Guiding Task Choice in Japanese Voice Interfaces through Vocalization Cost: Click-based vs. Voice-based Selection

Ryunosuke Shigematsu
Meiji University
Nakano, Japan
sryu131@icloud.com

Ryuto Oishi
Meiji University
Nakano, Japan
ryuto0115fms@gmail.com

Yuki Nakagawa
Meiji University
Nakano, Japan
yukkxi5050@gmail.com

Satoshi Nakamura
Meiji University
Nakano, Japan
satoshi@snakamura.org

Takeshi Torii
SUBARU CORPORATION
Mitaka, Japan
torii.takeshi@subaru.co.jp

Hideyuki Takao
SUBARU CORPORATION
Mitaka, Japan
takao.hideyuki@subaru.co.jp

Abstract

Intrinsic motivation is known to improve task performance when individuals make their own choices. However, when multiple tasks are available, people often choose easier ones even when more difficult or troublesome tasks may be more beneficial. This study investigates whether the phrasing of spoken options can influence such decisions in Voice-based interfaces by leveraging the cognitive and articulatory effort required for vocalization. We conducted a controlled experiment with 40 participants, systematically varying the linguistic complexity of Japanese adverbial phrases in a pointing task and comparing Voice-based and Click-based selection. Results indicated a clear tendency in the voice condition to avoid the most complex phrase and revealed a modality-specific positional tendency in which left-positioned options were chosen more often and right-positioned options were avoided. To our knowledge, this is the first empirical study to demonstrate that vocalization cost can systematically bias task selection in Japanese voice interfaces. These findings suggest that carefully designed spoken language can subtly guide task selection, providing implications for fair and effective voice interface design.

CCS Concepts

• **Human-centered computing** → **User interface design**; **Interaction design**; **Empirical studies in HCI**; *Interface design prototyping*.

Keywords

Intrinsic Motivation, Click-based Selection, Voice-based Selection, Pointing Task, Selection Bias, Cognitive Load

ACM Reference Format:
Ryunosuke Shigematsu, Ryuto Oishi, Yuki Nakagawa, Satoshi Nakamura, Takeshi Torii, and Hideyuki Takao. 2025. Guiding Task Choice in Japanese Voice Interfaces through Vocalization Cost: Click-based vs. Voice-based Selection. In *ACM Multimedia Asia (MMAsia '25), December 09–12, 2025, Kuala Lumpur, Malaysia.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3743093.3771074

MMAsia '25, Kuala Lumpur, Malaysia

ACM ISBN 979-8-4007-2005-5/25/12
https://doi.org/10.1145/3743093.3771074

1 Introduction

When engaging in a task, intrinsic motivation is known to enhance performance, particularly when individuals are free to choose the task themselves [10]. However, when multiple tasks are available, people often prioritize easier tasks and postpone more difficult or troublesome ones. Encouraging the selection of such difficult yet beneficial tasks is therefore an important challenge in task design. One possible approach is to shape the decision environment so that natural choice tendencies guide users toward less-preferred but worthwhile tasks.

In guiding user choices, factors such as visual pop-out effects [6], the position of options [1, 16, 17] and the color of options [13] have been shown to influence selection behavior. In Voice-based selection, an additional factor may play a role. This is the cognitive and articulatory effort required to say a choice aloud, which we refer to as vocalization cost. This cost is affected by the linguistic complexity of a phrase, including its length, structure, and clarity. If vocalization cost systematically biases selection, carefully choosing the phrasing of spoken options could help promote the choice of more beneficial tasks.

To examine this possibility, we propose a method to encourage task selection using a Voice-based interface. We conduct a controlled experiment comparing Voice-based and Click-based selection in a pointing task, systematically varying the complexity of Japanese adverbial phrases while keeping task content constant. We also explore whether Voice-based selection exhibits positional tendencies that differ from those observed in graphical user interfaces, such as a greater likelihood of choosing left-positioned options and avoiding right-positioned ones. Our aim is to provide empirical evidence and design insights for building choice architectures in voice interfaces that are both fair and effective. This is the first empirical study to demonstrate that vocalization cost can bias task selection in Japanese voice interfaces.

Contributions are as follows:

- We introduced a novel use of *vocalization cost*, the cognitive and articulatory effort of speaking, in the context of voice interfaces as a potential factor to influence user selection.

- We designed and conducted a controlled experiment demonstrating that differences in vocalization cost could lead to measurable selection bias, and that linguistic complexity could be used to shift selection distributions without altering task content.
- We found that in Voice-based selection, options positioned on the left were chosen more frequently, while those on the right were chosen less frequently, revealing a positional bias unique to the voice modality.

2 Related Work

2.1 Selection Bias

How options are presented, known as choice architecture [7], can strongly influence decisions. Positional bias is one example: Wilson and Nisbett [1] found that, when identical items are arranged in a row, people often prefer those on the right, whereas Valenzuela and Raghubir [17] identified the "center-stage effect," in which centrally positioned items are more likely to be chosen.

Timing also acts as a nudge: earlier presentation of options tends to attract more attention and increase selection frequency. Even minimal cues before the presentation of options, such as a simple progress bar, can guide gaze direction and create selection biases [19].

Furthermore, psychological research has demonstrated the framing effect, in which identical information can yield different choices depending on whether it is presented as a gain or a loss [15]. Similarly, order effects [5] in belief updating reveal that the sequence of information, whether presented early or late, can disproportionately shape judgments. Together, these findings indicate that presentation timing, positional arrangement, and linguistic framing can interact to shape user preferences.

2.2 Intrinsic Motivation

Our experimental paradigm draws on Self-Determination Theory [12], which posits that granting autonomy enhances intrinsic motivation, thereby improving both performance and well-being. In our prior work on driving tasks, allowing participants to choose their own goals, whether via click or voice, consistently improved performance compared to externally assigned goals [10].

While many studies have examined how to foster intrinsic motivation, few have explored how the phrasing of choices interacts with the modality of selection (e.g., click vs. voice). As noted by Oviatt [11], each modality has distinct strengths and weaknesses. Our work identifies vocalization cost as a previously underexplored limitation of Voice-based choice tasks, reflecting the cognitive and articulatory effort required to verbalize a phrase.

This research is also informed by the theory of implementation intentions [3], which suggests that forming explicit intentions (e.g., "I will do X") can substantially increase follow-through. This highlights that the act of making a choice can itself serve as a psychological trigger for action, independent of the specific choice content. Our study thus extends existing work by examining how linguistic phrasing and input modality jointly influence motivation and performance in voice interfaces.

3 Proposed Method

We propose a design approach that leverages *vocalization cost*, the cognitive and articulatory effort required to speak a phrase aloud, to potentially guide user choices in voice interfaces. The concept is to assign higher vocalization cost to less desirable options, such as overly easy tasks, and lower cost to more beneficial options, thereby creating conditions that may discourage the former and encourage the latter (Figure 1).

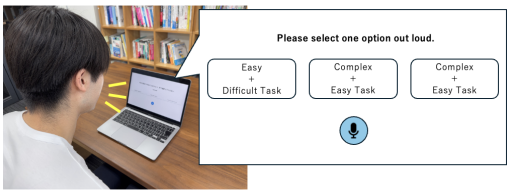

Figure 1: Proposed method: assigning high vocalization cost to less desirable options and lower cost to desirable options.

In this paper, we focus on the first step toward this approach: examining whether differences in vocalization cost can influence the distribution of user selections. The results can provide foundational evidence for future systems that aim to subtly nudge users toward more advantageous choices through linguistic complexity.

4 Experiment

4.1 Overview

This study aimed to examine how the linguistic phrasing of task options influences user selection and subsequent performance in Click-based and Voice-based interactions. Building on the premise that intrinsic motivation can enhance task engagement, we focused on the role of *vocalization cost*, which is the cognitive and articulatory effort required to speak a phrase aloud.

Hypothesis. Adverbial expressions with greater linguistic complexity will be chosen less frequently in Voice-based interactions than in Click-based interactions, even when the underlying task is identical.

To test this hypothesis, we conducted a controlled laboratory experiment. The experimental paradigm followed a repeated two-phase loop. In the first phase, participants selected an *attention prompt* (for example, "be fast") from three displayed options. In the second phase, they immediately performed a standard pointing task. This design allowed us to measure how differences in linguistic complexity affected both choice behavior and task performance across the two input modalities: a graphical user interface (Click-based) and a voice user interface (Voice-based).

Participants. 40 university students (20 male, 20 female; age range 19 to 23 years, $average = 21.1$, $SD = 1.2$) participated. They were randomly assigned to one of the two input modality conditions: Click-based ($n = 20$) or Voice-based ($n = 20$). All participants received a 500 JPY Amazon gift card as compensation. No participants met the exclusion criteria, which included task incompletion or recognition errors exceeding 20% of trials.

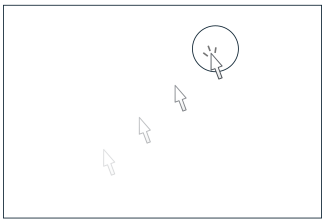

Figure 2: The user interface for the pointing task. During the task, target circles appear one by one at random positions within the designated area, and the participant clicks them in the order in which they appear. The numbers in the figure indicate the appearance order and are for illustration only.

4.2 Experimental Task

The experimental task was a Fitts's Law-style serial pointing task [2, 8] (Figure 2). Fitts's Law models the time required to move to a target as a function of the distance to the target and its size, and has been widely used in HCI research to evaluate pointing performance. The parameters were as follows:

- **Objective:** To correctly click a series of 20 circles presented sequentially.
- **Display Area:** 800×600 pixels.
- **Target Diameter:** Ranged randomly from 50 to 125 pixels.
- **Target Position:** Each new target appeared at a random location after the previous one was successfully clicked.

The task was implemented as a web application using JavaScript and the Web Speech API. All participants used same PC with a standard Mouse to ensure consistency.

4.3 Experimental Design

The experiment utilized a mixed design. Each selectable option presented to participants was created by combining one of three linguistic expressions with one of five attention prompts, resulting in 15 unique phrase–prompt pairs.

Linguistic expression. To investigate the effect of linguistic expression, we focused specifically on adverbial phrases. These phrases were used to modify a core task goal (e.g., "be fast") to alter its perceived nuance and required commitment, without changing the fundamental instruction. This allowed us to isolate the impact of phrasing itself. Based on agreement among the authors, we selected three Japanese expressions that, while all conveying a sense of "doing one's best," vary in their length, formality, and complexity. For clarity, we label them *Simple*, *Formal*, and *Complex*, with the specific details of each presented in Table 1.

Attention prompt. By author consensus, we adopted five attention prompts for participants to focus on. In addition to "be fast" and "be accurate," which were used in prior work on pointing tasks by Yamanaka et al. [18], we included three additional prompts that are objectively measurable: "keep a steady rhythm," "aim for the center," and "minimize movement."

We measured the selection rate and performance metrics corresponding to the five attention prompts:

- **Selection Rate:** The percentage of times each adverbial phrase (*Simple*, *Formal*, *Complex*) was chosen.
- **Task Performance:** Five specific metrics were recorded, each corresponding to one of the five attention prompts. The metrics were calculated as follows:
  - *Speed:* Total task completion time (s). (Corresponds to "be fast")
  - *Accuracy:* Error rate (%). (Corresponds to "be accurate")
  - *Rhythm:* The standard deviation of the elapsed time between consecutive clicks. (Corresponds to "keep a steady rhythm")
  - *Center Proximity:* The sum of Euclidean distances from each target's center to the actual click position. (Corresponds to "aim for the center")
  - *Movement Efficiency:* The ratio of the total mouse cursor travel distance to the optimal path distance. (Corresponds to "minimize movement")

In this paper, we compare the selection rates of *Simple*, *Formal*, and *Complex* expressions between Click-based and Voice-based selection. However, if all three expression types were always presented, participants might easily infer the experiment's purpose. To prevent this, we included dummy combinations such as {*Simple*, *Formal*, *Formal*} and {*Simple*, *Complex*, *Complex*}.

In each trial, participants were shown three options. To ensure a balanced comparison while masking the study's true purpose, the three options were generated as follows:

(1) Three distinct *Attention prompts* were randomly selected from the pool of five.
(2) The adverbial phrase *Simple* was always assigned to one of these prompts.
(3) For the other two prompts, adverbial phrases were chosen randomly from *Formal* and *Complex*, with replacement.
(4) The resulting three options were shuffled and displayed in a random order.

This algorithm ensured that the simple expression (*Simple*) always appeared alongside more complex expressions (*Formal*, *Complex*), allowing us to test whether simpler phrasing influenced choices. As a result, 50% of the choices were {*Simple*, *Formal*, *Complex*}.

4.4 Procedure

Participants were first briefed on the experiment. They then tried a short tutorial to familiarize themselves with the pointing task. The main experiment consisted of 20 trials. Each trial involved:

(1) **Choice Phase:** Participants were presented with three goal options on the screen. In the click condition (Figure 3), they used a mouse to select one. In the voice condition (Figure 4), they selected an option by speaking the entire phrase aloud into a microphone.
(2) **Task Phase:** After confirming their choice, participants viewed a screen that displayed (1) a confirmation message showing their selected instruction (e.g., "You have selected: As much as possible, keep a steady rhythm."), (2) the current task number, and (3) the number of remaining clicks. A Start button was shown, and participants were reminded that the task would begin immediately after pressing it.

Table 1: The three Japanese adverbial phrases used in the experiment, ordered by increasing linguistic complexity.

Condition	Japanese Phrase	Romaji	Description of Nuance
Simple	できるだけ	*dekirudake*	Simple and colloquial. The most basic phrase, written entirely in Hiragana and commonly used in daily speech.
Formal	可能な限り	*kanou na kagiri*	Formal and literal. Written with Kanji, it has a more structured and formal tone than *Simple*.
Complex	やれる範囲で最大限に	*yareru han'i de saidaigen ni*	Complex and descriptive. The longest phrase, containing the most Kanji, and explicitly elaborating the intended meaning.

Figure 3: The choice interface for the click condition. (1) Task instruction: "Please select one thing to pay attention to from now on." (2) Selection counter: Indicates the current selection number (e.g., "First selection"). (3) Clickable choice buttons: Participants select one of the three options by clicking with the mouse (e.g., "As much as possible, keep a steady rhythm"; "Aim for the center as much as possible"; "Minimize movement as much as possible within your ability").

Figure 4: The choice interface for the voice condition. (1)–(2) are the same as in Figure 3. (3) Spoken choice phrases: Participants select one of the three options by saying it aloud after pressing the voice input button (e.g., "Be as accurate to the center as possible within your ability"; "Be as fast as possible within your ability"; "Be as short in movement as possible"). (4) Voice input button: Participants press the button and then speak their chosen phrase aloud to make a selection.

Upon pressing Start, they performed the pointing task while keeping their chosen goal in mind.

(3) **Completion:** After successfully clicking 20 targets, participants saw a screen with a large "Completed" message and a button to return to the next trial's choice screen.

Participants took one 5-minute break after the 10th trial. The entire session concluded with a short post-experiment questionnaire.

5 Results

5.1 Analysis Strategy

We analyzed the data from two perspectives to examine biases in Click-based and Voice-based selection. First, to assess selection bias, we compared the selection rates of the three adverbial phrases (*Simple*, *Formal*, and *Complex*) across modalities and tested whether option position (Left, Center, Right) affected users' choices. Second, to evaluate task performance, we analyzed five metrics to compare performance between modalities and examine how the chosen phrase influenced subsequent task execution. This analysis also confirmed that participants were actively engaged with the prompts they selected.

5.2 Analysis of Selection Rates

We first analyzed the selection rates for trials where the three options presented consisted of one of each adverbial phrase type: *Simple*, *Formal*, and *Complex*. This subset of the data allowed for a fair comparison of the phrases' appeal. The results are summarized in Table 2. For the click condition, the selection rates for the three phrases were comparable, with only minor differences: *Simple* was chosen 30.9% of the time, *Formal* 33.3%, and *Complex* 35.8%. In contrast, the voice condition exhibited a significant selection bias. While *Simple* (38.4%) and *Formal* (37.4%) were selected at similar rates, the most complex phrase, *Complex*, was selected significantly less often, at only 24.2% of the time.

Across both modalities, we observed a consistent preference for specific attention prompts. The prompt to "be fast" was the most popular, particularly in the voice condition, where it was selected over half the time (51.3% on average). Conversely, the prompts "aim for the center" and "minimize movement" were consistently the least popular, with the former being chosen only 14.8% of the time in the voice condition.

The negative bias against the complex phrase, *Complex*, in the voice condition was particularly pronounced when combined with the least popular prompts. Notably, the option combining *Complex* with "aim for the center" was 4.7%, and with "minimize movement" was chosen just 9.8% of the time. A chi-square test of independence revealed no significant association between modality (Click vs. Voice) and phrase selection rates, $\chi^2(2, N = 800) = 2.09$, $p = .352$, Cramér's $V = 0.05$.

5.3 Task Performance

We analyzed the five task performance metrics to validate our setup and test our hypotheses. First, to confirm that participants understood and engaged with their chosen prompts, we compared performance on trials where a specific prompt was selected versus all other trials where it was not. As shown in Table 3, for all five metrics, performance was significantly better when a corresponding prompt was selected, confirming that the self-selection paradigm effectively directed participants' attention.

Table 2: Selection rates of each individual option when all three were presented, in both Click-based and Voice-based selection.

Modality / Phrase		be fast	be accurate	keep a steady rhythm	aim for the center	Minimize movement	Average
Click	*Simple*	42.2	30.6	34.2	23.1	23.3	30.9
	Formal	46.8	22.0	47.2	26.8	23.1	33.3
	Complex	40.5	38.1	38.5	33.3	29.0	35.8
	Average	43.4	30.3	39.7	28.1	25.0	33.3
Voice	*Simple*	55.3	50.0	34.0	19.4	29.0	38.4
	Formal	52.5	29.7	50.0	22.2	30.2	37.4
	Complex	45.7	34.1	31.4	4.7	9.8	24.2
	Average	51.3	38.6	38.7	14.8	22.6	33.2

Table 3: Performance metrics by phrase type and group, aggregated over all trials in both Click and Voice-based selection.

Modality / Group		Speed (s)	Accuracy (%)	Rhythm (SD)	Center (px)	Movement (Ratio)
Click	*Simple*	12.58	3.02	95.6	153.3	1.05
	Formal	12.38	2.43	102.9	153.1	1.09
	Complex	12.56	4.58	120.6	150.5	1.02
	Selected Group	12.51	3.36	105.8	152.2	1.05
	Non-selected Group	17.60	5.46	160.9	409.8	1.17
Voice	*Simple*	12.10	3.07	88.8	141.3	1.12
	Formal	12.24	2.13	95.7	137.5	1.21
	Complex	11.85	1.35	94.7	121.6	1.12
	Selected Group	12.07	2.37	93.4	135.1	1.15
	Non-selected Group	15.28	5.62	135.7	399.8	1.20

Next, we conducted a two-way mixed ANOVA with Modality (Click vs. Voice) as a between-subjects factor and Phrase (*Simple*, *Formal*, *Complex*) as a within-subjects factor for task speed and accuracy. For speed, there was a significant main effect of Modality, $F(1,794) = 31.13$, $p < .001$, indicating that participants in the voice condition completed tasks faster than those in the click condition. The main effect of Phrase was not significant, $F(2,794) = 1.99$, $p = .138$, and the Modality × Phrase interaction did not reach significance, $F(2,794) = 1.89$, $p = .151$. For accuracy, no significant effects were observed for Modality, $F(1,794) = 0.15$, $p = .699$, Phrase, $F(2,794) = 2.11$, $p = .122$, or their interaction, $F(2,794) = 0.45$, $p = .636$.

Finally, we conducted exploratory analyses on the rare cases where participants selected the *Complex* phrase in the voice condition. Given few such trials ($n = 109$), we used non-parametric Mann-Whitney U tests to compare performance against *Simple*/*Formal* selections ($n = 291$). Results indicated no significant differences in speed ($U = 15{,}178$, $p = .508$) or accuracy ($U = 14{,}716.5$, $p = .233$). However, descriptive trends suggested that participants tended to perform faster and more accurately when selecting *Complex*, indicating a possible link between selecting more demanding options and improved task execution. These observations are preliminary and warrant further study.

6 Discussion

6.1 Selection Bias

Our results support our central hypothesis: the phrasing of choices induces significant selection bias in voice interfaces but not in GUIs. While phrase selection in the click condition was evenly distributed, the voice condition revealed a strong aversion to the most *Complex* phrase. This may be due to three interacting factors: the cognitive load imposed by vocalization, visual pop-out effects in GUIs, and modality-specific processing patterns.

The primary driver for the bias in the voice condition appears to be what we term "vocalization cost." The *Complex* phrase was significantly longer and more linguistically complex than *Simple* and *Formal*. The act of planning and articulating this long phrase likely imposes a higher cognitive load on the user [14]. This additional effort, which is absent in the uniform physical action of clicking, may serve as a negative nudge, pushing users towards the more easily spoken options, even if the underlying task is the same. This effect was markedly amplified when the high-cost phrase was paired with an already unpopular task prompt (e.g., "aim for the center"), where the selection rate plummeted to just 4.7%.

In contrast, the analysis of choice set composition in the click condition revealed a different psychological phenomenon: a visual pop-out effect [6]. The selection rate of *Simple* was significantly higher when it was the only unique option presented alongside two identical phrases (e.g., *Simple*, *Formal*, *Formal*). In these sets, *Simple* becomes a visual singleton, attracting more attention and increasing its likelihood of being selected. This effect was absent in the voice condition, suggesting that this type of visual distinctiveness is not a salient factor when the selection method is speech.

Furthermore, our results revealed a modality-specific positional bias. As shown in Table 4, selection in the click condition showed a significant non-uniform distribution ($\chi^2(2, N = 460) = 6.53$, $p = .038$, $W = 0.12$), with a slight preference for the center option. In contrast, in the voice condition, participants appeared to prefer the

Table 4: Selection rate by position for Click-based vs. Voice-based selection (%).

	Left	Center	Right
Click	33.3	33.3	33.4
Voice	36.0	34.3	29.8

leftmost option (36.0%) and avoid the rightmost one (29.8%), but this distribution was not statistically significant ($\chi^2(2, N = 420) = 2.70$, $p = .259$, $W = 0.08$). We hypothesize this is due to different processing strategies: GUI users can visually scan all options in parallel, whereas VUI users may process the options serially from left to right, as one would read a sentence, potentially committing to an earlier option to minimize cognitive load. While the present findings suggest a tendency to prefer leftmost options in the voice condition, the underlying reason remains unclear. For example, future work could test whether this bias is related to reading direction by conducting an additional experiment in which options are arranged from right to left or vertically, as in traditional Japanese layouts. While our behavioral data supports this hypothesis, it remains speculative without direct evidence.

While these findings provide actionable insights for designing more efficient voice interfaces, they also highlight potential risks. Prior work has shown that dark patterns, interface designs that intentionally steer users toward certain choices, are widespread in commercial contexts [9] and can arise even from subtle presentation changes [4]. Techniques that leverage vocalization cost or positional biases could, if misapplied, function as such patterns, subtly guiding users without their full awareness. Future work should therefore consider not only performance and usability outcomes, but also the ethical implications of these design choices.

6.2 Influence of Modality on Task Performance

Beyond selection bias, our results indicate that the choice modality can influence task performance. Participants in the voice condition generally performed better than those in the click condition, suggesting that verbal commitment, which is a more personal and embodied action, may enhance self-determination and foster intrinsic motivation for the subsequent task. The one exception, lower *Movement* in the voice condition, may reflect cognitive spillover from the speech act and warrants further investigation.

Our two-way mixed ANOVA revealed a significant main effect of Modality on task speed, $F(1,794) = 31.13$, $p < .001$, confirming that Voice input generally improved performance. However, the Modality × Phrase interaction did not reach significance for either speed or accuracy, indicating that the performance boost observed for *Complex* in the voice condition remains inconclusive from an inferential standpoint.

Finally, we conducted exploratory analyses focusing on the rare cases where participants selected the Complex phrase (*Complex*) in the voice condition. Due to the limited number of such trials ($n = 109$), we employed non-parametric Mann-Whitney U tests to compare performance against *Simple*/*Formal* selections ($n = 291$). Results indicated no significant differences in speed ($U = 15{,}178$, $p = .508$) or accuracy ($U = 14{,}716.5$, $p = .233$). However, descriptive trends suggested that participants tended to perform faster and more accurately when selecting *Complex*. While preliminary, these observations indicate a potential association between choosing more demanding options and improved task execution.

6.3 Limitations

This study has several limitations. First, the number of *Complex* trials in the voice condition was small, so trends for that cell should be interpreted cautiously and replicated with more observations. Second, proposed mechanisms such as serial processing in voice interfaces remain inferential. Methods like eye tracking are needed to test processing strategies directly. Third, baseline performance differences between the click and voice groups were not controlled. Including a pre-test would enable more precise estimates of gains. Fourth, participants were Japanese university students and all phrases were in Japanese, which limits linguistic and cultural diversity. Future studies should include broader age ranges and other languages. Finally, the evaluation used a simplified laboratory task, applying the framework to everyday voice interactions such as smart assistants will improve ecological validity.

7 Conclusion

This study demonstrated that the vocalization cost, defined as the cognitive and articulatory effort required to verbalize a phrase, can significantly bias user choice behavior in voice interfaces. Through a controlled experiment comparing Click-based and Voice-based selection while systematically varying the complexity of Japanese adverbial phrases, we showed that differences in vocalization cost led to measurable selection bias without altering task content. Participants in the voice condition consistently avoided complex phrases, indicating a strong influence of vocalization cost. Moreover, we identified a positional tendency unique to Voice-based selection, with left-positioned options chosen more frequently and right-positioned options chosen less frequently. This work represents the first empirical demonstration of vocalization cost as a biasing factor in Japanese voice interfaces and provides a novel application of this concept as a design lever. Exploratory analyses further suggested that, although infrequently chosen, complex phrases were associated with a trend toward improved task performance, which warrants further investigation with larger samples.

These findings provide practical design implications. Reducing vocalization cost for tasks that are rarely chosen can promote their selection, while strategically introducing moderate cost through phrasing may encourage engagement with more cognitively demanding tasks. Future work should explore how to balance these factors to design voice interface choice architectures that are both effective and equitable.

Acknowledgments

The authors used ChatGPT (OpenAI, 2025) solely to improve the clarity and readability of the English text. All scientific content, including the research design, analysis, and conclusions, was developed entirely by the authors.

References

[1] Timothy de Camp Wilson and Richard E Nisbett. 1978. The accuracy of verbal reports about the effects of stimuli on evaluations and behavior. *Social Psychology* (1978), 118–131.

[2] Paul M Fitts. 1954. The information capacity of the human motor system in controlling the amplitude of movement. *Journal of Experimental Psychology* 47, 6 (1954), 381–391.

[3] Peter M Gollwitzer. 1999. Implementation intentions: strong effects of simple plans. *American psychologist* 54, 7 (1999), 493.

[4] Colin M Gray, Yubo Kou, Bryan Battles, Joseph Hoggatt, and Austin L Toombs. 2018. The dark (patterns) side of UX design. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–14.

[5] Robin M Hogarth and Hillel J Einhorn. 1992. Order effects in belief updating: The belief-adjustment model. *Cognitive psychology* 24, 1 (1992), 1–55.

[6] Mitsuki Hosoya, Hiroaki Yamaura, Satoshi Nakamura, Makoto Nakamura, Eiji Takamatsu, and Yujiro Kitaide. 2019. Does the pop-out make an effect in the product selection of signage vending machine?. In *Proceedings of the IFIP Conference on Human-Computer Interaction*. Springer, 24–32.

[7] Thomas C Leonard. 2008. *Richard H. Thaler, Cass R. Sunstein, Nudge: Improving Decisions about Health, Wealth, and Happiness*. Yale University Press, New Haven, CT.

[8] I. Scott MacKenzie. 1992. Fitts' Law as a research and design tool in human-computer interaction. *Human–Computer Interaction* 7, 1 (1992), 91–139.

[9] Arunesh Mathur, Gunes Acar, Michael J Friedman, Eli Lucherini, Jonathan Mayer, Marshini Chetty, and Arvind Narayanan. 2019. Dark patterns at scale: Findings from a crawl of 11K shopping websites. *Proceedings of the ACM on human-computer interaction* 3, CSCW (2019), 1–32.

[10] Yuki Nakagawa, Sayuri Matsuda, Takumi Takaku, Satoshi Nakamura, Takanori Komatsu, Takeshi Torii, Ryuichi Sumikawa, and Hideyuki Takao. 2023. A Study on the Effects of Intrinsic Motivation from Self-determination on Driving Skill. In *Proceedings of the International Conference on Human-Computer Interaction (HCII 2023) (CCIS, Vol. 1836)*. Springer, 73–81.

[11] Sharon Oviatt. 2007. Multimodal interfaces. *The Human-Computer Interaction Handbook* (2007), 439–458.

[12] Richard M Ryan and Edward L Deci. 2000. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist* 55, 1 (2000), 68.

[13] Yuto Sekiguchi, Riho Ueki, Kouta Yokoyama, and Satoshi Nakamura. 2023. Does the Average Color Influence Selection?. In *Proceedings of the International Conference on Human-Computer Interaction*. Springer, 485–496.

[14] John Sweller and Paul Chandler. 1991. Evidence for cognitive load theory. *Cognition and Instruction* 8, 4 (1991), 351–362.

[15] Amos Tversky and Daniel Kahneman. 1981. The framing of decisions and the psychology of choice. *science* 211, 4481 (1981), 453–458.

[16] Riho Ueki, Kouta Yokoyama, and Satoshi Nakamura. 2023. Does the Type of Font Face Induce the Selection?. In *Proceedings of the International Conference on Human-Computer Interaction*. Springer, 497–510.

[17] Ana Valenzuela and Priya Raghubir. 2009. Position-based beliefs: The center-stage effect. *Journal of Consumer Psychology* 19, 2 (2009), 185–196.

[18] Shota Yamanaka, Taiki Kinoshita, Yosuke Oba, Ryuto Tomihari, and Homei Miyashita. 2023. Varying subjective speed-accuracy biases to evaluate the generalizability of experimental conclusions on pointing-facilitation techniques. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–13.

[19] Kota Yokoyama, Satoshi Nakamura, and Shota Yamanaka. 2021. Do animation direction and position of progress bar affect selections?. In *Proceedings of the IFIP Conference on Human-Computer Interaction*. Springer, 395–399.