



Inter-Facing Difficulty



Group Members: Alejandro Marin (816035363), Pranav Soondar(816030105), Joelle Ramchandar(816035123)



VAL: Interactive Task Learning with GPT Dialog Parsing

Presented by Alejandro Marin

01



Background

Authors:

Lane Lawley



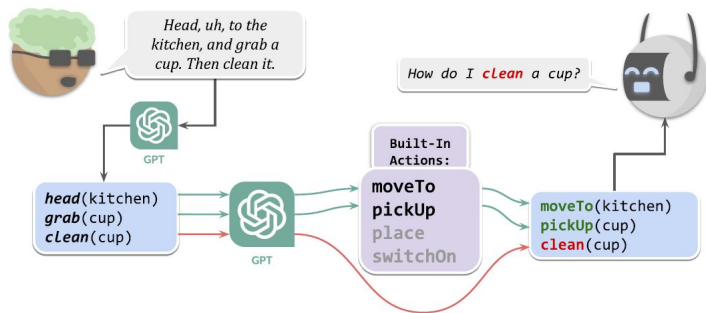
Christopher J. MacLellan



Authors: Both authors are affiliated with the Georgia Institute of Technology and contribute significantly to the Human-Computer Interaction (HCI) field, particularly in artificial intelligence (AI), and natural language processing

Publication Details: This paper was presented at CHI 2024, a prominent conference on Human Factors in Computing Systems during May 11–16, 2024, in Honolulu, Hawaii

Abstract



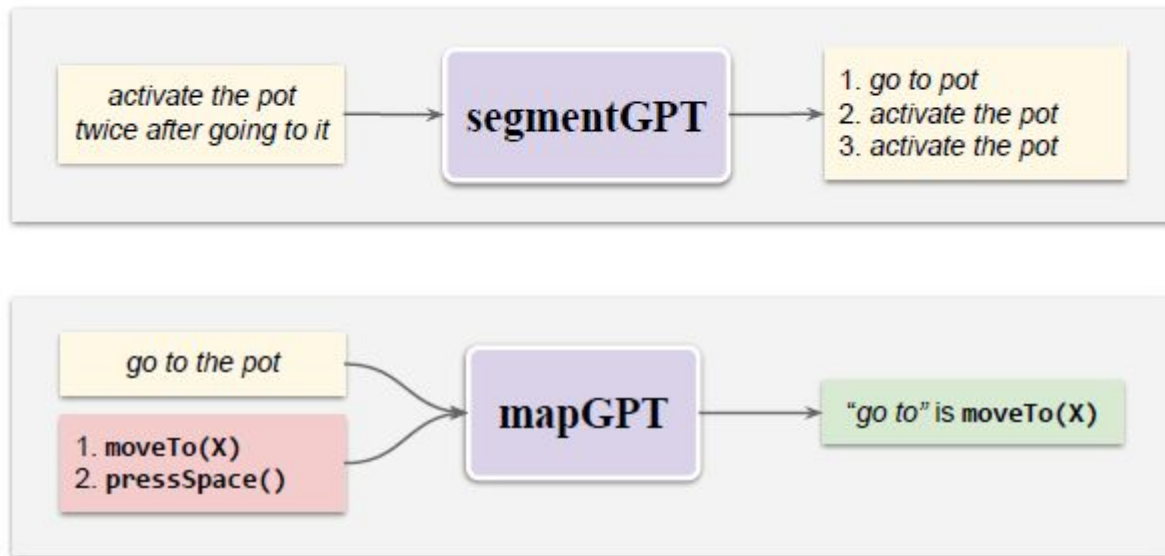
Findings: User studies showed that participants could successfully teach VAL to execute tasks using natural language. This approach mitigates brittleness in language parsing while allowing for task reusability.

Main Objectives: The paper introduces VAL (Verbal Apprentice Learner), a hybrid AI system that integrates GPT-based large language models (LLMs) with traditional learning algorithms to circumvent the limitations of interactive task learning (ITL) systems.

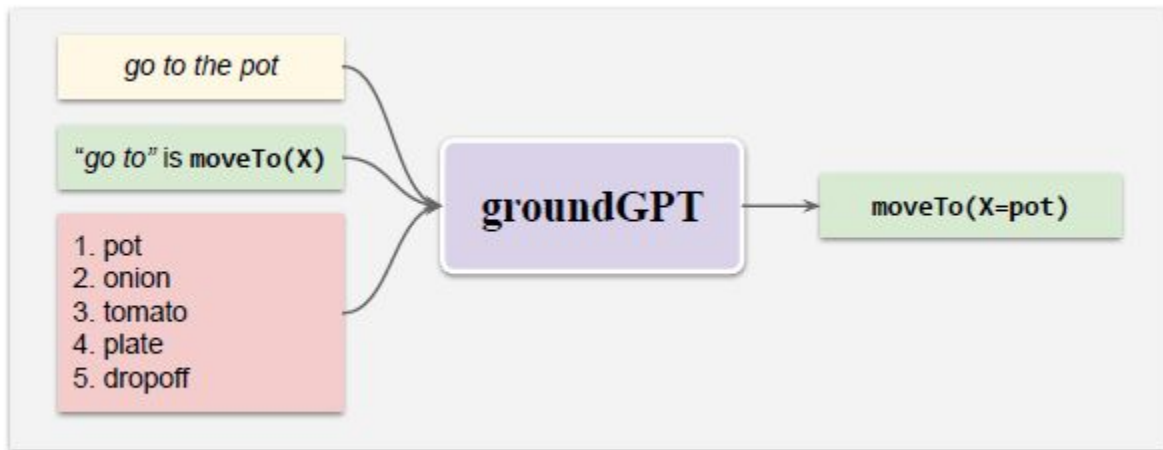
Contributions: The following key contributions are made in this paper:

- 1) It is a neuro-symbolic hybrid approach to ITL where natural dialogue is used.
- 2) The development of a VAL system which mitigates "cascading errors" by the use of confirmation and undo operations.
- 3) Is evaluated by a user study.

Abstract

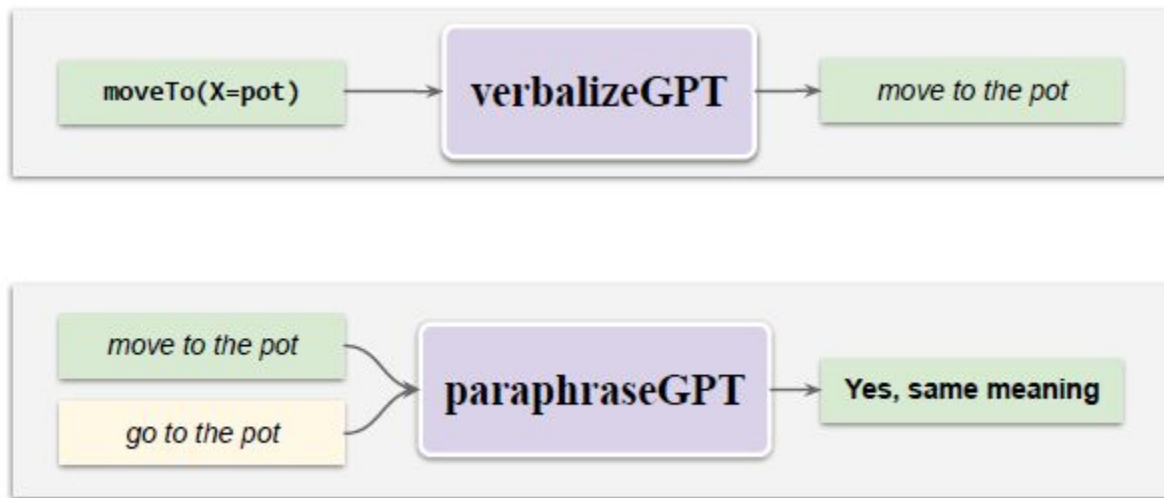


Abstract

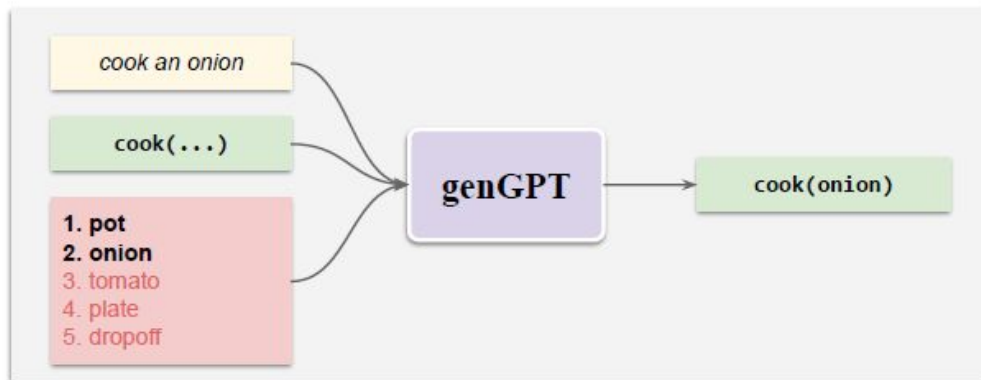




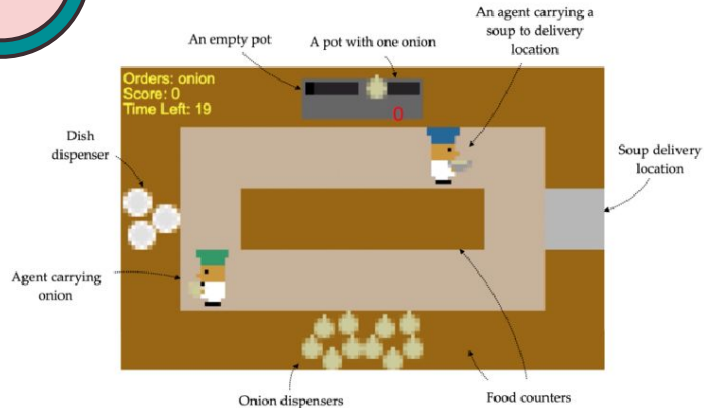
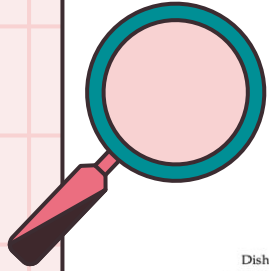
Abstract



Abstract



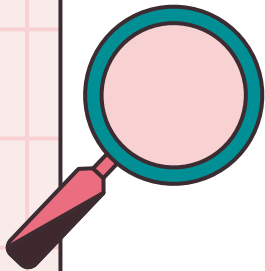
Methodology




Research Approach: Experimental design - usability study.

Experimental Design: The authors designed user studies using the Overcooked-AI game environment, where participants taught VAL to perform cooking tasks. They monitored user interactions, focusing on how effectively VAL learned tasks through clarification dialogs.

Methodology





VAL

15:46


I think these are the objects of "*get an onion*":

```
get ( onion )
```

Is that right?

Yes

No



VAL

15:46

Sorry about that. Could you help me pick the right objects?

```
get ( onion 
```

+

-

onion

▼

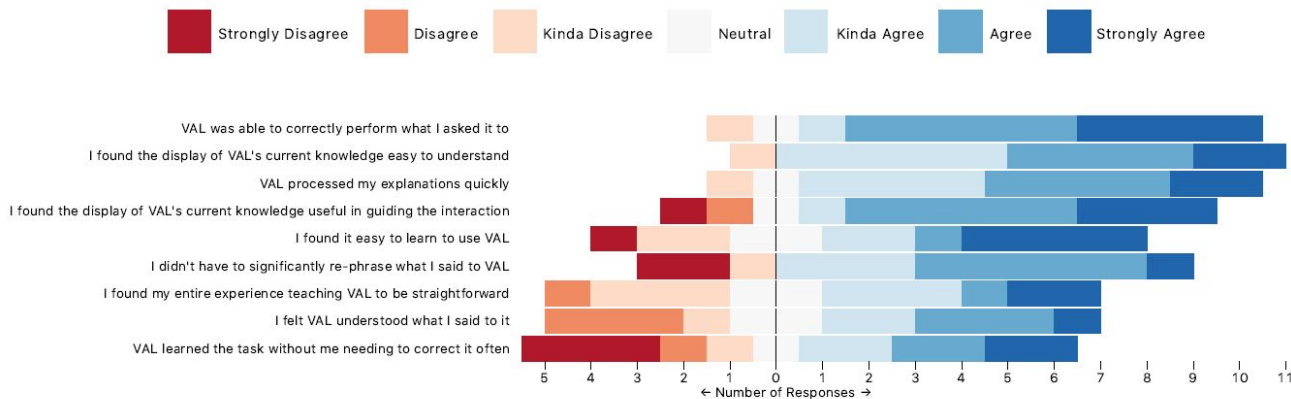
)

Submit

Data Collection: Objective performance metrics, such as success rates of GPT subroutines, use of confirmatory dialogs, and performance of different GPT models were collected while the experiment was conducted. The performance of the environment was also monitored. Subjective responses from participants regarding ease of use were also collected from a survey at the conclusion of the experiment.

Results

Key Findings: Most users could teach VAL effectively, although some tasks required frequent clarification. VAL demonstrated a high success rate in understanding and executing user commands, with some GPT subroutines achieving success rates up to 97%. Performance was better when GPT-4 was used instead of GPT-3.5



Results

Key Findings: Most users could teach VAL effectively, although some tasks required frequent clarification. VAL demonstrated a high success rate in understanding and executing user commands, with some GPT subroutines achieving success rates up to 97%. Performance was better when GPT-4 was used instead of GPT-3.5

GPT Subroutine	Success Rate
segmentGPT	93% user approval
mapGPT	82% user approval (gpt-3.5-turbo) 97% user approval (gpt-4)
groundGPT	88% user approval
genGPT	81% user approval
verbalizeGPT + paraphraseGPT	79% true positive rate 99% true negative rate



Discussion

Implications for HCI:

- VAL addresses key challenges in ITL by improving the flexibility and usability of task-learning systems for non-technical users.
- Its integration of LLMs allows for more natural language instruction while ensuring that the knowledge it acquires remains understandable and easy to interpret.

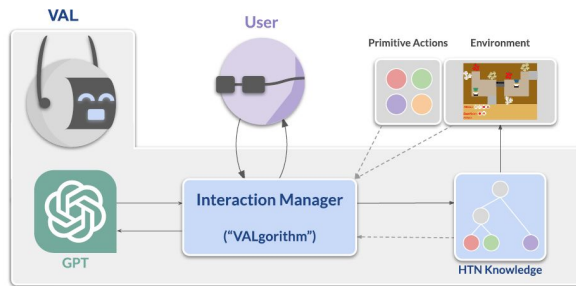
Limitations:

- Limited in current state to only text input whereas other models can use gestures to help clarify misinterpretations.
- Moreover, VAL's reliance on a proprietary LLM (GPT) raises ethical concerns as it is uncertain how exactly they work. These limitations may impact the system's scalability.

Conclusion

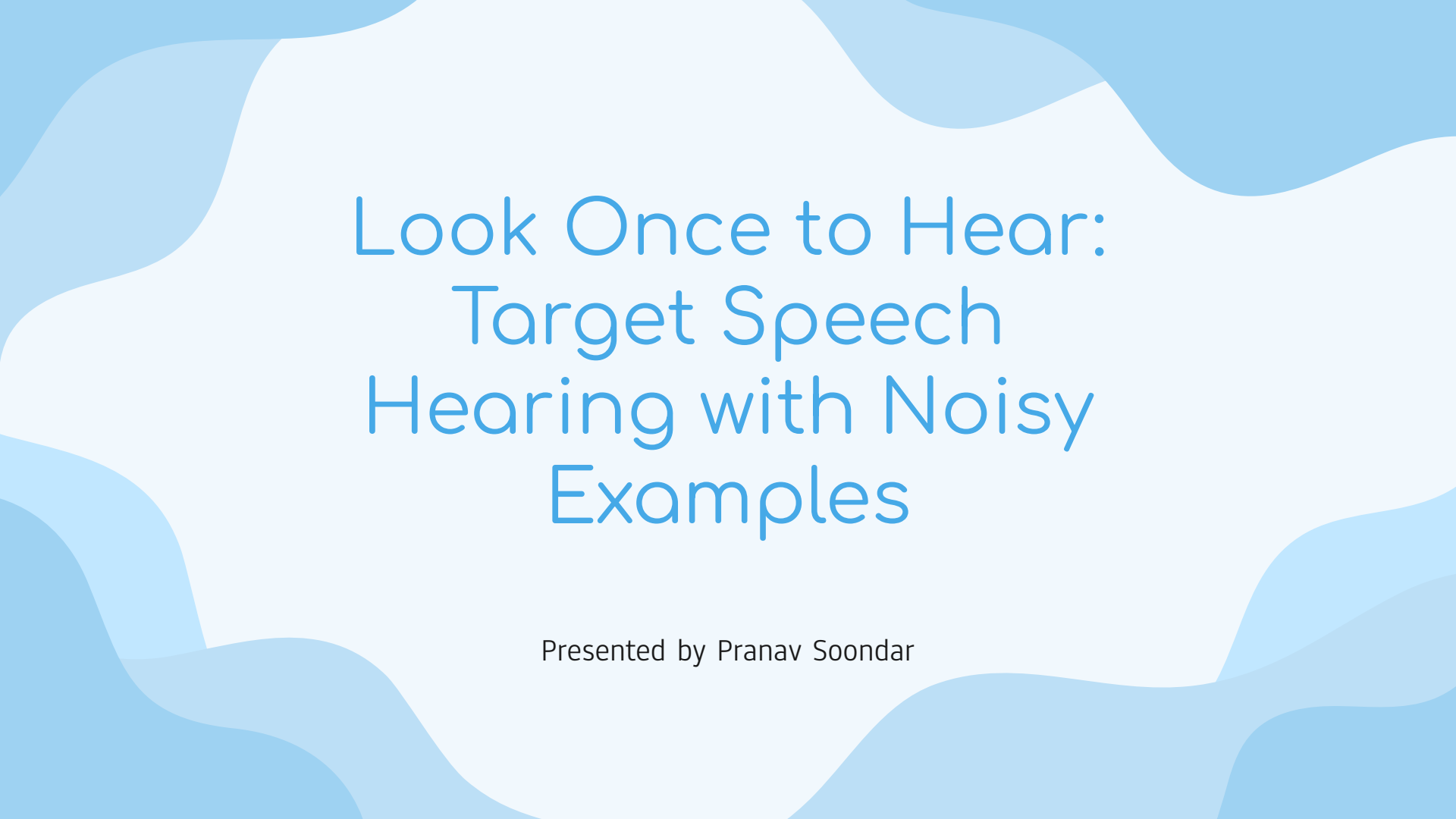
Main Takeaways: VAL is a neuro-symbolic AI system that leverages GPT for natural language parsing, having the ability to learn reusable, interpretable task knowledge from just a few examples. This contrasts with traditional machine learning systems that often require large datasets.

Relevance and Impact: The VAL system is a significant step toward making interactive task learning more accessible to non-technical users by means of natural language interactions. The study shows that VAL has the potential to bridge the gap between human users and AI.



References

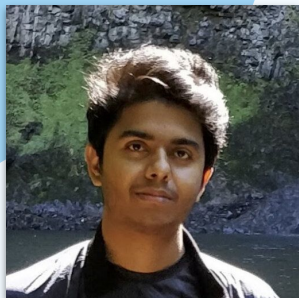
Lane Lawley and Christopher J. MacLellan. 2024. VAL: Interactive Task Learning with GPT Dialog Parsing. In Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24), May 11–16, 2024, Honolulu, HI, USA. ACM, New York, NY, USA, 18 pages. <https://doi.org/10.1145/3613904.3641915>

The background of the slide features abstract, flowing shapes in various shades of blue, creating a modern and clean aesthetic.

Look Once to Hear: Target Speech Hearing with Noisy Examples

Presented by Pranav Soondar

Background



Bandhav Veluri



Malek Itani



Tuochao Chen



Shyamnath Gollakota



Takuya Yoshioka

Bandhav Veluri and Malek Itania are the co-primary student. Bandhav Veluri, Malek Itania, Tuochao Chen and Shyamnath Gollakota are all affiliated with Paul G. Allen School, University of Washington, Seattle. All PhD holders with specialties in machine learning, mobile technologies, speech processing and embedded systems.

Researcher in:
Speech Recognition,
Speech Enhancement,
Speaker Diarization,
Machine Learning

—Presented in the ACM Conference
on Human Factors in Computing
Systems 2024, May 11-16, at Honolulu,
Hawaii, USA

Abstract

Target speaker with interference



Look once at the target speaker



Only hear the target speaker



Objective: To create a system that allows users to focus on the target speaker in noisy environments without prior knowledge of how the speaker sounds. The system must capture a single, short, highly noisy audio clip to be used for speech extraction and enrollment.

Contributions

01

Design and compare two enrollment methods (beamforming and knowledge distillation)

02

Implementing speech hearing neural networks to run in real-time on embedded CPU

03

Using synthetic data to train the system's ability to locate the target speaker in the real world

04

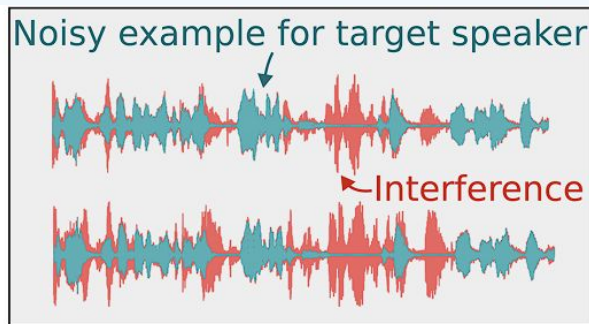
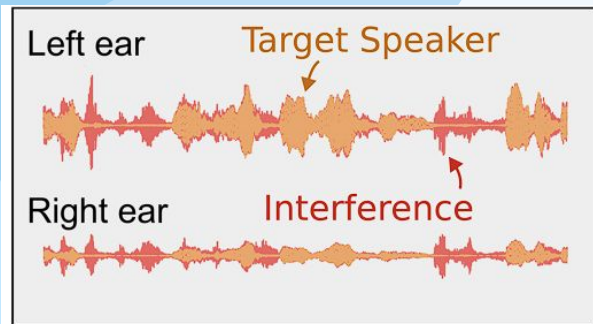
Fine tuning mechanisms which consider the user's head movements and target speaker movement

05

Training neural networks to create generalizations for indoor and outdoor environments

06

Development of an end to end hardware system



Findings

- Signal quality improvement of 7.01 dB using less than 5 seconds of noisy enrollment audio.
- Processing of 8 ms audio chunks in 6.24 ms on an embedded CPU.
- No performance degradation compared to clean enrollment audio examples.

Mixed-Method Research Methodology

Quantitative

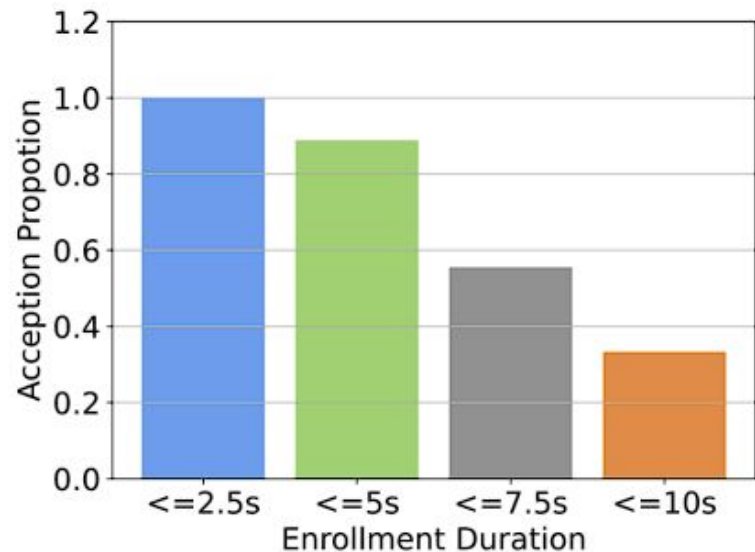
- Finding average target speaker's signal quality improvement in terms of scale invariant signal-to-noise ratio improvement (SI-SNRi) for different scenarios.
- Average runtime over 1000 forward passes

Qualitative

- 21 participants to take our survey and give their opinion to obtain a mean opinion score (MOS).
- System Usability Scale (SUS) questionnaire
- Determining acceptable enrollment time

User preferred enrollment time

Users prefer lower enrollment times with 2.5 seconds or less being the most desirable



Quantitative comparison of enrollment method effectiveness

Knowledge distillation was found to be the preferred method by users.

A low p value confirms the validity of the results

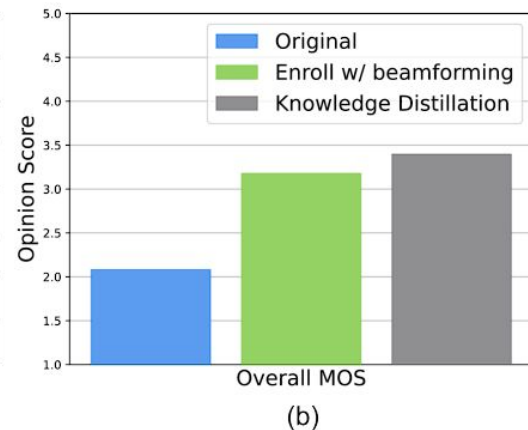
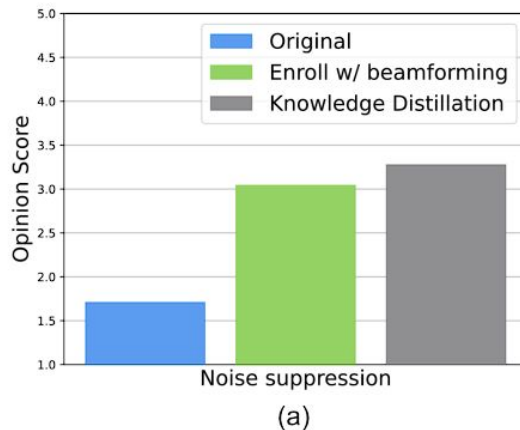
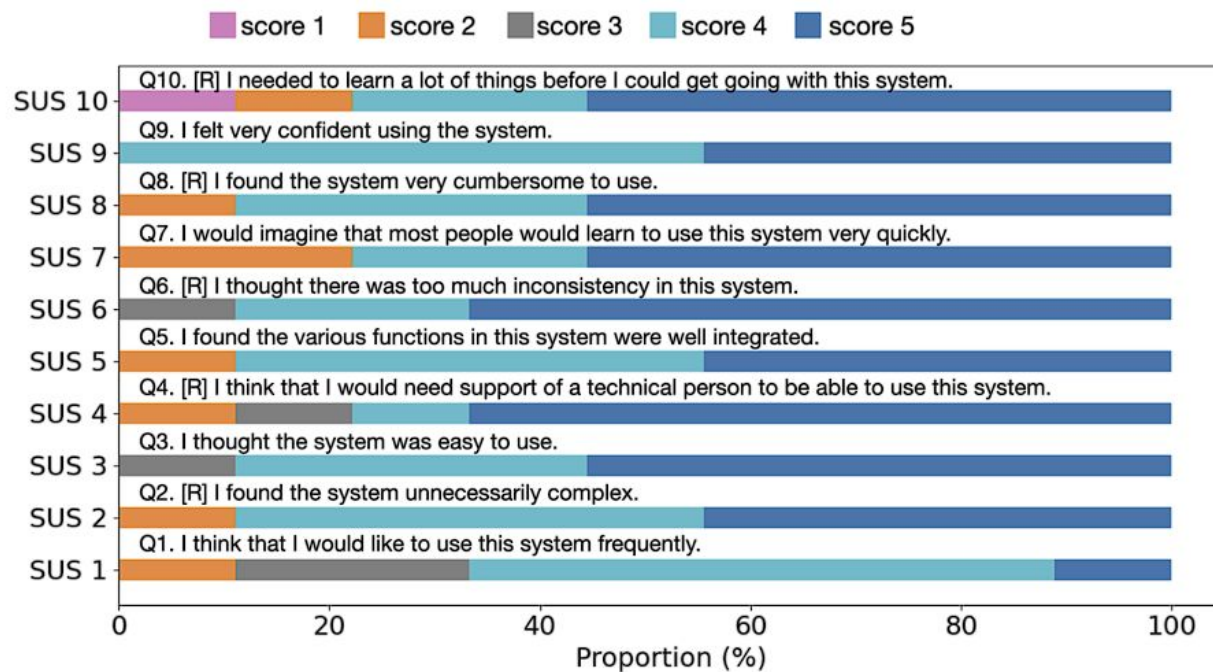


Figure 7: Subjective in-the-wild evaluations. (a) Mean opinion score for the noise suppression quality reported for the raw audio signal and the output using our two enrollment networks, and (b) overall reported mean opinion score. Paired t-tests between knowledge distillation and beamforming approaches resulted in p -values < 0.001 .



System Usability Survey

Prerequisite knowledge was needed, user confidence was high, the system was cumbersome, inconsistent results, good integration of functions, technical support was needed, the system was easy to use, the system was unnecessarily complex and users would like to use the system again

System Performance

Table 1: Benchmarking results on the generated test set. Proposed noisy enrollment methods are evaluated with 3 different audio/speech processing architectures. Performance with clean enrollments is also provided for reference.

Enrollment network	d-vector similarity	Real-time TSH backbone	SI-SNRi (dB)	Params (M)	MACs (GMAC)
Clean	1.0	Streaming TFGridNet	7.40	2.04	4.63
		Waveformer	4.94	1.6	2.43
		DCCRN	6.71	5.54	6.6
Beamformer	0.74	Streaming TFGridNet	4.53	"	"
		Waveformer	2.34		
		DCCRN	4.34		
Knowledge distillation	0.85	Streaming TFGridNet	7.01	"	"
		Waveformer	4.63		
		DCCRN	6.16		

d-vector used to show similarity.

Higher d-vector is better

Higher SI-SNRi is better

Results of Fine Tuning

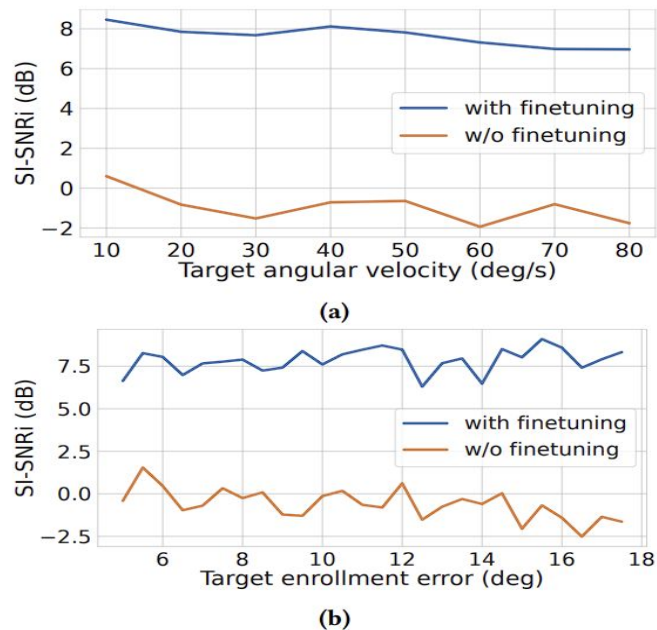


Figure 13: Comparison with and without fine-tuning, when relative motion and enrollment angle error is present.

Higher SI-SNRi is better

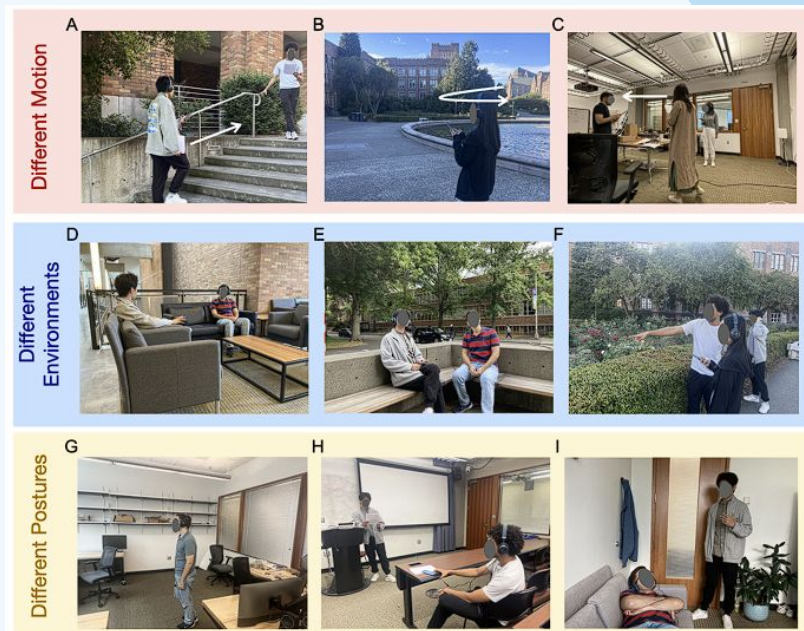


Figure 6: In-the-wild scenarios. Different scenarios in the real-world evaluation of our system.

Discussion

Implications

The system was highly successful in using noisy enrollment which can improve future hearable technology

Limitations

- Limited user group
- Inefficient use of hardware
- Single speaker environments

Conclusion

Presented a system that allows users to focus on the target speaker in noisy environments without prior knowledge of how the speaker sounds using neural networks

This research showcases a method to vastly improve human auditory perception through smart hearing devices.

Applicable scenarios include loud cities, lectures, social events and public spaces.

References

Veluri, B., Itani, M., Chen, T., Yoshioka, T., & Gollakota, S. (2024). Look once to hear: Target speech hearing with noisy examples. *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–16.
<https://doi.org/10.1145/3613904.3642057>

CONTROLLER:

An Auxiliary Tool for Controlling Your Avatar in PC Games Using Natural Motion Mapping

03

PRESENTED BY JOELLE RAMCHANDAR



BACKGROUND

Published - 11 May 2024



Categories : Natural mapping, NUI, Motion Tracking, Intuitive Interaction

2024 ACM CHI

Conference on Human Factors in Computing Systems



Junjian Chen

Professor Associate

5



Yuqian Wang

Professor Associate

3



Yan Luximon

Laboratory for Artificial Intelligence in Design

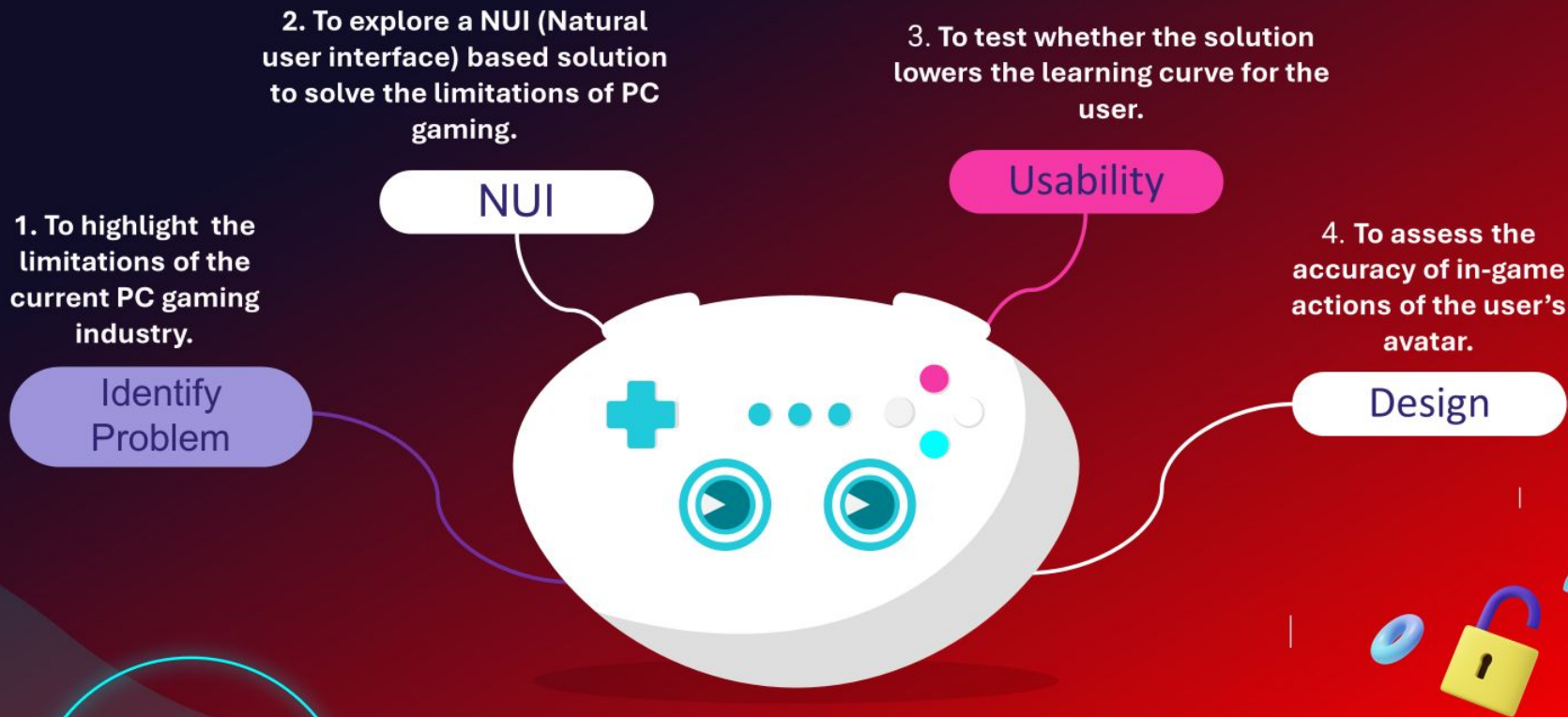
169

Authors are affiliated with The Hong Kong Polytechnic University Hong Kong SAR, China and have multiple publications in the field of HCI.

- Computer Graphics and Computer-Aided Design | Software
- Ergonomics
- Measurement Science and Technology
- Human Movement Science
- IEEE Transactions on Industrial Electronics
- Robotics

Abstract

Main Objectives



Abstract

CAMTROLLER

(1) The introduction of Camtroller (NUI concept) to solve the difficulty of performing complex avatar actions in traditional PC gaming.

Contributions



User Study

(2) The result of a subject study with non-professional players who practiced common operations (Basic), professional player's operations (Pro), and CamTroller to gauge the effectiveness and usability of the NUI design concept.

Abstract

Findings

↓ Difficulty

Camtroller alleviated the difficulty of complex actions, and solved memory burden for users.

↑ Speed

Performance time of in-game actions for Camtroller versus traditional mouse and keyboard was less.

↑ Usability

CamTroller achieved significantly higher intuitiveness than Basic and Pro, lowering the learning curve.

Accuracy

Camtroller reflected a player's live action, instantly to in-game avatar.





Methodology

❑ Research methods-

Qualitative - collecting user experiences and common avatar actions

Quantitative- statistical analysis of data sourced through natural mapping and motion tracking.

❑ Experimental design-

Performing an evaluation test

Executing the NUI concept using CAMROLLER in popular fps single avatar game, PUBG.



Methodology



Data collection techniques:

- ❑ Player Survey based on PUBG Questionnaire (Sojump)- Distributed among users to collect qualitative data
- ❑ Motion Selection and Analysis
Mapping human gestures and features (MediaPipe)- Using a webcam and mouse/keystroke emulator to map quantitative data through machine learning software
- ❑ Observation Activity
Direct observation of volunteers who were instructed to perform avatar actions

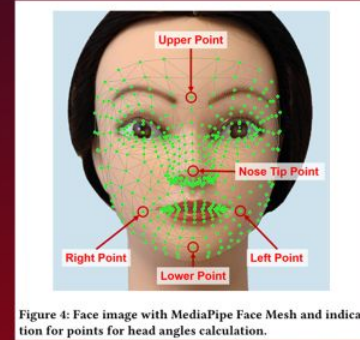


Figure 4: Face image with MediaPipe Face Mesh and indication for points for head angles calculation.

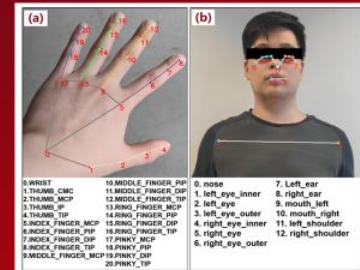


Figure 7: Landmarks of MediaPipe for (a) hand solution and (b) pose solution.

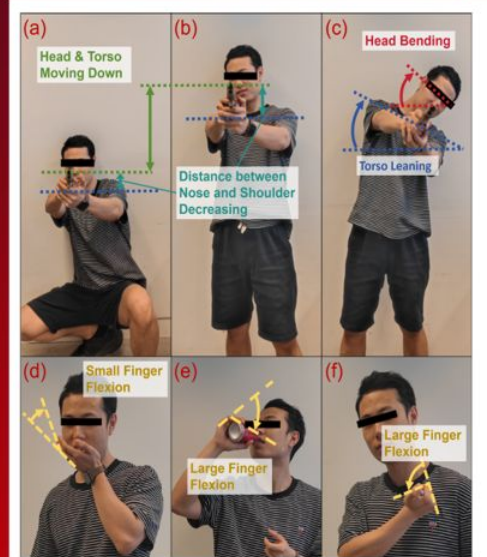


Figure 2: Avatar motions in the physical world and corresponding detectable features illustration for (a) crouching, (b) neutral position, (c) peeking, (d) taking pills, (e) taking a drink, and (f) injecting drugs.

Methodology



Experimental design

❑ Accuracy Evaluation Test

Given that MediaPipe's estimations for the head orientation angles were based on newly collected data, its accuracy requires validation

Equipment:

- Mock head model installed on a tripod with two degrees of freedom (DOF) (CIMAPRO LD-2R) to simulate the motion of a human head.

- Webcam (Rapoo C270AF) as an image-capturing device connected to a laptop (Zephyrus G14 2022) where the program is running.

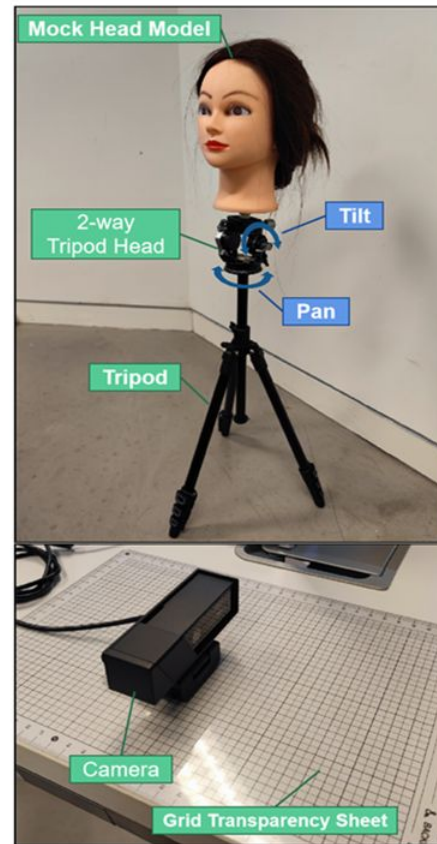


Figure 5: Apparatus for the evaluation of head orientation estimation accuracy

Methodology

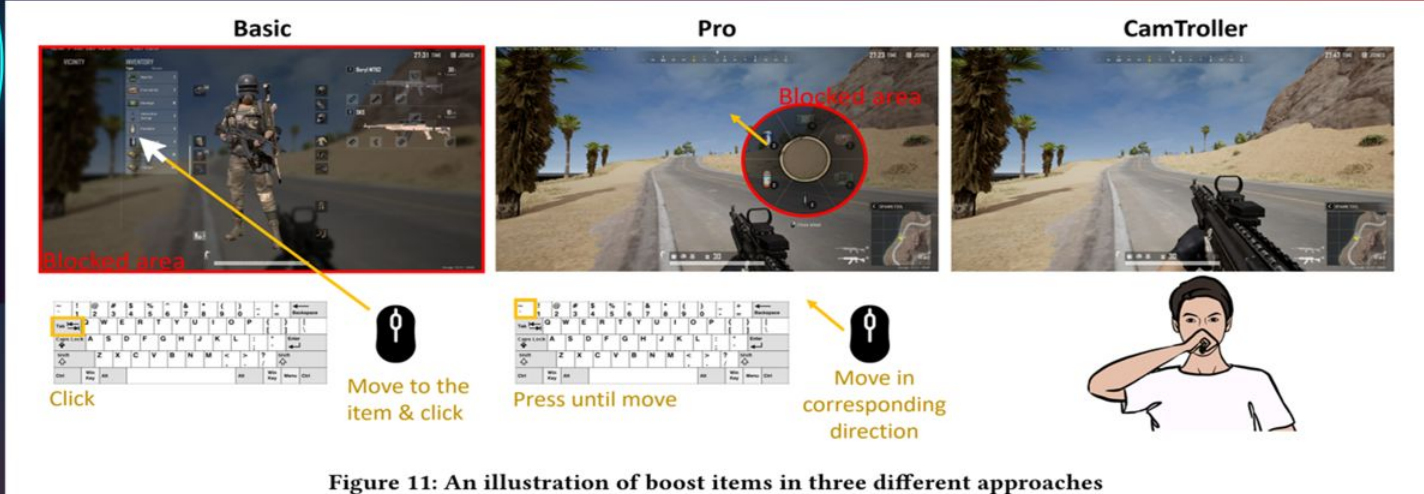


Usability study:

Group 1: Basic- beginners with little to no experience in the game

Group 2: Professional player's operation (pro)- high level of gaming experience

Group 3: Camtroller- 20 minutes training sessions for users and activities to ensure user is familiar with the equipment.



Results

Data Collection Techniques

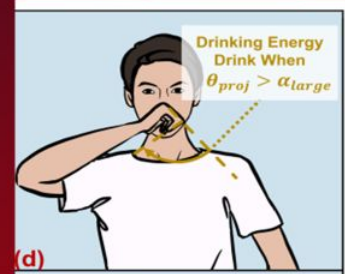
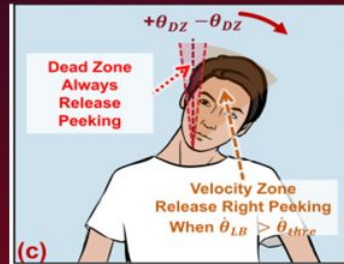
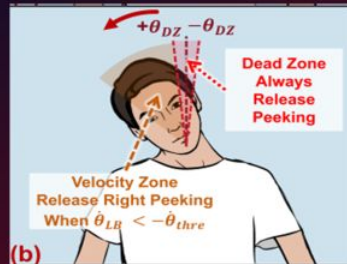
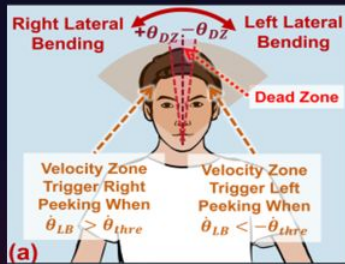


Table 1: Criteria for Different Motions

Motion	Criteria
Left Peeking	$\dot{\theta}_{LB} < -\dot{\theta}_{thre} \ \& \ \theta_{LB} < -\theta_{DZ}$
Release Left Peeking	$(\dot{\theta}_{LB} > \dot{\theta}_{thre} \ \& \ \theta_{LB} < -\theta_{DZ}) \ \text{or} \ (\theta_{LB} > -\theta_{DZ})$
Right Peeking	$\dot{\theta}_{LB} > \dot{\theta}_{thre} \ \& \ \theta_{LB} > \theta_{DZ}$
Release Right Peeking	$(\dot{\theta}_{LB} < -\dot{\theta}_{thre} \ \& \ \theta_{LB} > \theta_{DZ}) \ \text{or} \ (\theta_{LB} < \theta_{DZ})$
Crouching	$(\dot{y}_{nose} > \dot{y}_{thre} \ \& \ k_{thre} \times L_{s2n_0} < L_{s2n} < k_{DZ} \times L_{s2n_0}) \ \text{or} \ (L_{s2n} < k_{thre} \times L_{s2n_0})$
Release Crouching	$(\dot{y}_{nose} < -\dot{y}_{thre} \ \& \ k_{thre} \times L_{s2n_0} < L_{s2n} < k_{DZ} \times L_{s2n_0}) \ \text{or} \ (L_{s2n} > k_{DZ} \times L_{s2n_0})$
Jumping	$y_{shoulder_M} < y_{DZ} \ \& \ \dot{y}_{nose} < -\dot{y}_{thre}$
Drinking Energy Drink	$d_{h2m} < \Delta d_{thre} \ \& \ \theta_{proj} > \alpha_{large}$
Drinking Painkiller Pills	$d_{h2m} < \Delta d_{thre} \ \& \ \theta_{proj} < \alpha_{small}$
Injecting Adrenaline	$d_{h2s} < \Delta d_{thre} \ \& \ \theta_{proj} > \alpha_{large}$

- Qualitative data: the main motions gathered from the player survey.
- Quantitative data: the angles collected through motion mapping (MediaPipe). Camtroller used these resulting calculations for its technical implementation and calibration for gesture control.

Results

Accuracy Evaluation Test

Table 2: The success rate of motions in feasibility validation test

Motion	Success Rate
Head Moving Up for jumping	100%
Head Moving Down for crouching	96.3%
Head Left Bending for left peeking	100%
Head Right Bending for right peeking	100%
Head Left Rotation for free looking left	100%
Head Right Rotation for free looking right	100%
Head Flexion for free looking down	100%
Head Extension for free looking up	100%
Drink Energy Drink	100%
Take Painkiller	96.3%
Inject Adrenaline	100%

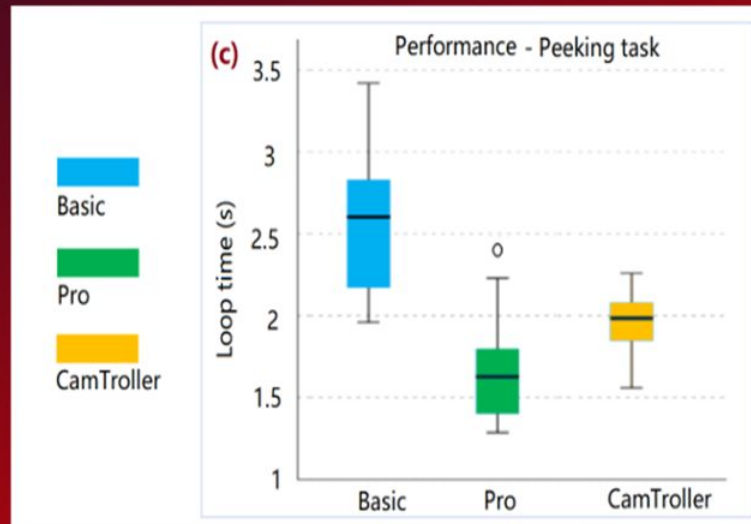
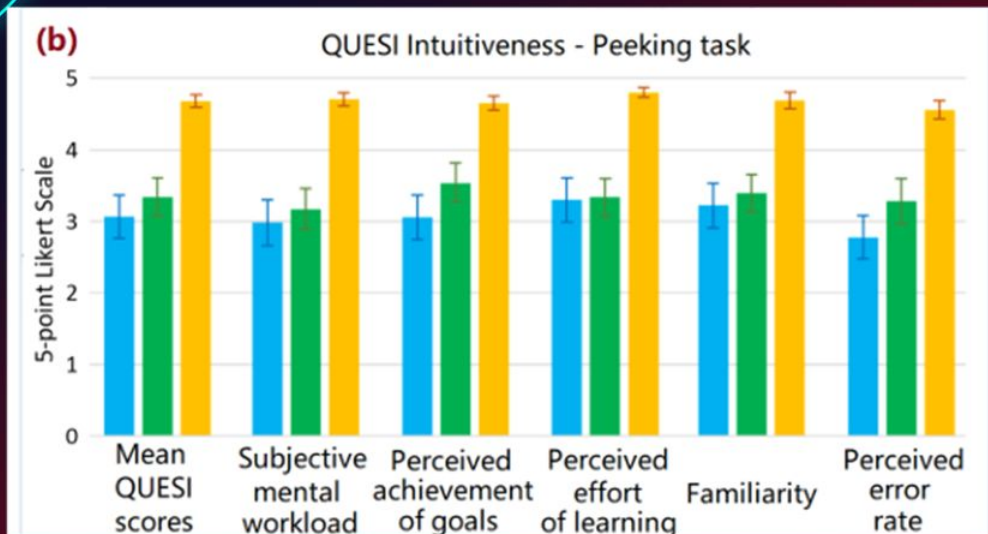
Real-time actions were reflected through the in-game avatar successfully with low margin of error/lag time. Users found their motions fluid in-game and corresponded to the trigger movements/control gestures.



Accuracy

Results

User Study data comparing each group's performance for certain tasks.



Usability & Effectiveness

- ❖ How it affects HCI
 - Enhanced Intuitiveness
 - Improved Performance
 - Generalizability

- ❖ Limitations of paper

Open-loop control

Camtroller is unable to determine whether the motion has been initiated successfully. The approach proposed by Yu-Hsin Lin et al. for detecting video game events solves this issue by monitoring the video, audio, and controller I/O.



CONCLUSION

To conclude, Camtroller expands the limits of PC gaming through NUI. It reduces the performance demand on the user.

Machine learning provides a more intuitive playstyle versus traditional PC. This concept contributes to using Computer Aided Design NUIs for everyday usage .



References

Junjian Chen, Yuqian Wang, and Yan Luximon. 2024. CamTroller: An Auxiliary Tool for Controlling Your Avatar in PC Games Using Natural Motion Mapping. In Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24), May 11–16, 2024, Honolulu, HI, USA. ACM, New York, NY, USA, 17 pages. <https://doi.org/10.1145/3613904.3642511>





THANK YOU