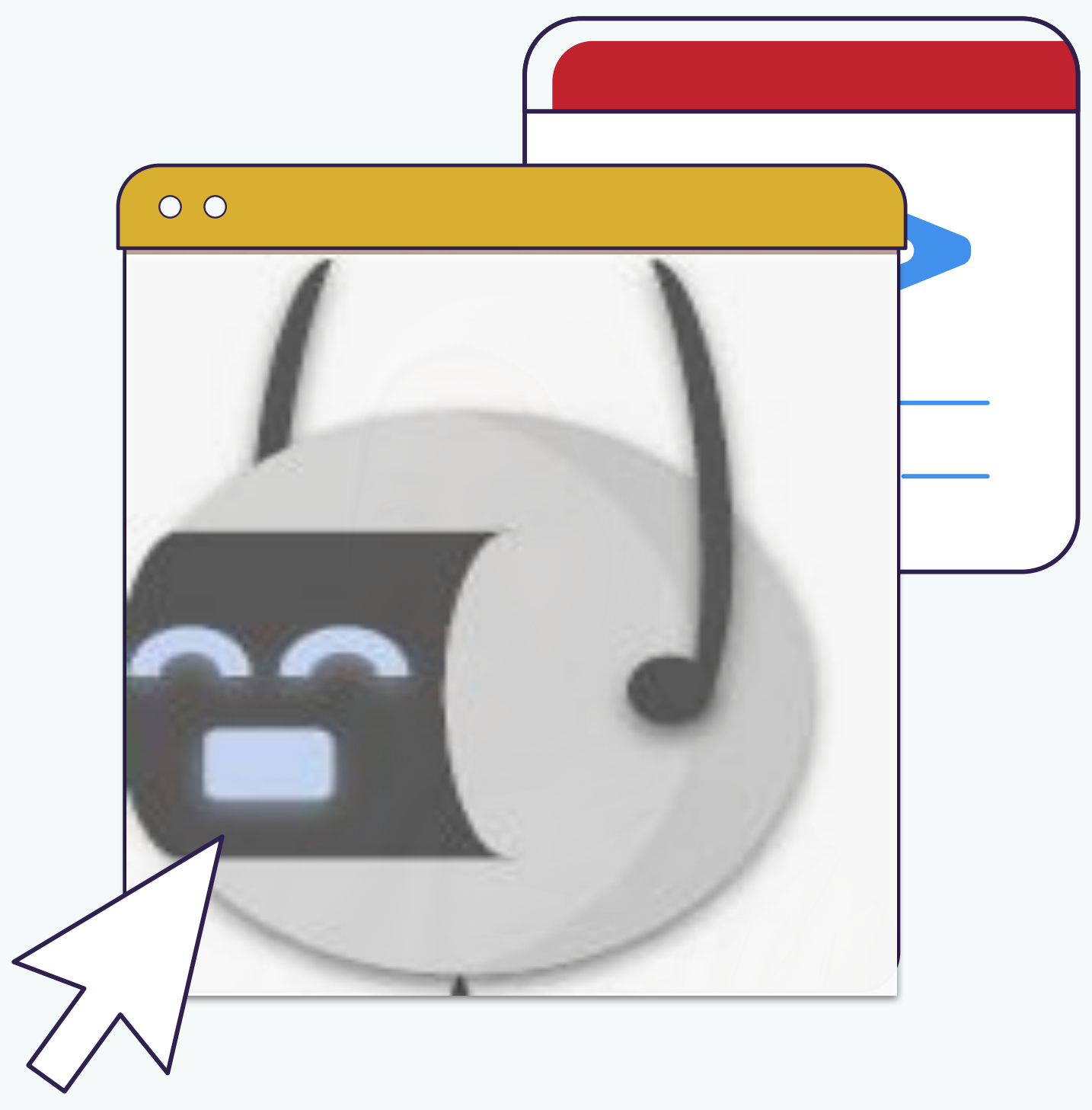


VAL: Interactive Task Learning with GPT Dialog Parsing

*Poster by Alejandro Marin **(816035363)***

Authored by Lane Lawley and Christopher J. MacLellan

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.

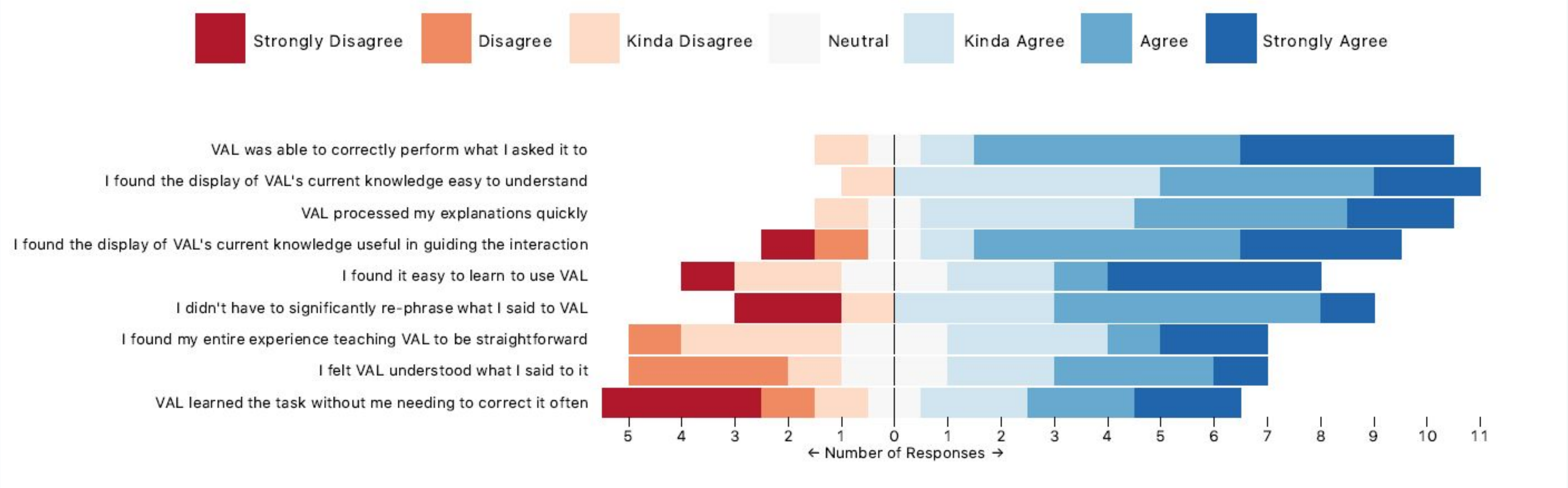
Background

- **Lane Lawley**
- A former postdoctoral researcher in Georgia Institute of Technology.
- Contributed AI related research to the field of HCI

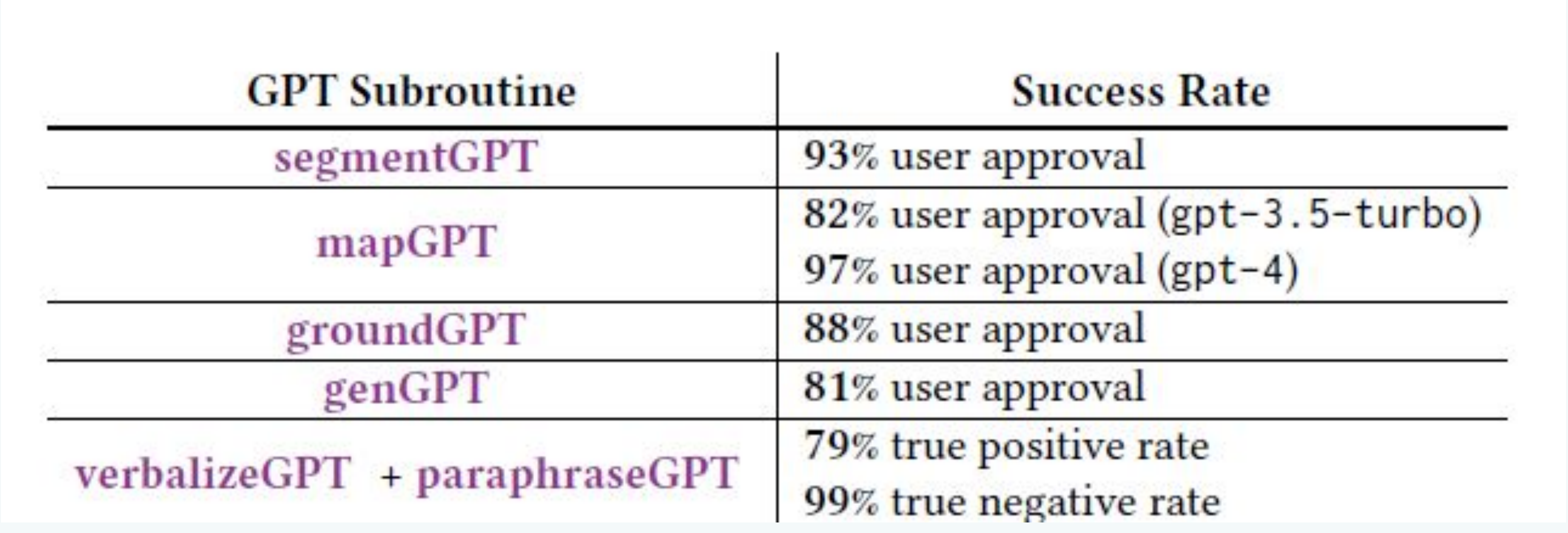
- **Christopher J. MacLellan**
- An assistant professor in the Georgia Institute of Technology.
- Contributed AI related and natural language research to the field of HCI

Results

Subjective Results: Most users thought that they could teach VAL effectively and shared positive views on the experience.



Objective Results: 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

Methodology

- **Research Approach**: Experimental design - usability study.
- **Experimental Design**: Used the Overcooked-AI game environment, where participants taught VAL to perform cooking tasks
- **Data Collection**: **Subjective -** Survey prior to experiment on ease of use. **Objective -** Success rates of GPT subroutines, use of confirmatory dialogues and performance of environment

Discussion

- **Implications for HCI:** VAL addresses key challenges in ITL by improving 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 does not deviate from the required syntax of ITL systems.
- **Limitations:** The study identified some challenges, including the limited modality of VAL, which would be expanded in future iterations. Moreover, VAL's reliance on a proprietary LLM (GPT) raises ethical concerns as it is uncertain how exactly they work.

Conclusion

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. 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.

Look Once to Hear: Target Speech Hearing with Noisy Examples

Poster by Pranav Soondar 816030105

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 specialities in machine learning, mobile technologies, speech processing and embedded systems.

Bandhav Veluri

Malek Itani

Tuochao Chen

Shyamnath Gollakota

Takuya Yoshioka

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

Mixed-Method Research Methodology

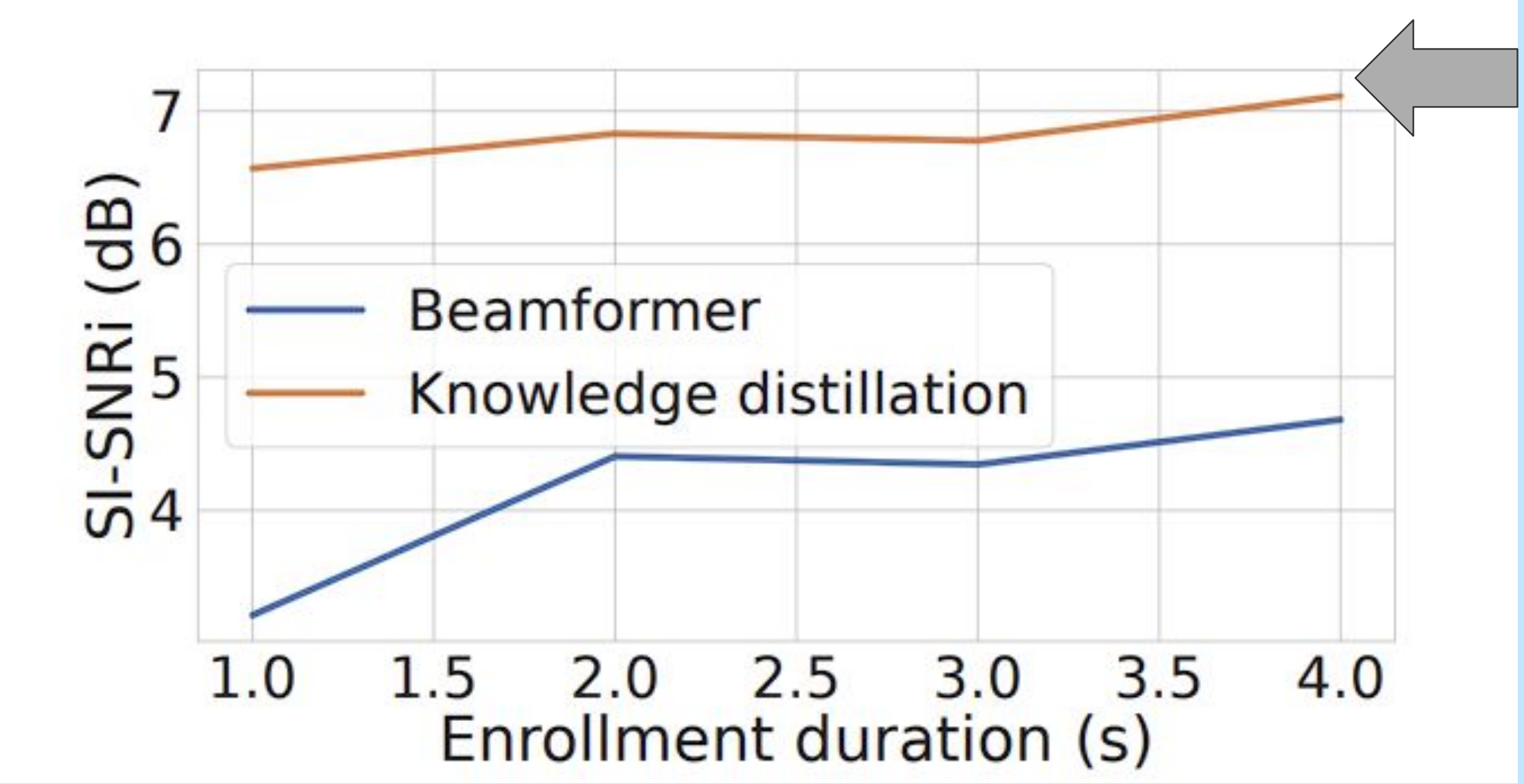
Quantitative:

- 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

Quantitative Results

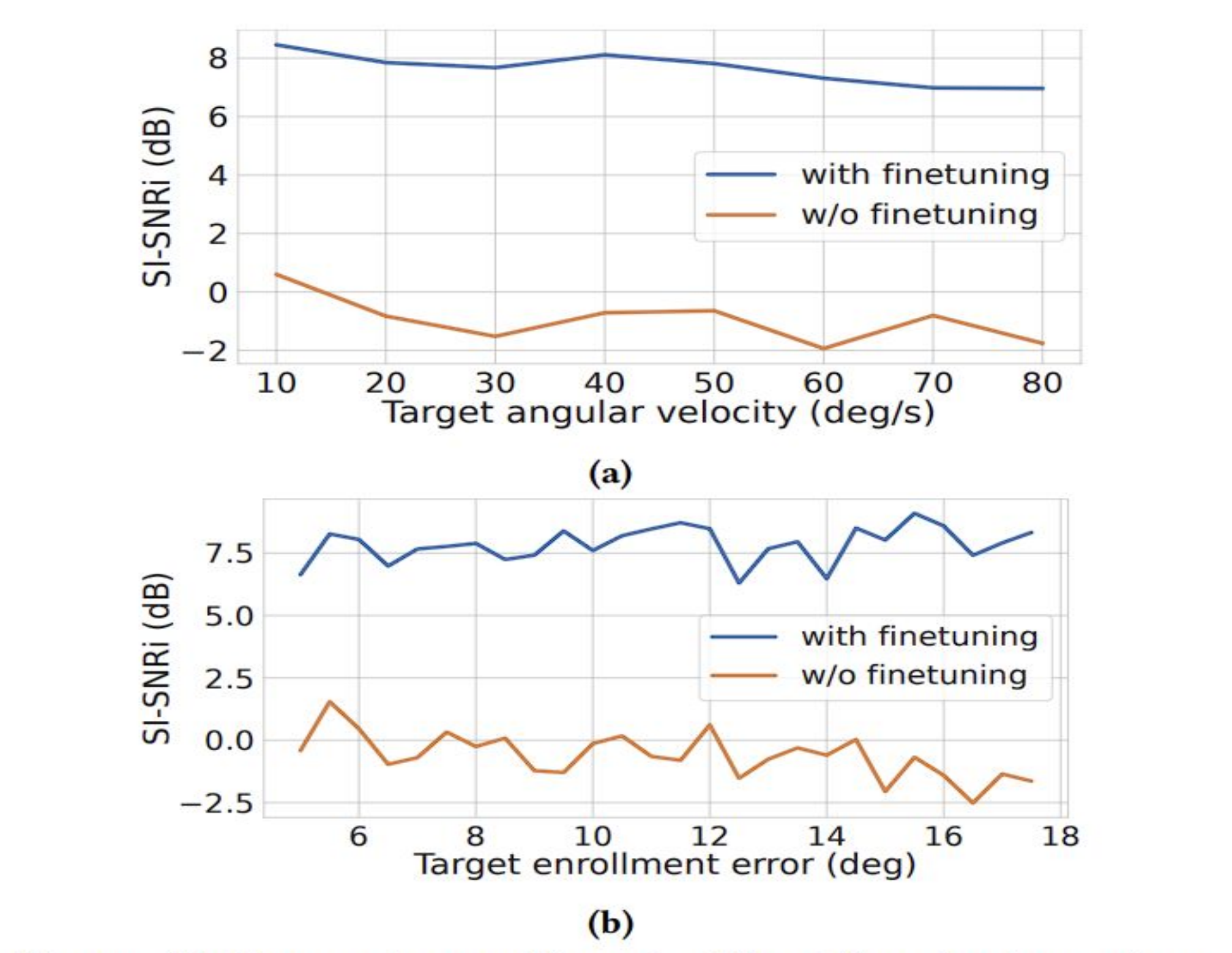


Knowledge distillation performs better even with less target speaker audio in the 5 second enrollment audio

Calculations show that knowledge distillation is superior to beam forming regardless of the neural network architecture used.

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		



(a)

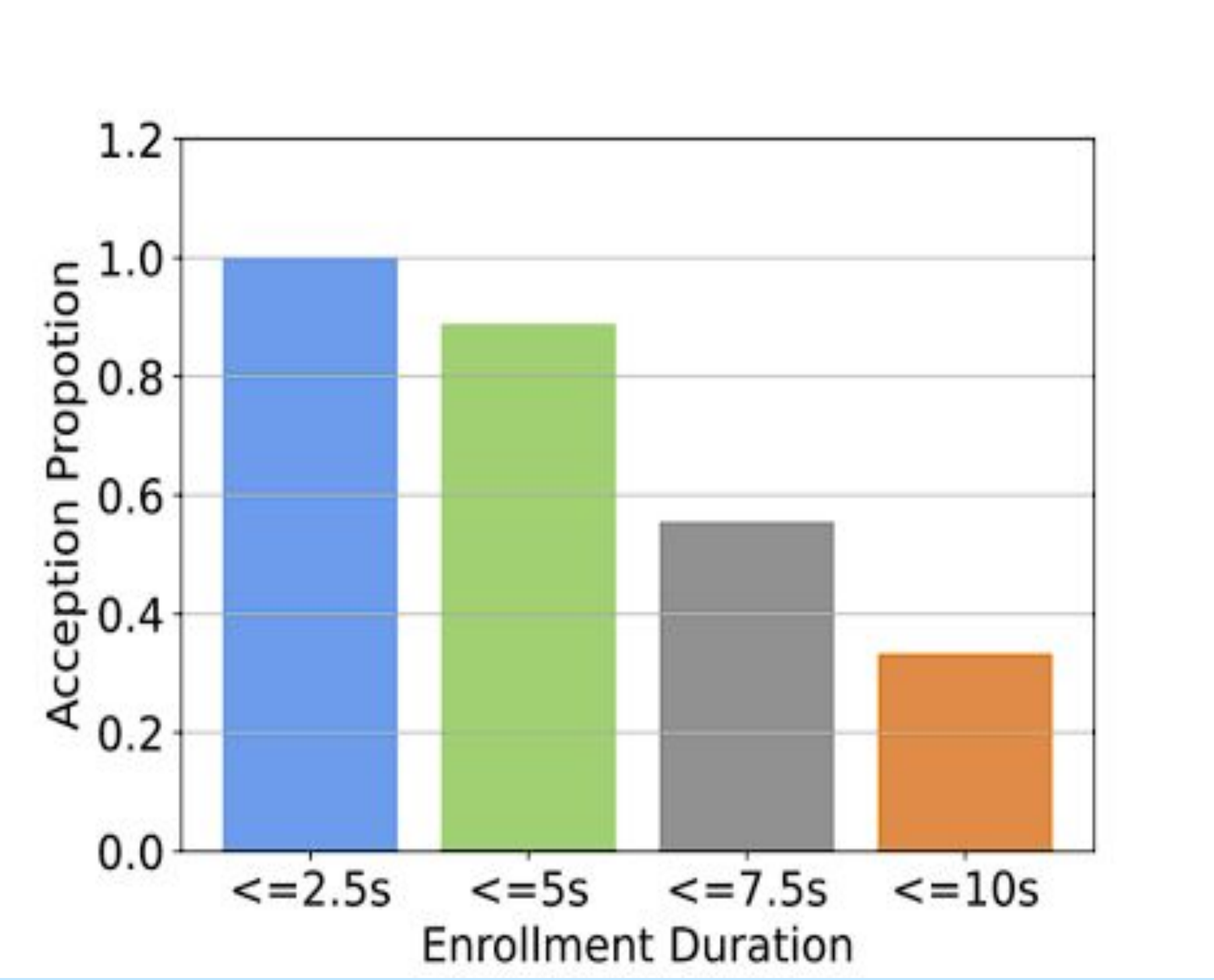
(b)

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

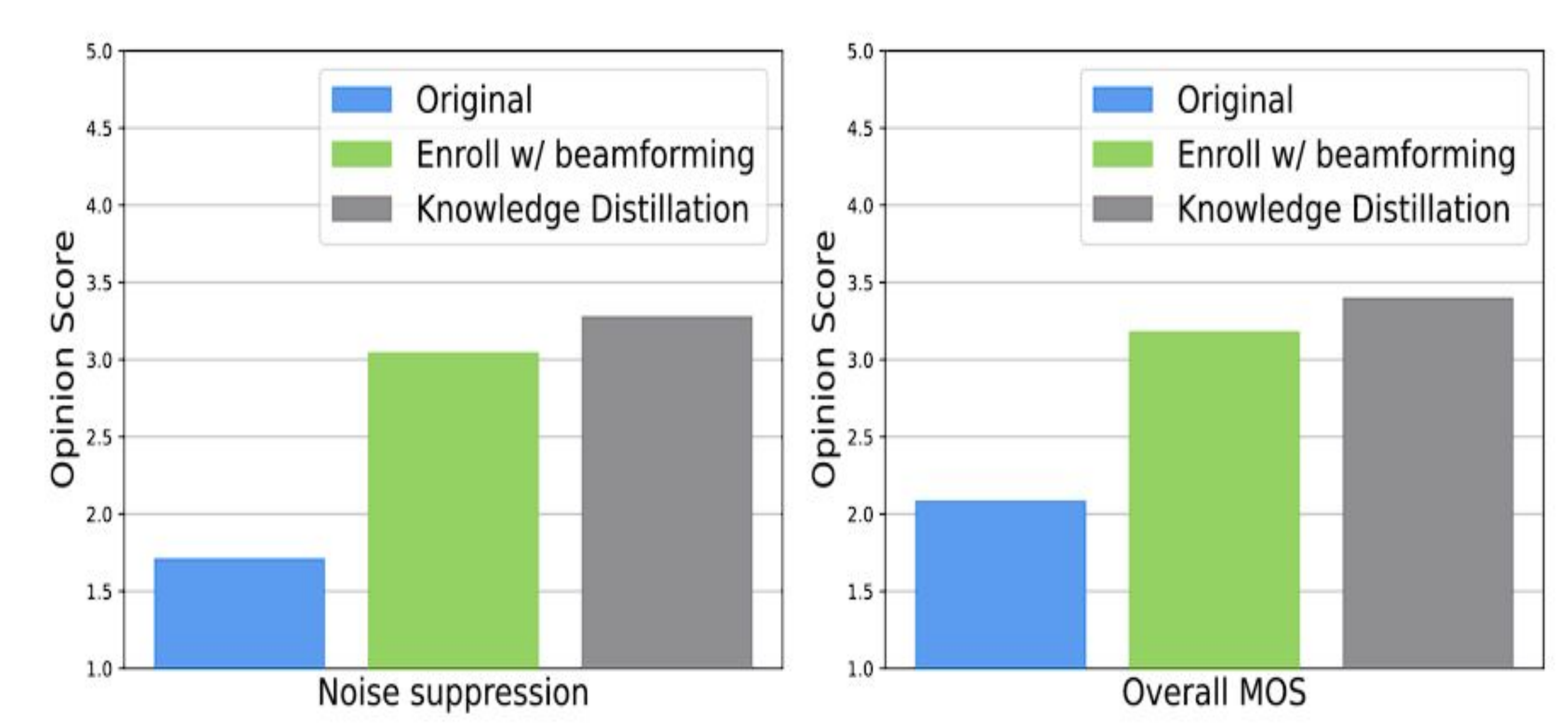
Fine tuning algorithms for wearer head movements and a moving target speaker provide significant improvement

UWI
ST. AUGUSTINE CAMPUS
TRINIDAD & TOBAGO, WEST INDIES

Qualitative Results



Users prefer lower enrollment time for the target speaker with 89% of users deeming 5 seconds to be acceptable.



Users find knowledge distillation to be most effective

Discussion

The research showed that neural networks could be successfully trained and implemented in real time environments for the purposes of smart hearing devices. Existing neural networks can be effectively utilized on embedded systems.

The implementations of these neural networks, however, did not utilize the neural processing unit (NPU)which is present on the embedded system used. No clear reason is outlined for this decision.

Users determined that the system had usability issues but performance was such that they would want to use it again. However, the sample size was relatively small for the qualitative research.

Future applications discusses in the paper indicate use cases for hearing aids, lectures, muting individuals, and improving social events for wearers.

Conclusion

Researchers presented a novel solution on an end to end hardware solution that effectively allows a wearer to look in the general direction of a speaker for 5 seconds and then look away while the system suppresses background noises and enhances the target speaker's voice.

This system is novel and one of the very firsts of its kind. The synthetic data used to train the models successfully created generalizations for indoor and outdoor environments.

This research provides a suitable starting point for future smart hearing devices and can contribute greatly to the field of HCI by improving the human auditory experience in a multitude of scenarios.

CAMTROLLER:

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

Poster by Joelle Ramchdandar 816035123

Background

Published - 11 May 2024

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

Junjian Chen
Professor Associate
5 Publications

Yuqian Wang
Professor Associate
3 Publications

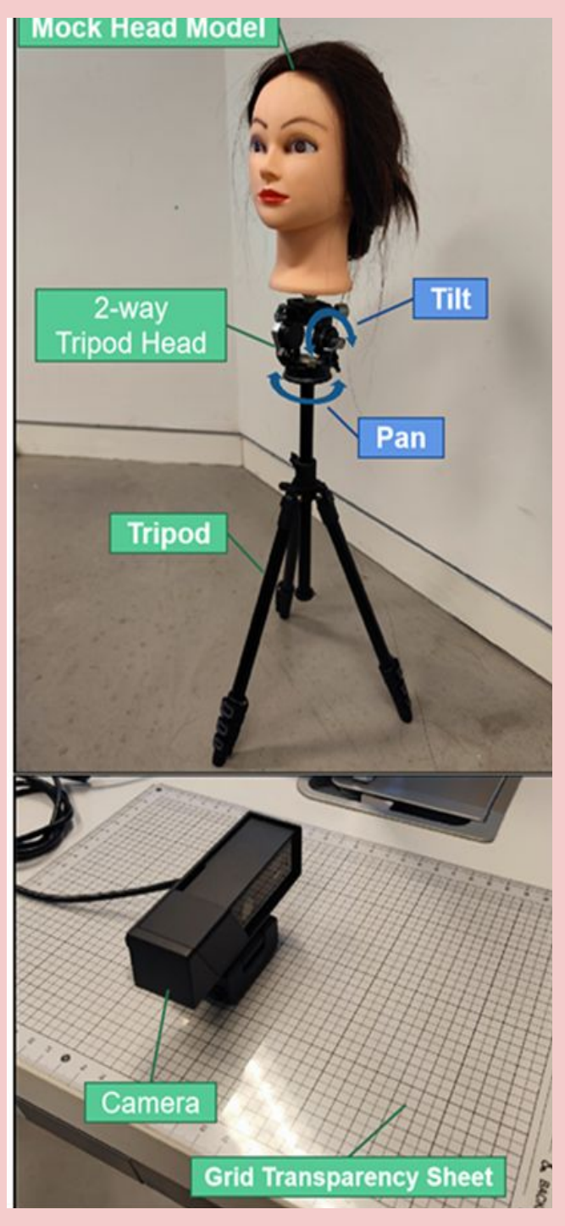
Yan Luximon
Laboratory for Artificial Intelligence in Design
169 Publications

……….Fields of study……...

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

Methodology

Research methods- To construct the auxiliary tool Camtroller, both qualitative and quantitative research methods were used. The qualitative aspect involved collecting user experiences and common avatar actions, while the quantitative aspect included statistical analysis of data sourced through natural mapping and motion tracking.



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.

Data collection techniques:

- Player Survey based on PUBG:Online 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

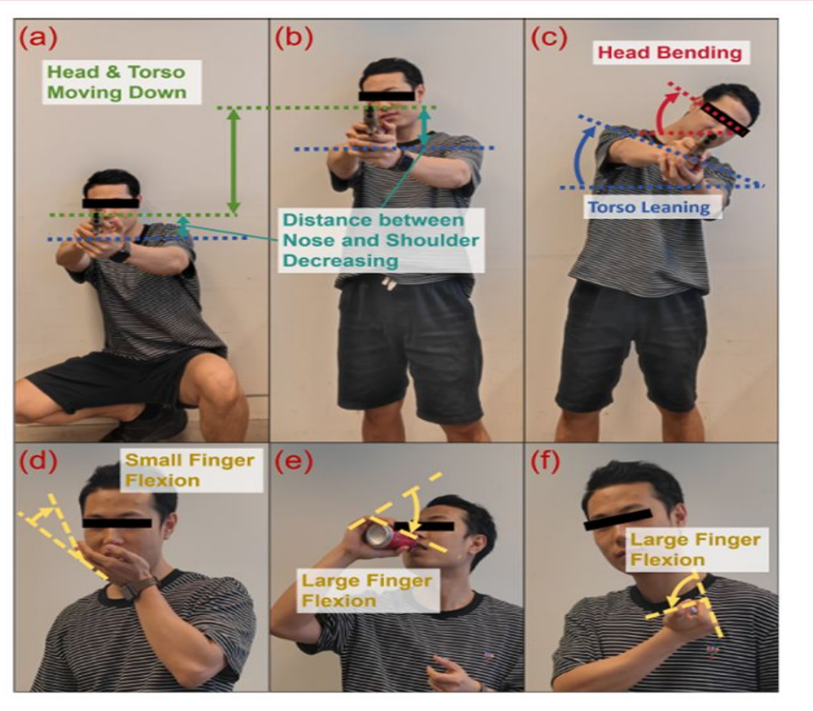


Figure 2: Avatar motions in the physical world and corresponding detectable features illustration for (a) crouching, (b) neutral position, (c) peeking, (d) taking pills, (e) taking a drink, and (f) injecting drugs.

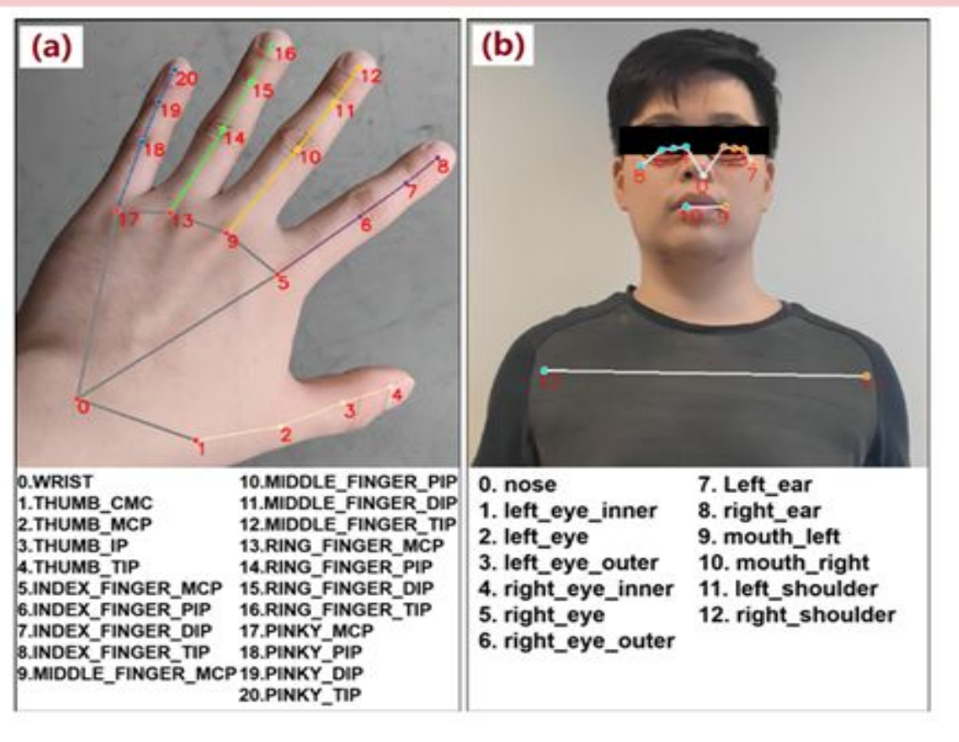


Figure 7: Landmarks of MediaPipe for (a) hand solution and (b) pose solution.

Usability study:

A study between 3 groups of users, gauging their performance and how well they adapted to using the controls during gameplay.

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.

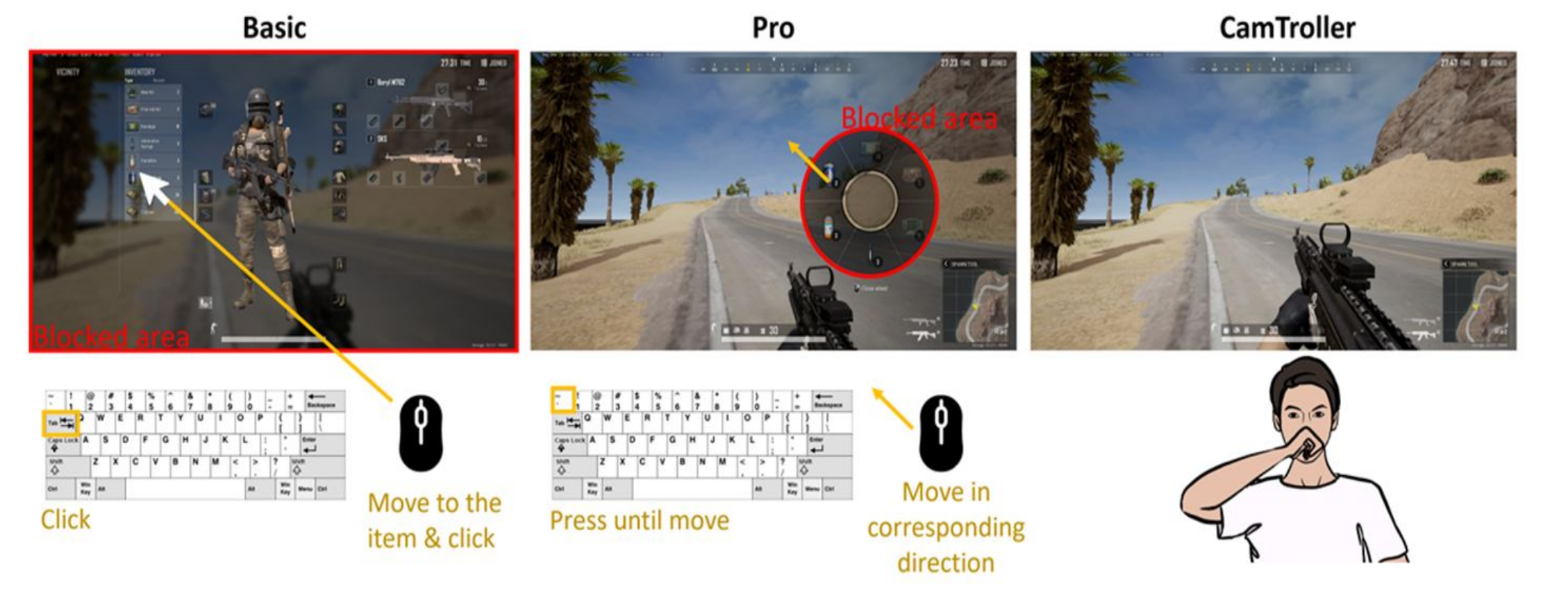


Figure 11: An illustration of boost items in three different approaches

Results

Quantitative data collected through motion tracking and mapping to create calculations. These calculations are then translated through machine learning to the in-game avatar to perform the action. Camtroller uses the processed data for its technical implementation and calibration for gesture control.

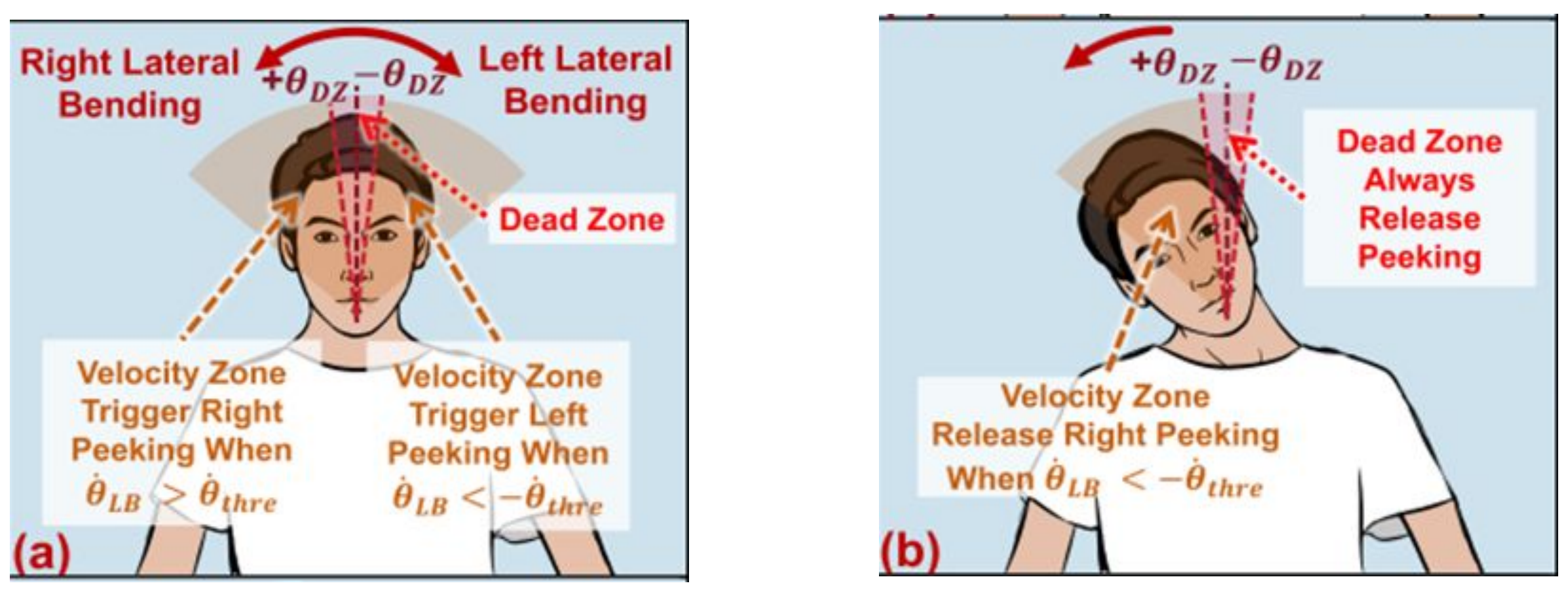
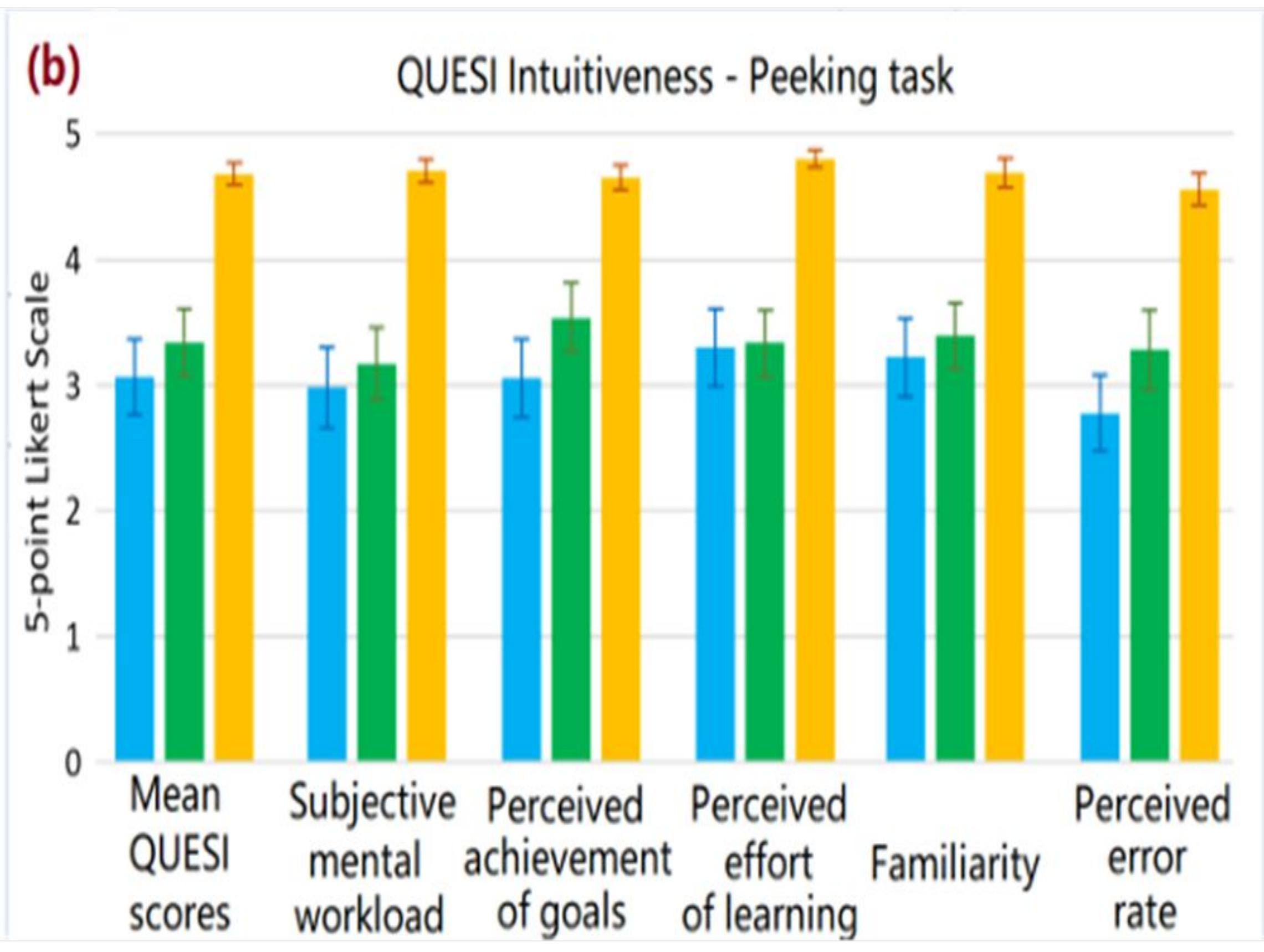
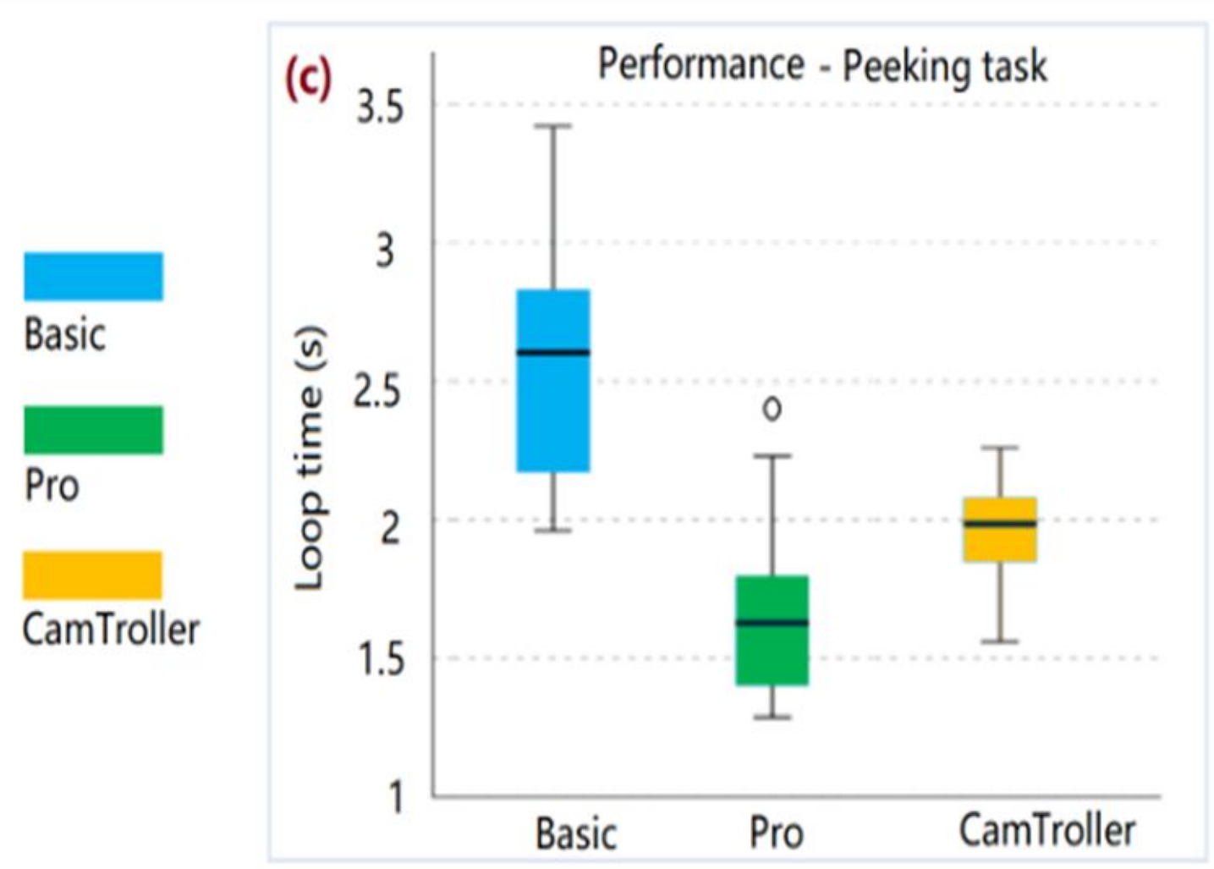

Right Lateral Bending
$+\theta_{DZ}$
$-\theta_{DZ}$
Left Lateral Bending
Dead Zone
Velocity Zone Trigger Right Peeking When $\dot{\theta}_{LB} > \dot{\theta}_{thre}$
Velocity Zone Trigger Left Peeking When $\dot{\theta}_{LB} < -\dot{\theta}_{thre}$
(a)
$+\theta_{DZ}$
$-\theta_{DZ}$
Dead Zone Always Release Peeking
Velocity Zone Release Right Peeking When $\dot{\theta}_{LB} < -\dot{\theta}_{thre}$
(b)

Table 1: Criteria for Different Motions

Motion	Criteria
Left Peeking	$\dot{\theta}_{LB} < -\dot{\theta}_{thre}$ & $\theta_{LB} < -\theta_{DZ}$
Release Left Peeking	$(\dot{\theta}_{LB} > \dot{\theta}_{thre}$ & $\theta_{LB} < -\theta_{DZ})$ or $(\theta_{LB} > -\theta_{DZ})$
Right Peeking	$\dot{\theta}_{LB} > \dot{\theta}_{thre}$ & $\theta_{LB} > \theta_{DZ}$
Release Right Peeking	$(\dot{\theta}_{LB} < -\dot{\theta}_{thre}$ & $\theta_{LB} > \theta_{DZ})$ or $(\theta_{LB} < \theta_{DZ})$
Crouching	$(\dot{y}_{nose} > \dot{y}_{thre}$ & $k_{thre} \times L_{s2n_0} < L_{s2n} < k_{DZ} \times L_{s2n_0})$ or $(L_{s2n} < k_{thre} \times L_{s2n_0})$
Release Crouching	$(\dot{y}_{nose} < -\dot{y}_{thre}$ & $k_{thre} \times L_{s2n_0} < L_{s2n} < k_{DZ} \times L_{s2n_0})$ 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	$d_{h2s} < \Delta d_{thre}$ & $\theta_{proj} > \alpha_{large}$



User study data comparing each group's performance for certain tasks. Camtroller performed significantly better than non-professionals (Basic) without the tool to aid them and close to pro, proving its ease of use for controls.



★★★★★★★ Usability

Summary table from the accuracy evaluation test: Real- time actions were reflected through in-game avatar successfully with low margin of error/lag time. Users found their motions fluid in-game and corresponded to the trigger movements/ common gestures.

★★★★★★★ Accuracy

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%

Discussion

How it affects HCI

•**Enhanced Intuitiveness**: CamTroller's natural motion mapping reduces the mental workload on the user, making interactions more familiar and error-free.

•**Improved Performance**: In complex gaming scenarios, CamTroller enhances player performance by easing cognitive load.

•**Generalizability**: CamTroller (NUI) concept can be applied to various one-to-one avatar mapping PC games, enhancing the gaming experience across different genres.

Limitations of paper

While CamTroller has demonstrated its potential to assist players in a more efficient and intuitive game experience, it is unable to determine whether the motion is being initiated correctly or not. 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 traditional PC gameplay through natural mapping techniques. Popular Esports platforms such as PUBG becomes more user friendly for players and lowers the learning curve, especially for beginners. A webcam, coupled with machine learning algorithms accurately translates in-game avatar movements demonstrating a more intuitive playstyle versus the previous gameplay of memorizing key binds. This concept is a significant step towards Computer Aided Designs and can be extended to other areas.

UWI
ST. AUGUSTINE CAMPUS
TRINIDAD & TOBAGO, WEST INDIES