim, & Park, 2003). Besides, the task related appropriate detection algorithms are chosen according to the available sensors.

In the example of the "search" command, a suitable search pattern is planned and an algorithm for detecting people in the UAV's sensor data stream is selected. If a person is detected, this is reported to the situation picture unit.

The Situation Picture

The situation picture contains all information about the current scenario, e.g. a map of the environment and points of interests like already detected persons. The information is used for planning flight paths, giving auditive output to the user and visualizing information on the AR-display. The situation picture unit is the main knowledge base of the system. It is updated by the UAV control unit and the audio input unit. Its data has to be precise, reliable and should be able to be restructured in a machine-readable form for automatic processing. The situation picture can be supplemented using AI (Gehlen, Govaers, Ulmke, & Fischer, 2023).

Output

The output unit processes the information of the situation picture. It controls the audio speaker and the AR-display. On the audio channel, alerts can be given to the user in form of spoken text or alert signals. The AR-display shows an abstract representation of the current situation. For example, the position of the UAV, the search pattern or detected persons can be displayed. It is important to apply an intuitive and abstract design scheme to prevent the user from being overwhelmed by the information as well as distracted from the current situation.

In the case of the "search" command, the user can see the search pattern on the AR-display and additionally the detected persons. If a person is detected, the user is informed by an auditive and visual signal.

As an example, Figure 3 illustrates a simplified process diagram of a "search person" command with the initial state of the UAV on the ground.

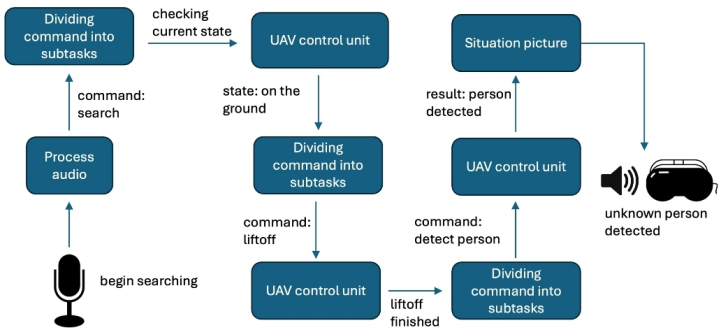

Figure 3: Process diagram of an example processing for the "search command".

FIRST EXPERIMENTAL VALIDATION

As a first step to validate the effectiveness of the system, the audio processing pipeline was constructed and tested for the experiments, we use the off-the-shelf software Whisper for the audio processing and GPT-4 as the LLM in combination with the OpenAI Function Calling API.

In the test, we limit the commands to "start", "land", "hover", "search", "return" and "abort". For automatic testing purposes, we use the LLM and a text to speech model to generate example input. Thus, the test pipeline consists of the following steps:

1. Command generation by the LLM based on the chosen command, for example "search."
2. Use text to speech model for audio input generation.
3. Process the audio input by firstly using a speech to text model and secondly extract the command.
4. Check the input and output command for equality.

For the command "search", the LLM generates for example the following commands:

- Please navigate the area and identify any notable landmarks or points of interest.
- Please activate the search mode and start scanning the area for any points of interest or specific targets that need to be located.
- Please begin a comprehensive scan of the designated area, carefully noting any anomalies or points of interest. Make sure to cover the entire perimeter and report back with high-resolution images and detailed analysis.
- Please initiate a thorough search of the designated area and identify any notable objects or anomalies.
- Could you please remotely access the surrounding area to see if you can locate the object we discussed?
- Begin a systematic search of the area.
- As our primary objective is to explore and gain deeper insights into the terrain, I would like you to commence a thorough survey of the designated area. Scan comprehensively and ensure to cover every detail for a more accurate analysis.
- Survey the surrounding area and relay any pertinent information back.
- Begin scanning the surroundings meticulously for any noteworthy details or objects.
- Please survey the area and gather any information you can about the surroundings, providing any pertinent details that could be of interest.

In our tests, we tested each command with 100 samples. The commands were recognized correctly for 595 of 600 examples even if the command did not appear verbatim.

The generalization and context-aware understanding offered by the LLM play a crucial role in enabling the proposed system. On one hand, it significantly reduces the learning curve for the operator, as there is no need to rely on predefined or rigid command structures. Instead, the system's

ability to interpret natural, flexible input allows the operator to interact more intuitively, enhancing usability and efficiency. This adaptive approach not only minimizes training time but also increases the overall effectiveness of the system by making control more seamless and accessible.

CONCLUSION

This paper introduces an innovative approach to achieve contactless control of UAVs using AI-powered audio and AR interfaces. By integrating deep learning powered speech recognition systems in combination with large language models instead of sophisticated language processing techniques, our method aims to enhance operational efficiency and situation awareness. This approach promises to mitigate risks associated with manual control in challenging environments.

Our objective is to establish a reliable framework that enhances safety and efficiency in UAV operations. By leveraging AI and augmented reality integration, we aim to address current operational challenges.

Looking ahead, our research will focus on refining system integration and optimizing task allocation algorithms within the operational framework. We aim to enhance the interaction model by exploring the capabilities of language processing technologies for real-time command interpretation and adaptive responses. Additionally, we plan to integrate a chat capability over the situation picture, leveraging AI models to facilitate enhanced communication and decision-making.

REFERENCES

Contreras, R., Ayala, A., & Cruz, F. (2020). Unmanned Aerial Vehicle Control Through Domain-based Automatic Speech Recognition. *CoRR, abs/2009.04215*. Retrieved from https://arxiv.org/abs/2009.04215.

Dhanjal, A. S., & Singh, W. (2024). A comprehensive survey on automatic speech recognition using neural networks. *Multimedia Tools and Applications, 83*, 23367–23412.

Furmanski, C., Azuma, R., & Daily, M. (2002). Augmented-reality visualizations guided by cognition: Perceptual heuristics for combining visible and obscured information. *Proceedings. International Symposium on Mixed and Augmented Reality*, (pp. 215–320).

Galvani, M. (2019, February). History and future of driver assistance. *IEEE Instrumentation & Measurement Magazine, 22*, 11–16. doi: 10.1109/MIM.2019.8633345.

Gehlen, J., Govaers, F., Ulmke, M., & Fischer, A. (2023). Architecture and design of AI based air situation assessment. *2023 IEEE Symposium Sensor Data Fusion and International Conference on Multisensor Fusion and Integration (SDF-MFI)*, (pp. 1–7). doi: 10.1109/SDF-MFI59545.2023.10361390.

Hwang, J. Y., Kim, J. S., Lim, S. S., & Park, K. H. (2003). A fast path planning by path graph optimization. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans, 33*, 121–129. doi: 10.1109/TSMCA.2003.812599.

Livingston, M., Rosenblum, L., Brown, D., Schmidt, G., Julier, S., Baillot, Y.,... Maassel, P. (2011, July). Military Applications of Augmented Reality. doi: 10.1007/978-1-4614-0064-6_31.

Luftfahrt-Bundesamt. (2023). *Statistik unbemannter Luftahrtsysteme - Registrierte Betreiber*. Retrieved June 9, 2024, from Statistik unbemannter Luftahrtsysteme - Registrierte Betreiber: https://www.lba.de/SharedDocs/Downloads/DE/SBl/SBl3/Statistiken/Betrieb/UAS_Betreiber.html?nn=4357312.

Mathew, T., Westhoven, M., Conradi, J., & Alexander, T. (n.d.). Touch-based Eyes-free Input for Head-Mounted Augmented Reality Displays. *cit. on*, 13.

Meta. (2024). *Meta Llama 3*. Retrieved from Meta Llama 3: https://llama.meta.com/llama3/.

OpenAI. (2023). GPT-4 Technical Report. *ArXiv, abs/2303.08774*. Retrieved from https://arxiv.org/abs/2303.08774.

Qiu, Z., Ashour, M., Zhou, X., & Kalantari, S. (2023). NavMarkAR: A Landmark-based Augmented Reality (AR) Wayfinding System for Enhancing Spatial Learning of Older Adults. *NavMarkAR: A Landmark-based Augmented Reality (AR) Wayfinding System for Enhancing Spatial Learning of Older Adults*.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). Robust Speech Recognition via Large-Scale Weak Supervision. *Robust Speech Recognition via Large-Scale Weak Supervision*.

Rajapaksha, S., Illankoon, V., Halloluwa, N. D., Satharana, M., & Umayanganie, D. (2019). Responsive Drone Autopilot System for Uncertain Natural Language Commands. *2019 International Conference on Advancements in Computing (ICAC)*, (pp. 232–237). doi: 10.1109/ICAC49085.2019.9103346.

Tezza, D., & Andujar, M. (2019). The State-of-the-Art of Human–Drone Interaction: A Survey. *IEEE Access, 7*, 167438–167454. doi:10.1109/ACCESS.2019.2953900

Zhang, Y., Han, W., Qin, J., Wang, Y., Bapna, A., Chen, Z.,... others. (2023). Google usm: Scaling automatic speech recognition beyond 100 languages. *arXiv preprint arXiv: 2303.01037*.