Design and Wizard-of-Oz Evaluation of a Tri-Modal Contactless In-Vehicle HMI for High-Noise and High-Vibration Cabins

Zhenghao Yu
Department of Eniegering, Faculty of Electronic and Electrical Engineering, The University of Hong Kong, Hong Kong, China
**Corresponding author.*

Abstract: Harsh cabins in mining, construction, and defense degrade touch human–machine inter- faces for secondary tasks, while single-modality contactless input stays fragile under coupled noise and vibration. A sequenced tri-modal protocol with dwell locking, head confirma- tion or cancellation, and optional voice refinement is specified and evaluated in a desktop Wizard-of-Oz browser under parameterized disturbance. An operator survey (N = 48) fixed timing constants; N = 30 participants compared touch (A), pooled single-modality contact- less (B), and tri-modal (C) under a fixed Hard preset. Holm–Bonferroni contrasts show C reduced error rate and NASA–TLX versus B (p <.01), shortened completion time versus B (p <.05), and improved SUS versus B, while A remained fastest with the lowest descriptive errors. Conclusions are interaction-level evidence under simulation, not ISO-calibrated cabin equivalence or production gaze or speech benchmarks.

Keywords: Multimodal Interaction; In-Vehicle HMI; Special-Purpose Vehicles; High-Noise Strong-Vibration; Gaze Dwell; Head Gesture; Voice Refinement; Wizard-of-Oz.

1. Introduction

1.1 Motivation and Research gaps

Mining, construction, and defense special-purpose vehicles expose operators to sustained 85– 100 dB occupational-scale cabin noise, whole-body vibration correlates, and persistent hand occu- pation with joysticks or levers [1], conditions under which touch-based HMIs for secondary tasks (climate, navigation, work parameters) become slow, error-prone, and safety-sensitive whenever visual and manual attention is diverted from primary operations [1]. Single-modality contactless alternatives are therefore attractive in principle, yet they remain inherently fragile in these cab- ins because gaze tracking degrades under motion, ASR degrades in high noise, and head-gesture discrimination degrades under vibration [2–5]. Against that background, the present work is motivated by three gaps in the literature:

(i) Few automotive multimodal designs publish a complete closed-loop temporal protocol (dwell duration, timeout cancellation, confirmation rules, and safety constraints) for harsh special-purpose cabins [6–8].

(ii) Empirical HCI under coupled high-noise and strong-vibration disturbance remains scarce relative to passenger-car scenarios [9].

(iii) Many studies conflate interaction-level user performance with engineering-level sensor bench- marks, which invites over-claiming and weak replication [10].

The remainder of this section states how those gaps are operationalised as research questions, contributions, and manuscript structure.

1.2 Objectives, Contributions, and Paper Outline

The empirical programme addresses the following questions:

(RQ1) Under coupled high-noise and strong-vibration simulator stress, how large are the gaps between touch (A) and a pooled single-modality contactless baseline (B) on error rate, workload, completion time, and perceived usability?

(RQ2) Relative to B, does a gaze–head–voice tri-modal temporal protocol with explicit dwell, timeout, and head confirmation (C) reduce errors and workload while keeping completion times acceptable?

(RQ3) Under the same disturbance preset, does C improve SUS relative to B, and how do all three policies relate to the conventional 68-point SUS reference [11]?

To keep those questions answerable without over-claiming, claims are intentionally bounded to interaction procedure and laboratory behaviour under a desktop Wizard-of-Oz simulator, leaving ISO-calibrated whole-body vibration, IEC-grade acoustics, production-grade gaze and ASR latency, and vocational operator samples outside the present evidence pass.

Within that scope, the paper contributes three replicable, interaction-focused items:

(1) A standardised tri-modal contactless protocol (3 s dwell, 5 s timeout, head confirm/cancel, voice refinement, safety-critical dual confirmation) aimed at unintended activation.

(2) Wizard-of-Oz empirical evidence under severe simulator disturbance that tri-modal policy C outperforms pooled single-modality contactless B on error rate, NASA Task Load Index (NASA–TLX) workload, completion time, and System Usability Scale (SUS) means.

(3) A parameterised web-based WOz prototype that decouples interaction logic from sensor performance, with exportable event logs and analysis scripts suitable for independent replication subject to privacy and institutional policy.

The tri-modal paradigm is evaluated end-to-end in the WOz experiment reported below, and the remainder of the manuscript is organised as follows: Section 2 situates non-contact in-vehicle interaction, multimodal sequencing, and WOz under stress; Section 3 formalises the protocol, documents the prototype and disturbance model, and specifies the controlled study; Sections 4– 5 present quantitative outcomes with interpretation; and Section 6 summarises contributions together with deployment-facing implications.

2. Related Work

2.1 Non-contact In-Vehicle Interaction

Gaze, head pose, and speech are now mainstream non-contact channels for in-vehicle secondary control [2,12,13], and dwell-based gaze with explicit confirmation in particular mitigates Midas- touch unintended activation [14, 15], while complementary automotive research combines gaze with supplementary manual input on HUD surfaces [16]. Much of this literature nevertheless centres on passenger-centric prototypes, so comparatively little work isolates *coupled* noise and vibration stressors in the special-purpose cabins that motivate the present study.

2.2 Multimodal Fusion and Late Sequencing

Turning from isolated channels to coordinated input, late, rule-governed multimodal sequencing reduces ambiguity relative to unconstrained early fusion [6], and representative in-car prototypes accordingly fuse speech, gaze, and micro-gestures for secondary tasks [7, 8]. Even so, published sequences seldom document a complete closed-loop policy (dwell, timeout, confirm, safety) along- side empirical behaviour under non-stationary, cabin-like disturbance, which limits direct transfer to harsh operating envelopes.

2.3 WOz Under Adverse Cabins

Because full sensing stacks are often unavailable during early cabin-HMI design, Wizard-of-Oz (WOz) methods place a human wizard in the loop to emulate unstable sensing and thereby support interaction validation before complete engineering integration [10], and automotive HCI increasingly applies WOz in lab and field studies [9], including remote prototyping [17], automated-vehicle field mediation [18], and language-technology WOz guidance [19]. Relatively few platforms nevertheless couple WOz with graded, non-stationary audio–visual–mechanical disturbance aimed at special-purpose cabins, which motivates combining a tri-modal policy with a parameterised perturbation console here as a first interaction-layer evidence pass [1].

In summary, prior work supports individual non-contact channels and multimodal combinations, yet three gaps recur for harsh cabins: published closed-loop temporal rules are often incomplete; empirical stressors are frequently decoupled or passenger-centric; and sensing-layer performance is easily conflated with interaction-layer outcomes. The present study targets those gaps through a protocol-first WOz evaluation that foregrounds replicable timing semantics.

3. Methods

Turning from positioning to procedure, the empirical work followed a sequential mixed-methods pipeline: survey-informed protocol parameters, a frozen implementation in a web Wizard-of- Oz console, and a within-subjects experiment (N = 30) under a fixed simulator preset. Sub- sections 3.1–3.5 follow that order so dependencies read forward-only (timing constants before behavioural testing). The browser console, analysis scripts, and

anonymised session logs can be shared for replication on reasonable request and subject to ethics clearance.

3.1 Operator survey (N = 48)

The pipeline begins with priors from operators rather than from arbitrary constants: forty-eight operators of special-purpose vehicles completed a structured questionnaire whose descriptive summaries were used only to set engineering constants for the prototype rather than to test hypotheses. Items covered perceived touchscreen difficulty under vibration, acceptable dwell durations for contactless locking, preferred non-contact modality bundles, tolerance for delayed or failed inputs, and qualitative safety concerns about unintended activation. Vibration-related touchscreen errors dominated complaint scores, the modal acceptable dwell band was 2–3 s, and a gaze–head–voice combination was the most endorsed non-contact bundle, which jointly motivated a 3 s dwell lock, a 5 s post-lock timeout to *Idle* (logged error), and sub-1 s head-response targets where hardware allowed.

3.2 Tri-modal Contactless Protocol

Those timing commitments were then frozen into the tri-modal policy described here. Formally, the interaction is modelled as a finite transition system $M = (S, s_0, \mathrm{A}, T, F)$. The state set is $S = \{Idle, Dwelling, Locked\}$, with a transient *Execute* phase for committed commands. The initial state is $s_0 = Idle$, and the action alphabet A comprises dwell ticks, a 5 s timeout, nod, chin-down, lateral module switches, and wizard-mediated voice submits. The transition function T encodes the protocol (3 s dwell to *Locked* ; timeout to *Idle* with logged error; nod advances to *Execute* and thereafter to *Idle*). Accepting quiescence F marks safe return to *Idle* after execute or cancel, so the tuple fixes semantics for replication independent of visual skinning.

Operationally, the same semantics are realised in the reference console as follows: steady cursor dwell (3 s) serves as a gaze proxy that locks a control until true eye tracking is integrated. Head nod commits, chin-down cancels, and lateral motion (or keys) cycles modules under a 900 ms refractory that limits duplicate triggers in line with the artefact cooldown constant. Typed utterances are mapped by a wizard from a closed codebook for voice refinement. The same stochastic latency and failure rules as other inputs emulate unreliable ASR. Channels are fused *late* under explicit rules rather than in parallel at the signal level, so spatial commitment precedes linguistic refinement [6]. Navigation-type deactivation requires dwell plus confirm with bypass disabled, consistent with the safety-critical dual-confirmation rule used elsewhere in the protocol. Figure 1 summarises the state machine: states advance from *Idle* through dwell (*Dwelling*) and *Locked* to command execution or cancellation, then return to *Idle*; gaze is simulated with steady cursor dwell until a 3 s lock; solid connectors denote mandatory transitions and dashed connectors denote optional wizard-mediated linguistic refinement after locking; timeout and chin- down cancellation return to *Idle* (timeout logged as a protocol error). Table 1 lists operator- visible sequencing rules.

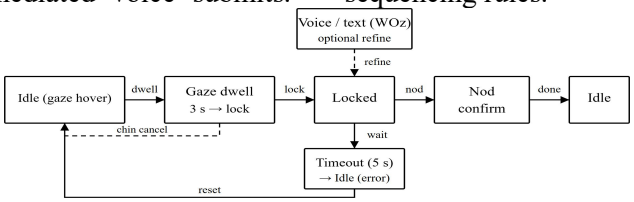


Figure 1. Tri-Modal Sequential Interaction State Machine (WOz Console).

Table 1. Tri-Modal policy (C): Operator-Visible Sequencing Rules.

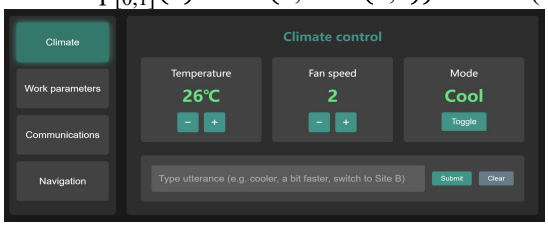
Stage	Trigger	Outcome / logging
Target lock	Steady cursor dwell (3 s) on a control	Highlight lock;enterLocked
Confirm	Head nod surrogate (default:keyboard)	Execute committed action;return toIdle
Cancel	Chin-down surrogate	Clear lock without execution
Timeout	No confirm within 5 s after lock	Return toIdle;protocol error counted
Voice refine	Wizard maps typed codebook utterance after lock	Same stochastic latency/failure draw as other inputs
Module browse	Lateral head-turn surrogate between modules	900 ms refractory between successive triggers

3.3 WOz Web Console and Disturbance Model

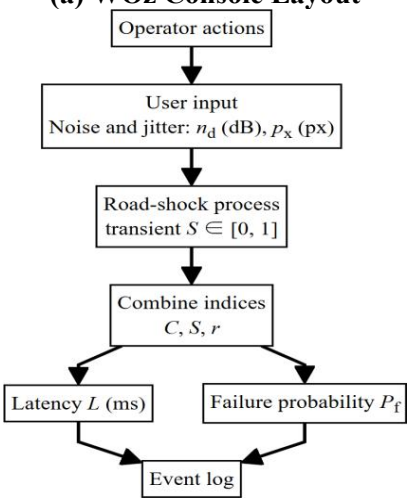
Figure 2 sketches layout and coupling for a browser implementation (HTML/CSS/JavaScript) intended for desktop use. Figure 2a shows the operator-facing console: secondary-task modules, dwell targets, and disturbance controls. Figure 2b depicts the disturbance–reliability coupling in which non-stationary perturbation drives stochastic latency and failure draws (formalised in this subsection). A mouse supplies dwell and touch emulation; keys supply head surrogates. Optional webcam head pose uses MediaPipe Face Landmarker with smoothed thresholds (implementation detail omitted here). On-screen "dB" values, pixel-scale viewport jitter, and "road-shock" spikes are labelled only as perceptual correlates of harsh cabins. They are not ISO 2631 whole-body vibration, IEC acoustics, or NIOSH noise dosimetry.

Given that layout, the remainder of this subsection specifies how slider-level stress maps into stochastic latency and failure draws on each input. The baseline combined disturbance index fed into the latency and failure models clamps an unbounded weighted sum onto [0, 1]. The clamping map and combined index are defined as:

$$\mathrm{clip}_{[0,1]}(x) = \min(1, \max(0, x)) \quad (1)$$



(a) WOz Console Layout



(b) Disturbance–Reliability Pipeline

Figure 2. WOz Experimental Platform

$$C = \mathrm{clip}_{[0,1]}\left(0.45\left(\frac{n_d - 70}{40}\right)^{1.35} + 0.55\left(\frac{p_x}{24}\right)^{1.7}\right) \quad (2)$$

where clip is given by (1), nd is the display-scale noise label in dB and px the jitter amplitude in pixels read from the simulator UI sliders, the weights 0.45 and 0.55 combine noise- and jitter-derived partial indices before clamping, and the exponents 1.35 and 1.7 impose concave gains on each channel as implemented in the reference browser artefact.

Graded presets (*Standard*, *Hard*, *Extreme*) combine looped industrial audio, periodic viewport translation, and stochastic shocks that raise transient severity $S \in [0, 1]$. Each accepted operator input triggers stochastic response latency L (ms) and failure probability P_f according to

$$L = 110 + C^{1.15} \times 540 + r + S \times 280 \quad (3)$$

$$P_f = \min(0.58,\ 0.015 + C^{1.25} \times 0.24 + S \times 0.16) \quad (4)$$

where C is the combined index from (2), $r \sim U(0, 120)$ ms injects per-input micro-jitter, $S \in [0, 1]$ encodes transient "road shock" severity from the non-stationary perturbation loop, Equation (3) maps (C, r, S) to realised latency L in milliseconds, Equation (4) yields the capped per-input fail- ure probability P_f prior to each Bernoulli draw, and the outer $\min(\cdot, 0.58)$ prevents unrealistically high failure rates under extreme shocks.

Failed draws flag trials, increment error tallies, and are logged with timestamps and active disturbance parameters.

3.4 Prototype Implementation and Instrumentation

The preceding subsection defines the disturbance–reliability mapping; the present subsection summarises how that mapping was instantiated and logged in the laboratory artefact. The reference console is a single-page HTML application executed locally in the laboratory. Pointer
dwell on controls uses a mouse cursor as a gaze proxy; the 3 s dwell and 5 s post-lock timeout trace to the operator survey (Section 3.1). Head nod, chin-down, and lateral turns are bound to dedicated cursor keys on the keyboard (up, down, left, and right). Optional head pose uses the MediaPipe Face Landmarker (reduced-precision model) with exponential smoothing on nose–eye geometry proxies. Analysed laboratory sessions used keyboard head surrogates by default so that head timing was consistent across participants; webcam tracking remained an

optional engineering path in the artefact. Non-stationary "road shock" intervals modulate translation gain, oscillation period, and audio gain according to preset-dependent hazard, supplying the shock factor S that feeds Equations (3)–(4). Climate, work-parameter, communications, and navigation modules reuse the same widgets across policies so that task isomorphism holds.

Each session exports UTF-8 logs with timestamped operator actions, success flags, active preset parameters, realised delay, failure draws, module identifiers, and running operational and proto- col error counts (browser event clock; laboratory wall-clock alignment noted per session). Exports bundle scalar summaries with a chronological operLog array for offline parsing and merge with anonymised participant identifiers and wizard-side timestamps where applicable. Wizards followed a single codebook, applied Equations (3)–(4) without post-hoc tuning, and did not coach beyond scripted training. Wizard timestamps were merged into the participant log stream. The analysed sample comprises $N = 30$ participants who completed policies A, B, and C under the nominal *Hard* preset without withdrawal.

3.5 Controlled Experiment

The within-subjects design compared three modality policies under the protocol and logging conventions above. Policy A (touch) enabled direct pointing with dwell and head routing disabled. Policy B (single-modality contactless) exposed participants to counterbalanced blocks in which only one contactless channel was available at a time (voice-only alternating with dwell- only); trials were pooled at the participant × policy level to estimate a pragmatic single-modality contactless baseline under session-length constraints. Policy B is therefore a *pooled* baseline rather than a factorial decomposition of isolated channels: it answers whether sequenced tri- modal interaction improves on "best-effort single channel at a time" under the same disturbance, without claiming channel-level omnibus effects. Policy C (tri-modal) enforced the full sequence of dwell locking, head confirmation or cancellation, and optional voice refinement, with direct click-through disabled.

Task tier (basic versus advanced) was crossed with policy in a linear mixed-effects model, using eight isomorphic tasks (Table 2) that mixed discrete controls with colloquial phrasing; *NL* denotes a colloquial prompt entered in the Wizard-of-Oz voice field and mapped by the wizard from a closed codebook (Section 3.2). Each trial began at task prompt onset; *completion time* ran from prompt onset until the first successful attainment of the goal state, with unsuccessful trials (including timeout cancellations where applicable) excluded from completion-time summaries, whereas *error rate* tallied unsuccessful trials, including timeouts where applicable, relative to all presented trials within each policy block.

To keep disturbance comparable across policies and sessions, internal validity relied on holding the preset fixed at *Hard* throughout all analysed sessions. Participants were instructed not to change preset sliders during blocks, and the experimenter verified the logged preset and slider fields at block boundaries. The prototype UI still exposes mute and reset controls for engineering demonstrations, but those controls were disabled or left unused during analysed trials, so any unintended change would appear as a logged preset transition rather than as silent drift.

Table 2. Task Battery (T1–T8)

ID	Mod.	Goal (summary)	Tier
T1	Clim.	Temp. 26 °C	Basic
T2	Clim.	Fan = 2	Basic
T3	Work	Speed 60%	Basic
T4	Work	Mode Auto	Basic
T5	Clim.	NL "cooler", then Cool on	Adv.
T6	Work	NL "faster", then verify %	Adv.
T7	Comm.	Message "work normal" to Dispatch	Adv.
T8	Nav.	Nav. off, then dest. Site B	Adv.

With the task battery fixed as above, thirty adults (18–55 years; self-reported normal or corrected-to-normal vision and hearing) were recruited from the university community. They provided written informed consent, received monetary compensation, and could withdraw without penalty. Identifiers were pseudonymised in logs; optional webcam video was processed locally and was not retained as recordings. Participants were not vocationally certified equipment operators, so con- clusions concern laboratory behaviour under a browser simulator and motivate future field studies with professional crews. Procedures complied with The University of Hong Kong human-subjects requirements; formal ethics identifiers are available outside the anonymised submission pipeline on request. With three repeated policies per participant, $N =$

30 meets conventional sensitivity for medium main effects at $\alpha = 0.05$ under the usual $1 - \beta \approx 0.80$ target for repeated-measures designs.

For the fitted models, the independent variable was modality policy (levels A, B, C). Dependent variables were completion time, error rate, NASA Task Load Index (NASA–TLX) global score (unweighted mean of six 0–100 subscales) [20], and System Usability Scale (SUS) [11], reported alongside prior multimodal in-cabin usability practice [21]. The frozen *Hard* UI defaults in the reference artefact are a 95 dB display-scale noise label, 6 px nominal jitter amplitude, and a

0.055 s viewport oscillation period before non-stationary shock modulation; exact realised draws are logged per trial.

Sessions lasted approximately 100 minutes. They included brief demographics, roughly 10 min- utes of training per policy, practice to criterion, Latin-square task order within each policy block, NASA–TLX after each policy block, SUS after each policy block (same questionnaire items, pol- icy label on the form), and debrief. Block order (A/B/C) was counterbalanced; the disturbance preset was fixed to *Hard*.

Inference relied on linear mixed-effects models (R package lme4) with participant random intercepts, fixed effects of modality policy, task tier, and their interaction, and REML variance estimation. Estimated marginal means (EMMs) were obtained using emmeans. Holm–Bonferroni adjustment controlled family-wise error across the three pre-registered pairwise policy contrasts on each endpoint (two-tailed $\alpha = 0.05$), using the step-down procedure shipped with emmeans. For completion time, NASA–TLX, and SUS, omnibus tests are reported together with adjusted pairwise contrasts and descriptive means ± SD; for error rate, model-based means are reported with the same contrast structure (binomial mixed models are a natural sensitivity analysis for trial-level binary outcomes and can be reported when trial logs are aggregated accordingly).

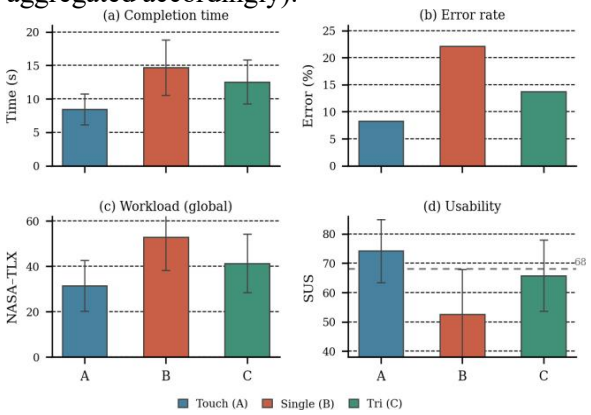


Figure 3. Within-Subjects Outcomes by Modality Policy (*Hard*)

4. Results

Building on the design in Section 3, Figure 3 plots mixed-model estimated marginal means for policies A/B/C under the fixed *Hard* preset described in Section 3.3. The four panels share an A/B/C axis (touch, pooled single-modality contactless, tri-modal): *(a)* completion time in seconds (mean ± SD); *(b)* error rate in percent (model means; SD not plotted); *(c)* NASA–TLX global score on 0–100 (mean ± SD); *(d)* SUS (mean ± SD) with the conventional 68-point norm line [11]. Holm–Bonferroni *p*-values for policy contrasts are reported in this section rather than encoded graphically in the figure. Table 3 summarises descriptive means ± SD ($N = 30$, fixed *Hard* preset) alongside the inferential contrasts discussed below; standard deviations are omitted for model-based error-rate means as in Figure 3*(b)*. The fitted models additionally included policy

× task-tier interactions as fixed effects, but the present narrative foregrounds policy contrasts because they directly answer RQ1–RQ3.

Table 3. Descriptive Outcomes by Modality Policy.

Policy	Time (s)	Error (%)	NASA–TLX	SUS
A touch	8.42 ± 2.31	8.3	31.4 ± 11.2	74.2 ± 10.8
B pooled single-modality	14.67 ± 4.15	22.1	52.8 ± 14.6	52.6 ± 15.3
C tri-modal	12.53 ± 3.28	13.7	41.3 ± 12.9	65.8 ± 12.1

For completion time, touch (A) yielded the shortest mean, followed by tri-modal C and pooled single-modality B (Table 3), and a significant main effect of policy was obtained ($F(2, 58) = 18.74$, $p < .001$). Holm–Bonferroni pairwise contrasts indicated that A was faster than B and C ($p < .01$) and that C was faster than B ($p < .05$), corresponding to a 15% lower mean completion time for C than for B.

The ordering of policies is mirrored in errors: mean error rates were 8.3% (A), 22.1% (B), and 13.7% (C), with the C versus B contrast significant at $p < .01$ and reflecting a 38% proportional reduction in mean error rate relative

to B (standardised mean differences were not computed from the published marginal summaries alone). Workload and usability measures tell the same comparative story at a subjective level: mean NASA–TLX global scores were 31.4 ± 11.2 (A), 52.8 ± 14.6 (B), and 41.3 ± 12.9 (C), where C was lower than B ($p < .01$), a 22% proportional reduction relative to the B mean, while mean SUS scores were 74.2 ± 10.8 (A), 52.6 ± 15.3 (B), and 65.8 ± 12.1 (C); C lay nearer the conventional 68-point acceptability norm than B [11], with pairwise ordering consistent with the fitted contrasts rather than descriptive means alone.

5. Discussion

Taken together, the outcomes in Section 4 answer RQ1–RQ3 in order: clear policy gaps between A and B on workload, errors, time, and SUS; statistically supported advantages of C over B on the same endpoints; and a shift of mean SUS for C toward the conventional 68-point norm relative to B [11], with A remaining the fastest low-error condition.

5.1 Mechanisms and Empirical Interpretation

Interpreting those contrasts, tri-modal policy C plausibly benefits from dwell plus head confirmation, which reduces unintended Midas-touch commits under vibration and noise [15], together with post-lock voice refinement that supplies a structured path for underspecified phrasing on advanced tasks even when ASR is WOz-simulated rather than production-grade. Exploratory inspection of trial logs suggests that most error reductions under C co-occur with the dwell-and- confirm path, whereas voice submissions cluster on advanced tasks; this decomposition is treated as hypothesis-generating because channel blocks were pooled in policy B and wizard timing is embedded in the procedure trace.

5.2 Limitations

Those interpretations must, however, respect four limitations that bound generalisation. First, participants were university volunteers rather than vocationally certified equipment operators, which limits direct transfer to professional crews. Second, disturbance was implemented as a desktop browser simulation rather than ISO 2631-calibrated vibration or NIOSH noise dosime- try [22], so magnitudes should be read as controlled stressors rather than cabin metrology. Third, policy B pooled voice-only and dwell-only blocks, so contrasts are policy-level rather than channel-level factorial evidence. Fourth, Wizard-of-Oz mediation embeds wizard reaction time in the observed procedure trace, so latency outcomes bound the documented protocol rather than commercial gaze trackers or noise-robust ASR stacks.

5.3 Implications and Future Work

Despite those constraints, the protocol already suggests practical guidance for harsh-cabin HMIs through a 3 s dwell, a 5 s timeout, and explicit head confirmation for contactless commitment, with sequenced tri-modal input appearing more robust than pooled single-modality contactless use under the tested stressor, while the WOz prototype itself supports rapid policy iteration before sensor integration. Future work should therefore field-test with professional crews in real vehicles, replace WOz channels with production gaze, head pose, and noise-robust ASR while freezing timing semantics, and examine adaptive interaction parameters driven by real-time cabin disturbance sensing once calibrated telemetry exists, in line with context-sensitive HMI lines [23] and modality-choice work under automation [24].

6. Conclusion

A closed-loop sequenced tri-modal contactless interaction protocol for harsh special-purpose ve- hicle cabins was specified and evaluated with a desktop Wizard-of-Oz console under fixed *Hard* disturbance. In particular, the within-subjects study ($N = 30$) showed that tri-modal policy C improved error rate, NASA–TLX workload, completion time, and SUS relative to pooled single- modality contactless B, while touch A remained fastest overall with the lowest descriptive mean errors. These results support interaction-level design guidance for dwell locking, explicit con- firmation, and optional linguistic refinement before investing in production sensing stacks. The browser console, analysis scripts, and anonymised session logs can be shared for replication on reasonable request and subject to ethics clearance.

References

[1] Lorenz, S., Helmert, J.R., Anders, R., Wölfel, C., Krzywinski, J. (2020) UUX evaluation of a digitally advanced human–machine interface for excavators. *Multimodal*

Technologies and Interaction, 4(3): 57.

[2] Tan, Z., Dai, N., Su, Y., et al. (2022) Human–machine interaction in intelligent and con- nected vehicles: A review of status quo, issues, and opportunities. *IEEE Transactions on Intelligent Transportation Systems*, 23(9): 13954–13975.

[3] Murali, P.K., Kaboli, M., Dahiya, R. (2022) Intelligent in-vehicle interaction technologies. *Advanced Intelligent Systems*, 4(2): 2100122.

[4] Khan, M.Q., Lee, S. (2019) Gaze and eye tracking: Techniques and applications in ADAS. *Sensors*, 19(24): 5540.

[5] Dua, M., Akanksha, Dua, S. (2023) Noise robust automatic speech recognition: Review and analysis. *International Journal of Speech Technology*, 26: 475–519.

[6] Zimmermann, M., Bengler, K. (2013) A multimodal interaction concept for cooperative driv- ing. In: *2013 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, Piscataway, NJ. pp. 1285– 1290.

[7] Nesselrath, R., Moniri, M.M., Feld, M. (2016) Combining speech, gaze, and micro-gestures for the multimodal control of in-car functions. In: *Proceedings of the 12th International Conference on Intelligent Environments (IE)*. IEEE, Piscataway, NJ. pp. 190–193.

[8] Aftab, A.R. (2019) Multimodal driver interaction with gesture, gaze and speech. In: *Pro- ceedings of the 21st ACM International Conference on Multimodal Interaction (ICMI '19)*. ACM, New York. pp. 487–492.

[9] Ayoub, J., Zhou, F., Bao, S., et al. (2019) From manual driving to automated driving: A review of 10 years of AutoUI. In: *Adjunct Proceedings of the 11th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*. ACM, New York. pp. 70–90.

[10] Dahlbäck, N., Jönsson, A., Ahrenberg, L. (1993) Wizard of Oz studies—why and how. *Knowledge-Based Systems*, 6(4): 258–266.

[11] Brooke, J. (1996) SUS: A "quick and dirty" usability scale. In: Jordan, P.W., Thomas, B., Weerdmeester, B.A., McClelland, A.L. (Eds.), *Usability Evaluation in Industry*. Taylor & Francis, London. pp. 189–194.

[12] Ojsteršek, T.C., Topolšek, D. (2019) Eye tracking use in researching driver distraction: A sci- entometric and qualitative literature review approach. *Journal of Eye Movement Research*, 12(3): Article 5.

[13] Roider, F., Rümelin, S., Pfleging, B., Gross, T. (2017) The effects of situational demands on gaze, speech and gesture input in the vehicle. In: *Proceedings of the 9th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutoUI)*. ACM, New York. pp. 94–102.

[14] Khamis, M., Alt, F., Bulling, A. (2018) The past, present, and future of gaze-enabled handheld mobile devices. In: *Proceedings of the 20th International Conference on Human– Computer Interaction with Mobile Devices and Services (MobileHCI '18)*. ACM, New York. pp. 1–17.

[15] Mohan, P., Goh, W.B., Fu, C.-W., Yeung, S.-K. (2018) DualGaze: Addressing the Midas touch problem in gaze mediated VR interaction. In: *2018 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*. IEEE, Piscataway, NJ. pp. 79–84.

[16] Prabhakar, G., Ramakrishnan, A., Murthy, L.R.D., et al. (2020) Interactive gaze and fin- ger controlled HUD for cars. *Journal on Multimodal User Interfaces*, 14: 101–121.

[17] Martelaro, N., Ju, W. (2017) WoZ Way: Enabling real-time remote interaction proto- typing. In: *Proceedings of the 2017 ACM Conference on Computer Supported Cooper- ative Work and Social Computing (CSCW '17)*. ACM, New York. pp. 169–182.

[18] Detjen, H., Pfleging, B., Schneegass, S. (2020) A Wizard of Oz field study to understand non- driving-related activities, trust, and acceptance of automated vehicles. In: *Proceedings of the 12th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutoUI)*. ACM, New York. pp. 19–29.

[19] Schlögl, S., Doherty, G., Luz, S. (2024) Wizard of Oz experimentation for language technol- ogy applications: Challenges and tools. *arXiv* :2402.14563.

[20] Hart, S.G., Staveland, L.E. (1988) Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In: Hancock, P.A., Meshkati, N. (Eds.), *Human Mental Workload*. North-Holland, Amsterdam. pp. 139–183.

[21] Detjen, H., Geisler, S., Schneegass, S. (2020) Maneuver-based control interventions during automated driving:

Comparing touch, voice, and mid-air gestures as input modalities. In: *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, Piscataway, NJ. pp. 3268–3274.

[22] National Institute for Occupational Safety and Health (NIOSH). (1998) *Criteria for a Recommended Standard: Occupational Noise Exposure*. DHHS (NIOSH) Publication No. 98-126. U.S. Department of Health and Human Services, Cincinnati, OH. Available: https://www.cdc.gov/niosh/docs/98-126/

[23] Hussain, J., Ul Hassan, A., Muhammad Bilal, H.S., et al. (2018) Model-based adaptive user interface based on context and user experience evaluation. *Journal on Multimodal User Interfaces*, 12(1): 1–16.

[24] Detjen, H., Geisler, S., Schneegass, S. (2021) Driving as side task: Exploring intuitive input modalities for multitasking in automated vehicles. In: *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, New York. Article 180, pp. 1–6.

