User-Defined Co-Speech Gesture Design with Swarm Robots

Minh Duc Dang
duc_dang@sfu.ca
School of Computing Science
Simon Fraser University
Burnaby, BC, Canada

Samira Pulatova
pulatova.samira@gmail.com
School of Computing Science
Simon Fraser University
Burnaby, BC, Canada

Lawrence H Kim
lawkim@sfu.ca
School of Computing Science
Simon Fraser University
Burnaby, BC, Canada

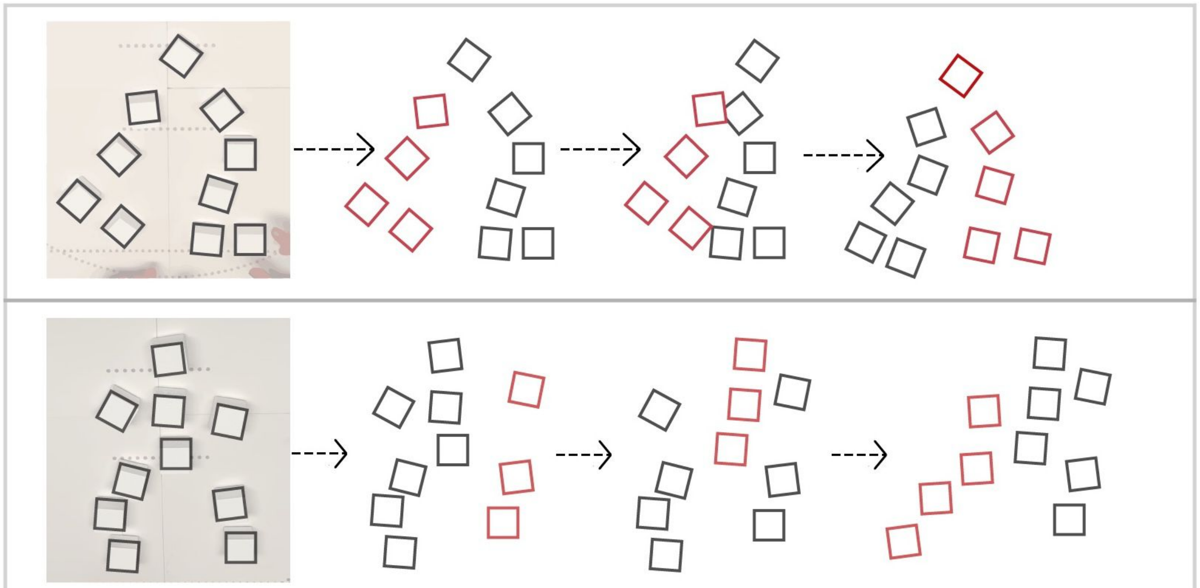


Figure 1: Two user-defined patterns of swarm robot co-speech gestures corresponding to the word "run." Each pattern includes multiple phases, with red squares indicating the robots that have moved relative to the previous phase.

ABSTRACT

Non-verbal signals, including co-speech gestures, play a vital role in human communication by conveying nuanced meanings beyond verbal discourse. While researchers have explored co-speech gestures in human-like conversational agents, limited attention has been given to non-humanoid alternatives. In this paper, we propose using swarm robotic systems as conversational agents and introduce a foundational set of swarm-based co-speech gestures, elicited from non-technical users and validated through an online study. This work outlines the key software and hardware requirements to advance research in co-speech gesture generation with swarm robots, contributing to the future development of social robotics and conversational agents.

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

CCS CONCEPTS

• **Human-centered computing** → *Empirical studies in interaction design.*

KEYWORDS

Swarm Robotics, Co-speech Gesture, Elicitation Study

ACM Reference Format:
Minh Duc Dang, Samira Pulatova, and Lawrence H Kim. 2025. User-Defined Co-Speech Gesture Design with Swarm Robots. In *Proceedings of ACM Conference (Conference'17)*. ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

In interpersonal communication, humans use various non-verbal cues such as eye gaze, facial expressions, body posture, and gestures to convey information and complement verbal discourse [15]. Among these non-verbal behaviors, gestures that synchronize with speech - known as co-speech gestures - have garnered significant research attention. Psychology and Neuroscience studies have shown that co-speech gestures enable interlocutors to exchange nuanced meanings beyond what is conveyed through speech alone [13, 75]. In Human-Robot Interaction (HRI) literature, it has been shown that robots using co-speech gestures are perceived as more likable and lively, and they also enhance the task performance of their human listeners [9, 60, 64]. Given these benefits of co-speech gestures in human-human and human-robot conversations, research has explored designing co-speech gesture systems for conversational agents, including social robots and virtual chatbots [42]. Furthermore, recent advancements in deep learning have accelerated the development of these co-speech gesture systems, enabling the agents to perform more expressive gestures [50].

Although there is growing interest in co-speech gestures for conversational agents, existing research has predominantly focused on human-like agents, such as humanoid robots or realistic virtual avatars. This focus is understandable, as the human-like appearance of these agents allows them to apply principles and insights from the study of human communication more readily, thereby replicating the dynamics observed in human-human conversations [64]. Nevertheless, there are notable drawbacks to human-like conversational agents. Firstly, a resemblance to humans does not always equate to appeal - agents that look too similar to humans can evoke the uncanny valley phenomenon, causing users to perceive them as eerie or unsettling [49]. Secondly, in the field of social robotics, developing realistic humanoid robots is both difficult and costly. A survey on speech-based gesture generation for conversational agents found that most studies used virtual agents rather than humanoid robots, citing the limited degrees of freedom and high development costs as significant obstacles to creating realistic co-speech gestures [42].

Despite the challenges in developing co-speech gestures for human-like conversational agents and the significant effects of robot's anthropomorphism on humans such as task performance and emotion [30, 31, 43], there has been limited exploration of non-humanoid alternatives. Indeed, existing studies on non-humanoid robots as conversational agents have primarily focused on using specialized bodies and movements that convey specific meanings, often overlooking gestures as a means of more generalized non-verbal communication [26, 28, 46]. To address this gap in research, our project investigates the potential of using swarm robots as conversational agents. This choice is inspired by prior works that have proposed swarm robot systems as flexible and scalable user interfaces with abundant degrees of freedom [29, 33, 34, 38, 41, 43, 58, 71, 78, 80]. Additionally, research in animation has shown that simple movements of non-humanoid bodies, such as basic shapes or everyday objects, can convey complex intentions and emotions effectively, similar to how gestures do. [22, 72]. Therefore, we hypothesize that the movements of a swarm robot system can be synchronized with speech to function as co-speech gestures.

While designing gestures for humanoid conversational agents mainly involves imitating human arm and hand movements, it is unclear how these gestures can be replicated in swarm robots. Building on prior research in gesture design [32, 77], we conducted an elicitation study to generate user-defined co-speech gestures in swarm robot systems and to understand how non-technical users envision swarm robots as conversational agents. In summary, the current research makes two key contributions:

- We present an elicitation study that establishes a foundational set of swarm-based co-speech gestures and offers insights into their hardware and software requirements for their implementation.
- We introduce the novel application of swarm robot systems as conversational user interfaces, laying the groundwork for future research in social robotics and conversational agents.

2 RELATED WORKS

This section offers a comprehensive overview of existing research literature, focusing on several key areas: previous studies on co-speech gestures in human-robot interaction, challenges associated with these gestures in social robots, the application of swarm robots as user interfaces, and the use of animation principles and elicitation studies in designing co-speech gestures.

2.1 Co-speech Gestures in Human-Robot Interaction

Co-speech gestures have been established in the Psychology and Neuroscience literature as essential to verbal communication [75]. Studies show that gestures can convey information not mentioned in speech or add nuances to spoken information [17, 18], and their use enhances communication and reduces the cognitive load in conversations [19]. Furthermore, brain scan studies have found that co-speech gestures influence neural activity in regions associated with processing semantic information, suggesting that during a conversation, listeners seek meaning not only in the spoken words but also in the accompanying hand movements [13].

Building on the established benefits of co-speech gestures in human-human conversations, researchers in Human-Robot Interaction have implemented these gestures in humanoid robots to explore their impact on interactions with humans. Studies have shown that incorporating co-speech gestures enhances participants' evaluations of robots, improving perceptions of sympathy [62], liveliness [61], anthropomorphism [60], engagement [3], and social ability [64]. Additionally, robots using co-speech gestures have boosted human task performance, such as improving participants' ability to recall information in storytelling contexts [74].

2.2 Challenges to Implementing Co-speech Gestures in Social Robots

Despite the well-documented benefits of co-speech gestures in robots, significant challenges still hinder the widespread adoption of this technology. Firstly, some humanoid robots with co-speech gesture systems, such as ASIMO [60–62] and KASPAR2 [67], were developed in collaboration with technology companies for academic or industrial purposes [68]. As a result, these robots often have high production costs and large sizes, making them unsuitable as social companions or everyday robotic assistants. Additionally, while efforts have been made to design co-speech gestures for smaller, less expensive humanoid robots like NAO [3, 66], these robots often lack the sophisticated actuator systems necessary for replicating the complex arm, hand, and finger movements found in human gestures. This limitation is underscored by the fact that research on co-speech gesture generation has primarily been conducted with virtual human avatars rather than humanoid robots - these studies cite the limited movement capability of humanoids as a significant barrier to creating realistic gestures [42].

Furthermore, while studies have predominantly focused on co-speech gestures for humanoid robots, their anthropomorphic appearance can sometimes cause discomfort due to the "uncanny valley" effect [49, 76]. For example, Thaler et al. [73] found that participants perceived humanoid virtual agents as more eerie compared to their non-humanoid counterparts. Additionally, Luria et al. [44] discovered that although human-like behaviors in robots are easily understood, they are not always the most effective - non-human elements can sometimes feel more natural.

Despite these challenges of developing co-speech gesture systems for humanoid robots, few studies have examined the alternative of non-humanoid conversational agents, and to our knowledge, none have explored the implementation of co-speech gestures in such robots. For instance, Kim et al. [28] studied a block-based, shape-changing robot in a storytelling context, while other research focused on minimal conversational agents in healthcare settings [26, 46]. There were also studies on non-humanoid robotic gestures, such as Anderson-Bashan et al. [5] on their impact during opening encounters, Rifinski et al. [10] on enhancing human interactions, and Press & Erel [57] on reducing social awkwardness. However, these studies have primarily focused on using specialized bodies and gestures for specific objectives, rather than exploring the broader conversational context of co-speech gestures.

2.3 Swarm Robotics

Given this gap in the research literature, the current study explores a new approach to implementing co-speech gestures in non-humanoid conversational agents - using swarm robots. In recent years, researchers have studied the use of small, tabletop robots as a flexible and scalable user interface for applications such as data visualization [38, 39] and education [40, 80]. Additionally, the movement of these tabletop swarm robots has proven effective for conveying information [21, 70] and expressing fundamental emotions [7, 25, 63]. Since these functions are central to co-speech gestures in human conversations [75], swarm robots should be explored as a potential interface for generating co-speech gestures in human-robot interactions. Moreover, Kim et al. [35] found that human observers can reliably and quickly interpret the intent of a robotic swarm's collective movement with just a glance. This further supports the use of swarm robots for creating co-speech gestures, as users do not need to fully focus on the swarm robots to grasp their gestures, similar to how people primarily concentrate on speech during conversations with other humans [75].

Nevertheless, the current literature has not considered the implementation of co-speech gestures in swarm robots. The closest related study is Ichihashi et al. [24], which examined how a swarm of tabletop robots functioning as a hand affects the sense of embodiment in the user controlling the swarm. Although this study did not explore how the swarm hand could generate different co-speech gestures, it supports our research by validating the idea of using a group of swarm robots as a body part to achieve various objectives — in our case, gesture generation.

2.4 Animation as Inspiration for Designing Co-speech Gestures in Swarm Robots

To further support the use of swarm robots for co-speech gestures, we consider how these gestures might be realized. While a collective group of swarm robots could mimic a human hand to replicate gestures, this approach would under-utilize the scalability and flexibility of swarm robots [33]. Theoretically, swarm robots, with their numerous units and degrees of freedom, could form a wider variety of shapes and execute more complex movements than a human hand or arm [14, 38]. Therefore, exploring methods beyond merely mimicking human movements is essential to fully leverage the unique capabilities of swarm robots for co-speech gestures.

One source of inspiration for designing co-speech gestures in swarm robots is animation. In the past, animation principles have influenced improvements in robot behavior design. For example, Takamaya et al. [72] found that incorporating animation techniques like anticipation and reaction can make robot behaviors more understandable. Another relevant study is Heider & Simmel [22], which found that people interpret simple movements of geometric shapes as actions of animate beings, often attributing personalities and motives to them. Citing such human tendency to interpret movements as emotional, intentional, and social, Hoffman & Ju [23] suggested that non-humanoid robot designs should consider movement as a key element. Furthermore, Erel et al. [16] even showed that people automatically interpret any robotic movements as social cues, regardless of whether the robot has a social role. These principles underlie the potential use of swarm robots for expressing co-speech gestures, with their movements and formations conveying semantic information that aligns with speech [63, 70].

3 ELICITATION STUDY

While animation principles can help envision how co-speech gestures with swarm robots might look, it remains unclear what specific gestures would accompany different types of speech. For example, designing a gesture for the speech "It is under the table" could involve various movements, such as moving a group of robots downward or having them form a table with one robot placed underneath. One solution to this uncertainty is to hire expert animators to design gestures for specific speech instances, though this may not reflect the preferences of most users [59]. Alternatively, we can use elicitation methods [1, 45, 51, 54] from gesture design, as popularized by Wobbrock et al. [77] study on touchscreen controls, where participants were shown the result of a gesture on a touchscreen and then instructed to perform the gesture they believed would create that outcome. Kim et al. [32] employed a similar method to develop user-defined gestures for controlling swarms of robots. We adapted this elicitation approach for the current research by showing participants various speech instances and asking them to design movements and formations for swarm robots as co-speech gestures, which we then recorded and analyzed to identify design insights and develop an initial set of swarm-based co-speech gestures.

3.1 Hypotheses

In addition to collecting the co-speech gestures that participants generated for different speech instances, we hypothesized that the content of these instances would influence the patterns of the generated gestures. Specifically, we categorized the speech content into four semantic types - action (verbs), object (nouns), relation (prepositions), and characteristic (adjectives) - then examined two dependent variables: the average preferred speed and quantity of robots used in the generated gestures. These two characteristics were chosen because previous research on swarm user interfaces has shown that the number of robots and their speed can greatly influence how users interact with a robot swarm [32, 33]. The detailed hypotheses are as follows:

***H1**: Different semantic types will result in generated co-speech gestures with significantly different preferred speeds.*
We anticipated that semantic types would significantly influence participants' preferences for the average speed of swarm robots. This expectation is based on the notion that different speech content evokes distinct contexts, requiring different types of movements as gestures [75]. If this hypothesis is correct, future swarm-based conversational agents will need hardware that supports a broad range of speeds.

***H2**: Semantic types do not significantly impact preferences for the number of robots used in generated co-speech gestures.*
We hypothesized that the quantity of swarm robots would not vary significantly across different semantic types. This assumption is based on the idea that participants would prefer a consistent number of robots for their gestures, regardless of the speech content. Additionally, Podevijn et al. [55] found that a larger number of robots elicits a stronger physiological response from users. Therefore, if this hypothesis is correct, future swarm-based conversational agents can simplify their hardware requirements by using a uniform number of robots for gesture generation.

3.2 Methods

Adapting the methods described in Kim et al. [32], we presented participants with an audio recording of a referent word, simulating the speech of a swarm robot system. Participants then brainstormed and generated movements and formations that the swarm robots could use as co-speech gestures for the given word.

3.2.1 Apparatus.

For the swarm robots, we used Toio, a miniature multi-robot platform developed by Sony Corporation [69]. The setup included 10 Toio robots (Figure 2.B), though participants could use any number of robots, up to 10, for each gesture. We employed a $40 \times 40\text{cm}^2$ tracking mat as a dedicated area where participants could manipulate the robots by hand (Figure 2.A). To simulate a speaking robotic system, we projected the audio of each speech instance, generated using Google Cloud Text-to-Speech AI [20], from a laptop positioned near the mat (Figure 2.C).

To ensure that participants understood the context of the speech instances — single words in our study — we displayed their definitions and parts of speech (verb, noun, adjective, preposition) on a 27-inch monitor positioned behind the mat. Additionally, to accurately capture the gestures that participants created, we recorded their physical manipulations and verbal explanations with a video camera mounted on a tripod above the mat (Figure 2).

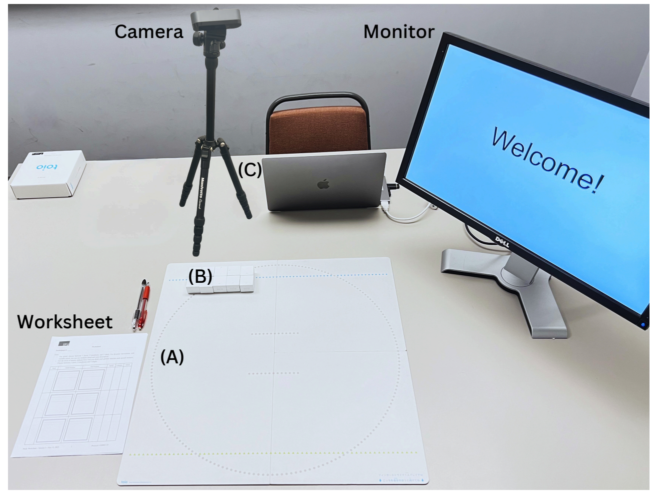


Figure 2: Setup for the elicitation study. Participants sat in front of a tracking mat (A) and manually manipulated 10 Toio robots (B). Audio for single-word speech instances was projected from a laptop near the mat (C).

3.2.2 Word Selection.

For the speech instances, we selected 40 English words, divided into four semantic types - action, object, characteristic, and relation - with 10 words in each type. We chose individual words rather than phrases or sentences for this study to establish a foundational gesture set for common semantic content since these words could be used to build longer phrases and sentences. To ensure that the generated gestures are broadly applicable, we selected words that are frequently used in spoken language and have distinct meanings.

We achieved this by analyzing the top 200 most frequently spoken words in each category (verb, noun, adjective, preposition) from the Corpus of Contemporary American English [11]. To handle words with similar meanings (e.g., "home" and "house"), we grouped these words using a pre-trained word2vec model [48] and K-means algorithm into 10 distinct clusters. We then manually chose one word from each cluster so that each word had a unique meaning. Table 1 shows the final set of referent words.

Table 1: We selected 40 referent words for the elicitation study, categorized into four semantic types with 10 words representing each type.

Verbs (Actions)	**Nouns** (Objects)	**Adjectives** (Characteristics)	**Prepositions** (Relations)
want	world	late	around
help	home	wrong	between
thank	money	best	into
go	time	low	through
think	family	serious	versus
grow	police	big	until
run	people	medical	after
speak	question	difficult	below
write	health	beautiful	towards
join	book	different	within

3.2.3 Participants.

20 participants (8 males, 12 females) aged 18 to 35 years ($M = 21.5 \pm 0.97$) were recruited from our institution. All participants were fluent or native English speakers and reported little to no prior experience with swarm robotics. Each participant provided informed consent, and the study was approved by the institutional review board.

3.2.4 Procedure.

Participants were first briefed on the study's background and purpose of designing co-speech gestures for swarm robots. They then interacted with the Toio robots to familiarize themselves with their weight and shape. We demonstrated the robots' movement speeds — slow (100 mm/s), medium (200 mm/s), and fast (300 mm/s) — by showing a robot moving in a straight line at each speed. Following this, participants were introduced to the study procedure.

In each trial, participants listened to an audio recording of a single-word speech instance and viewed the word, its definition, and its part-of-speech tag on the monitor display. For instance, if the word "home" was played, the display showed "home," its definition ("The place where a person or animal dwells"), and its part-of-speech tag ("n" for noun). The screen remained up for two minutes, during which participants brainstormed a co-speech gesture for the swarm robot that matched the presented word. This involved physically manipulating the swarm robots within the designated tracking mat while describing and explaining the intended gestures. We recorded these gestures, descriptions, and explanations with the video camera. Once the two minutes elapsed, the display automatically switched to indicate the end of the brainstorming period. Participants then specified the preferred robot speed, noted the number of robots used, and drew the start and end positions of each generated gesture on a provided worksheet.

To familiarize themselves with the task, participants began with a practice trial and could repeat this practice as needed. In the main trials, each participant worked with 20 words. The order and selection of these words were randomly balanced for each participant, ensuring that each word received 10 gestures once all 20 participants completed the study. The study's randomization, timing, and automatic procedures were programmed with JsPsych, a framework for running behavioral studies [12]. After finishing the 20 trials, each participant filled out a brief post-study survey, which included questions about their task strategies, encountered challenges, recommendations for the robots' appearance (colors, shape, and size), and any additional suggestions or concerns.

3.3 Analysis

Based on the methodologies outlined in Kim et al. [32] and Wobbrock et al. [77], we developed a systematic taxonomy to analyze qualitative data from video recordings, worksheets, and survey responses. After validating the reliability of our coding scheme, we calculated the agreement score for each generated gesture and extracted other statistics.

3.3.1 Taxonomy.

As shown in Table 2, we categorize each co-speech gesture generated by participants based on its function, group characteristics, and individual characteristics as follows:

1. **Function:** The function of a gesture describes its role in communication. McNeill [47] identified four categories: iconic (representing objects or actions), metaphoric (conveying metaphors), abstract (representing abstract concepts), and deictic (pointing). Additionally, Kendon [27] included a fifth category of symbolic gestures.
2. **Group Characteristics:** When designing gestures for swarm robots, we anticipated that participants would group some robots to behave collectively rather than independently. Based on this assumption, we counted the number of robot groups used in each gesture and recorded their collective formations, which included lines, simple shapes (e.g., squares, circles, triangles, rectangles), and complex symbols (e.g., hearts, S-shapes, arrows). We also noted the collective movements of each group, which could involve transitioning from one location to another, repetitive or back-and-forth motion, or tracing a shape.
3. **Individual Characteristics:** For each individual robot, we considered two types of movements: external movements involve relocating from one position to another, including straight or curved paths, while internal movements involve stationary actions such as jittering, oscillating, or spinning in place. Additionally, we recorded the number of individual robots used in each gesture.

3.3.2 Reliability.

To ensure consistency in applying the taxonomy for analyzing each gesture, all three authors first discussed and jointly coded 10 of the 400 recorded videos. Next, two authors independently coded 20 videos from four different participants. We calculated the unweighted Cohen's Kappa for seven items to assess inter-rater reliability, which yielded $\kappa = 0.76$ with a standard deviation of $\sigma = 0.16$. Given the high level of agreement between raters, the remaining videos were divided into two sets, with each set rated by a single rater.

3.3.3 Agreement Score.

To establish a common set of co-speech gestures, we identified gestures with the highest agreement among participants for each referent word. Although participants often generated similar gestures for a given word, these gestures exhibited slight variations. For example, both P3 and P14 used a circular motion for the word "around," but P3 included an additional robot in the center. Due to these subtle differences, finding completely identical gestures was challenging. Instead, we grouped similar gestures based on their taxonomic categories, specifically their function and movements. We then calculated an agreement score for each word to reflect participant consensus, using the formula from Wobbrock et al. [77]:

$$A = \frac{1}{|W|} \sum_{w \in R} \sum_{G_s \subseteq G_w} \left(\frac{|G_s|}{|G_w|} \right)^2$$

In the equation, w is a referent word in the set of all referent words W, G_w is the set of proposed co-speech gestures that accompanies referent word w, and G_s is a subset of similar gestures from G_w.

3.3.4 Statistics.

We reported mean and standard error of agreement score, speed preference, and robot quantity for the generated co-speech gestures.

Table 2: Taxonomic analysis of swarm co-speech gestures generated by participants.

Taxonomy		Categories	Description
Function		Iconic	The gesture depicts objects or actions.
		Metaphoric	The gesture represents a metaphor.
		Abstract	The gesture conveys an abstract concept with an arbitrary connection to the referent word.
		Deictic	The gesture indicates pointing.
		Symbolic	The gesture refers to a symbol.
Group Characteristics	Quantity		The number of swarm robot groups.
	Collective Formation	Singular	The robot group consists of a single robot.
		Line	The robot group forms a line.
		Shape	The robot group forms a basic shape, such as a square, a circle, a triangle, or a rectangle.
		Symbol	The robot group forms a complex symbol, like an arrow, an S-shape, or a dollar sign.
	Collective Movement	Stationary	The robot group does not move from its initial position.
		Transform	The robot group moves from one location to another, possibly changing formation.
		Repetitive	The robot group moves repetitively, including back-and-forth motions.
		Shape-tracing	The robot group traces a shape or symbol with its movement.
		Other	The robot group moves in a complex manner that does not fit into the other categories.
Individual Characteristics	Quantity		The number of individual swarm robots.
	External Movement	Stationary	The swarm robot does not move from its initial position.
		Straight	The swarm robot moves in a straight path.
		Curved/Circular	The swarm robot moves in a curved or circular path.
		Other	The swarm robot moves in a complex manner that does not fit into the other categories.
	Internal Movement	Stationary	The swarm robot does not have any internal movements.
		Spin	The swarm robot rotates in place.
		Jitter	The swarm robot performs small, rapid back-and-forth movements over a short distance.
		Oscillation	The swarm robot sways in place rhythmically from side to side.

The distribution of each taxonomy category across all gestures was also computed. For hypotheses H1 and H2, we conducted Kruskal-Wallis tests, followed by Bonferroni-corrected post-hoc Dunn's tests, to identify significant differences in participants' speed preferences and the number of robots used across semantic types.

3.4 Results

This section presents a statistical analysis of the agreement scores, taxonomic distribution, and effect of semantic type. We also provide key insights from participants' feedback in the post-study survey and prototype a co-speech gesture set for swarm robots.

3.4.1 Agreement Score.

Figure 3 shows the calculated agreement scores for the generated co-speech gestures across all referent words. The average agreement scores with their standard errors for each semantic type are as follows: $A_{\text{Adjective}} = 0.26 \pm 0.07$, $A_{\text{Noun}} = 0.43 \pm 0.09$, $A_{\text{Preposition}} = 0.37 \pm 0.07$, and $A_{\text{Verb}} = 0.18 \pm 0.03$.

3.4.2 Taxonomic Distribution.

Figure 4 presents the percentage breakdown of the taxonomy for all generated co-speech gestures. For a detailed taxonomic distribution by semantic type, refer to Appendix I. Iconic gestures were the most common (30%), closely followed by metaphoric gestures (29%). The remaining categories included abstract (15%), deictic (13%), and symbolic gestures (13%). Notably, deictic gestures were more prevalent for propositions (31%), likely due to the need for pointing to represent relational concepts. The use of iconic gestures increases to 42% for nouns, likely because these gestures often represent the object the word refers to. Similarly, metaphoric gestures rise to 38% for adjectives, as adjectives are more abstract and harder to convey through simple objects or actions. Regarding the collective formation of robot groups, we found that the distribution is balanced between a single robot (22%), a robot line (28%), basic shapes (19%), and complex symbols (22%).

While we initially anticipated complex movements for co-speech gestures, the analysis revealed that most movements were simple. For robot groups, aside from being stationary, the most common collective movement was transform (37%), where robot groups moved between locations and sometimes changed their formation. Some group movements also displayed back-and-forth (15%) and shape-tracing (6%) behaviors. Individual robot movements were also straightforward, with most external movements being either straight (40%) or curved/circular (17%). Internal movements were rare, with jittering (5%), oscillation (2%), and spinning (1%) being the most common.

3.4.3 Effect of Semantic Type.

There was a significant difference in preferred speed among gestures based on semantic types ($p^{***} < 0.001$), as shown in Figure 5.A. Verbs ($M = 2.04 \pm 0.08$) and prepositions ($M = 1.99 \pm 0.08$) elicited faster speeds compared to nouns ($M = 1.57 \pm 0.10$, $p^{**} = 0.003$ and $p^{*} = 0.011$ respectively) and adjectives ($M = 1.59 \pm 0.10$, $p^{*} = 0.015$ and $p^{*} = 0.04$ respectively). However, no significant differences were found between verbs and prepositions ($p = 0.75$), or between nouns and adjectives ($p = 0.66$). This supports hypothesis H1, suggesting that speech instances with different content require varying swarm gesture speeds.

Regarding the number of robots (Figure 5.B), while there is significant variation across semantic types ($p^{***} < 0.001$), contradicting H2, post-hoc tests showed that nouns ($M = 8.78 \pm 0.19$) prompted the use of more robots compared to all other groups ($p^{*}_{A} = 0.012, p^{**}_{P} = 0.002, p^{**}_{V} = 0.006$). However, excluding nouns, no significant differences were observed across semantic types ($p = 0.88$). This can be attributed to the observation that gestures accompanying nouns are more likely to be iconic (41%), which tend to require a higher number of robots.

No significant differences were found in participants' judgments of speech-gesture matching ($p = 0.09$) or the number of robot groups used ($p = 0.51$) across different semantic types. On average,

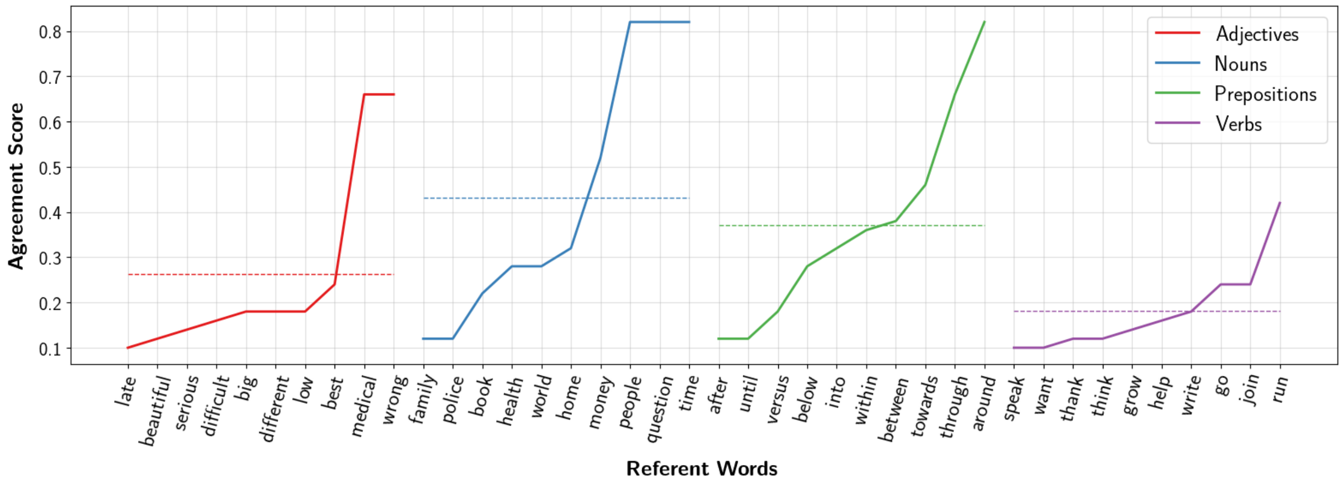


Figure 3: Agreement scores for the generated co-speech gestures are presented for each referent word, categorized by semantic type. Within each semantic type, words are sorted from the lowest to the highest agreement score. The dotted line represents the average agreement score for each semantic type.

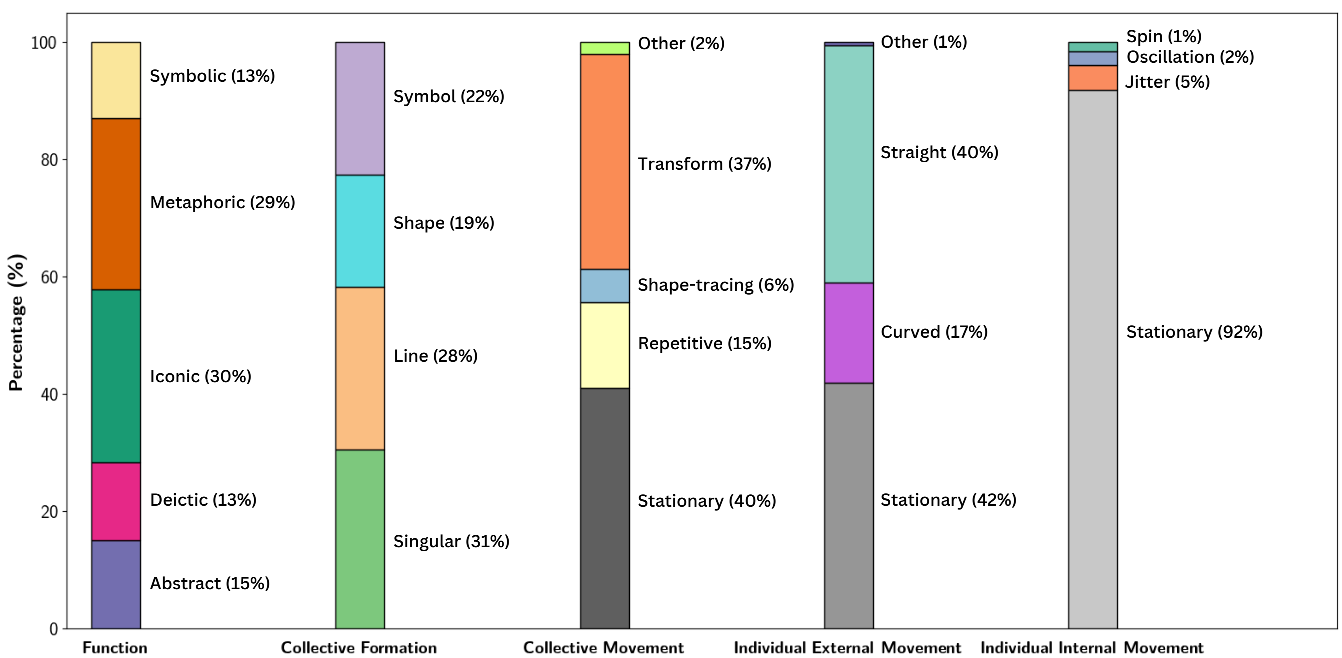


Figure 4: Taxonomic breakdown of the generated co-speech gestures across all conditions.

participants rated the speech-gesture match at 2.03 ± 0.037, suggesting a reasonable alignment between gestures and speech instances. The average number of robots used was 8.00 ± 0.12, with an average of 2.40 ± 0.076 robot groups. The preferred speed averaged 1.80 ± 0.048, indicating a preference for medium to slow speeds.

3.4.4 Post-study Survey.

The most common strategy reported by participants for the brainstorming task was using simple and familiar visual representations of words, adopted by 10 participants. For example, P5 explained, "*My approaches to this were to try to make something visually familiar, such as for the word 'people,' I put 3 robots into lines that would make a human stick figure.*" Similarly, P1 suggested that for the word "best," the optimal approach was "*to just use a pictorial representation (a literal 1).*" The second most common approach, mentioned by 6 participants, was drawing inspiration from human gestures. P9 shared, "*For the word 'towards,' I imagined an arrow*

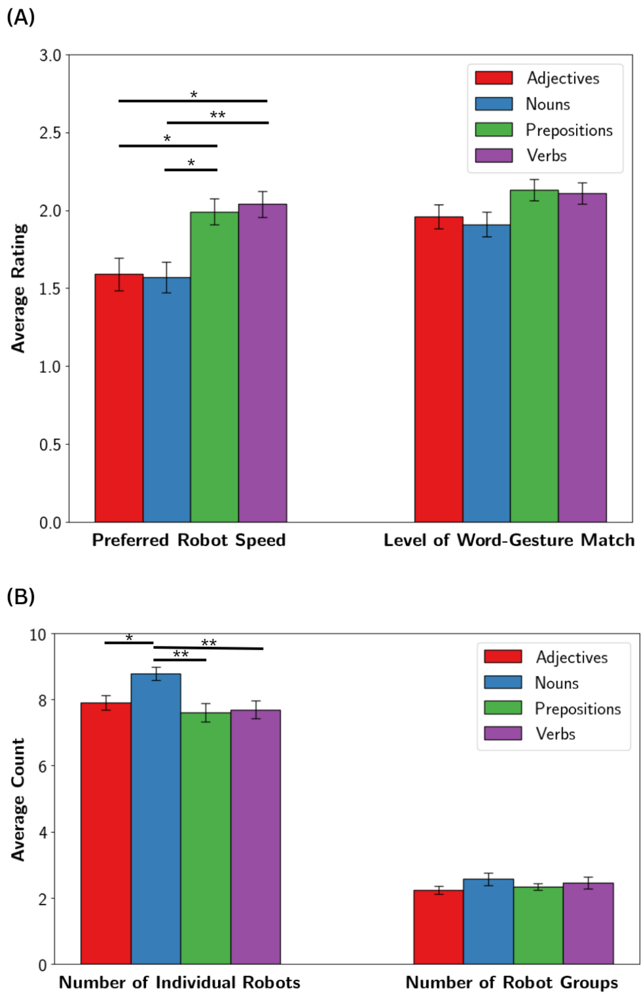


Figure 5: Figure A shows the average participant preference for robot speed (where 0 indicates no movement, 1 is slow, 2 is medium, and 3 is fast) and average participant judgment of how well the brainstormed gesture matches the referent word (where 0 indicates no match, 1 is a low match, 2 is a medium match, and 3 is a high match). Verbs and prepositions elicited significantly faster speeds than nouns ($p < 0.01$, ** and $p < 0.05$, *, respectively) and adjectives (both $p < 0.05$, *). Figure B presents the average number of individual robots and robot groups used in the gestures. The number of individual robots used was significantly higher for nouns than for other semantic types ($p_A < 0.05$, *, other $p < 0.01$, **).

pointing to a single robot, then expanding to surround it, similar to how we use our pointer finger to indicate an area."

11 participants highlighted difficulties in brainstorming gestures for abstract words. For instance, P19 said that "*some words like 'book,' 'think,' and 'beautiful' were hard to create because you normally don't use gestures to describe them.*" Additionally, 7 participants noted challenges related to the limited number of robots, which constrained their ability to create more complex gestures. P20 shared that for the word "money," they envisioned a dollar sign, but the limited number of robots made it difficult to form the desired shape. Nevertheless, 3 participants suggested that the limitation in the number of robots fostered creativity in their brainstorming process, with P13 noting that the constraint "*is a good way to get myself thinking creatively and fast.*"

Regarding the robots' appearance, 9 participants suggested adding color to the robots, either through the robots themselves or via LED lights, as the current robots are all white. P6 highlighted the impact of color, arguing that "*colors could make a big impact for emotions and to stress importance — for medical, a Red Cross would be recognizable and not confused with a white cross from church.*" In addition, 7 participants proposed using circular shapes instead of squares for the robots, as they would allow for greater flexibility in gesture creation. P2 remarked, "*maybe circle would be a better way, more flexible for representing gestures more accurately like pixels.*"

3.4.5 Co-speech Gesture Set.

The co-speech gesture set was developed by selecting referent words with agreement scores around or above the average for each semantic type and choosing the gestures most frequently generated by participants. Moreover, based on participants' preference for simple and familiar visual representations indicated in the post-study survey, we selected the gesture set accordingly. For example, between two gesture options for the word "best" — the number 1 shape or a podium shape with a robot at the top — we chose the former option due to its simplicity. Figure 1 shows two possible gesture designs for the referent word "run," while Figure 6 presents the rest of the gesture set.

4 EVALUATION STUDY

To assess the quality of the co-speech gesture set (Figure 6) generated in the elicitation study, we conducted an online evaluation study within a voice assistant (VA) context. We compared VAs equipped with swarm-based co-speech gestures to those using simple animated movements, similar to indicators on devices like Amazon Alexa and Google Home Assistants [37], to evaluate whether the swarm-based gestures enhance perceptions of the VA's animacy, likability, and intelligence, as well as improve its movement in terms of fluidity, semantic alignment, and temporal synchronization.

4.1 Hypotheses

Previous studies comparing physical and virtual implementations of robots [79] and personal assistants [65] have shown that physical embodiment enhances human perception of these agents, while incorporating co-speech gestures improves the perception of their movements [36]. Drawing on these findings and the current context of evaluating co-speech gestures in physical swarm robots, we formulated the following hypotheses:

***H1**: VAs with swarm-based gestures will be perceived as more animated, likable, and intelligent than animated VAs.*

Similar to prior works comparing physical and virtual agents [65, 79], the current study adapts the Godspeed questionnaire [6]. This hypothesis excluded the anthropomorphism and perceived safety metrics, as the primary goal of implementing co-speech gestures

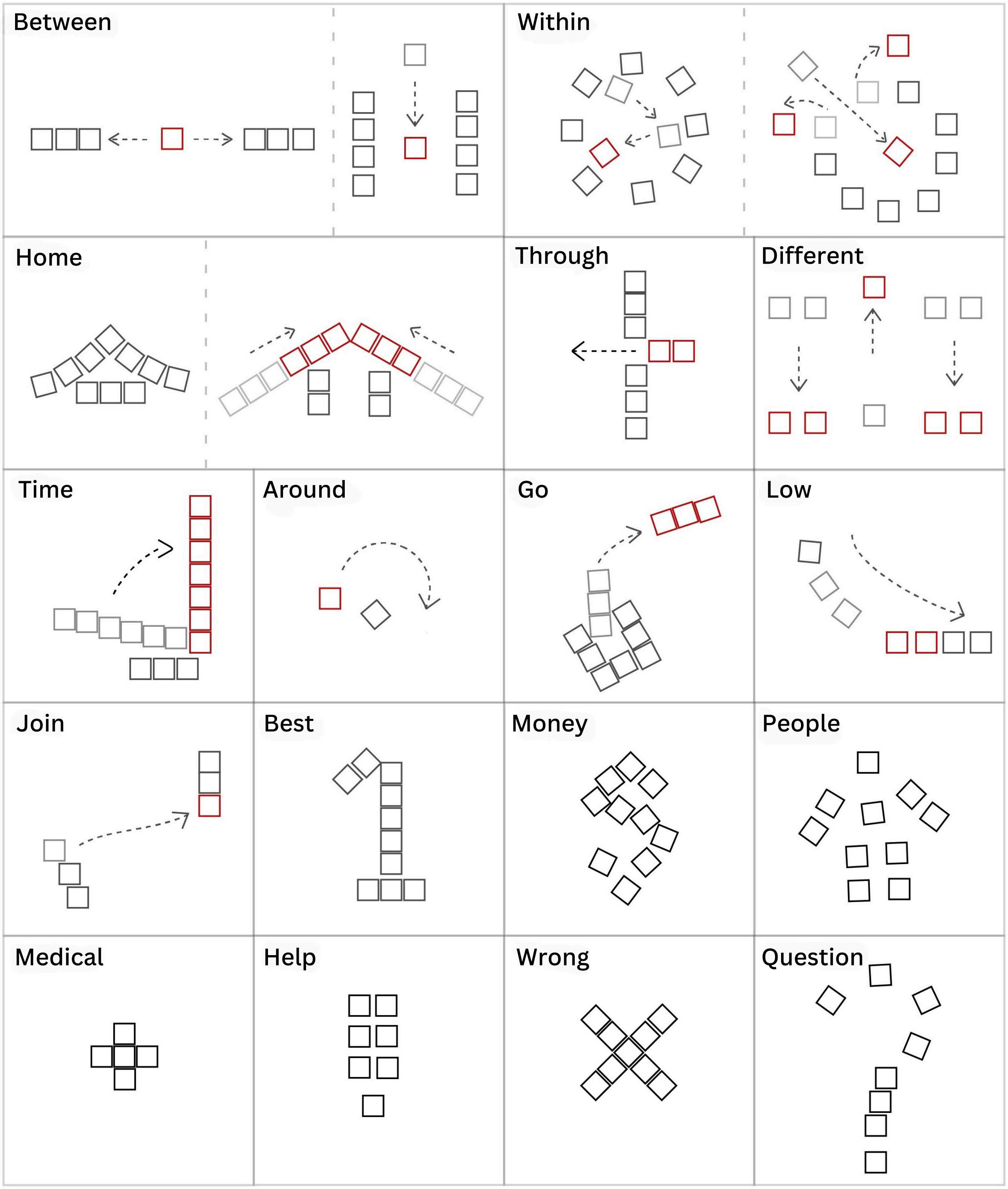


Figure 6: Co-speech gestures for referent words with high agreement scores. For some referent words, multiple co-speech gesture designs are shown, separated by dotted lines. In each design, red squares indicate the robots that have moved from their initial positions. Designs without red squares represent the final formation of the swarm robots after their movements from any previous positions.

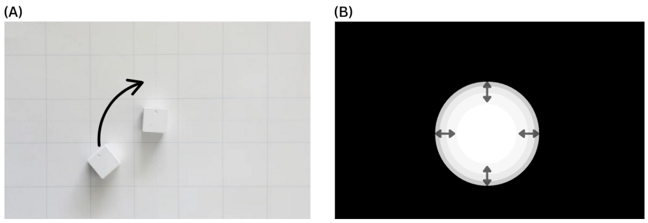


Figure 7: Side-by-side videos of VAs delivering the phrase "Go around the building." In Figure A, the swarm-based VA performs a gesture represented by the curved arrow. In Figure B, the animated VA features a simple contracting and expanding circle that reflects the audio's volume.

in swarm robots is to enhance animacy, likability, and perceived intelligence, rather than the other two aspects.

***H2**: The gestures of swarm-based VAs will be viewed as more fluid, semantically aligned, and temporally synchronized than the movements of animated VAs.*
This hypothesis is based on Salem et al. [61], which found that implementing gestures in physical robots improves human perception of their movements, particularly by enhancing fluidity, semantic alignment of gestures with speech, and synchronization between gestures and audio.

4.2 Methods

We adopted a within-subject methodology inspired by Kurenchenko et al.'s large-scale evaluation of gesture generation models [36]. Participants watched two side-by-side videos of VAs — one featuring the swarm-based VA (Figure 7.A) and the other showing the animated VA (Figure 7.B) — then completed a Likert-scale survey comparing the two VAs across six metrics: animacy, likability, perceived intelligence, fluidity, semanticity, and synchronization.

4.2.1 Participants.

We recruited 30 participants (13 male, 17 female) aged between 20 and 62 years ($M = 34.0 \pm 2.06$) through Prolific [52]. All participants were fluent in English and received CAD$4 for their participation. Each study session took an average of 22.37 ± 1.76 minutes to complete. Participants were required to enable their audio and provide informed consent before participating. The study was approved by our institutional review board.

4.2.2 Stimuli.

The experimental stimuli consisted of 17 pairs of videos, each corresponding to one of the 17 words in our proposed co-speech gesture set (Figure 6). Each video pair features two VAs answering the same question with identical responses but differing in their movements (Figure 7). The question-response pairs were selected to ensure concise answers with a direct word-gesture correlation. For example, in response to "How do I get to the supermarket?" the answer "Go around the building" would trigger a gesture aligned with the word "around." Both the question and response audio were generated using Google Cloud Text-to-Speech AI [20], and the VA movements began only during the response audio. The gestures of the swarm-based VA were implemented using the same Toio robot platform

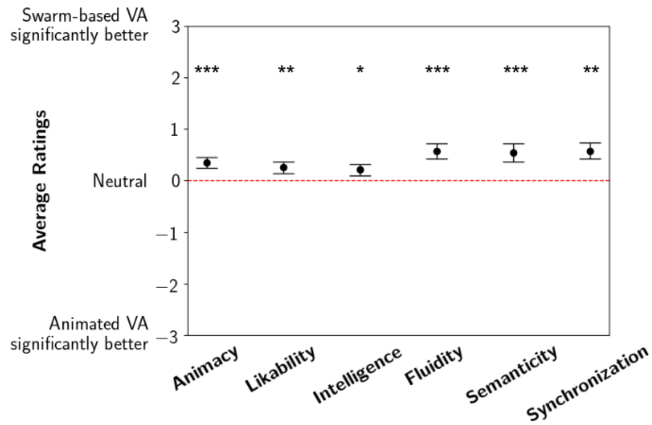


Figure 8: The average comparative ratings indicate that the swarm-based VA is perceived as more animated ($p < 0.001$, *), likable ($p < 0.01$, **), and intelligent ($p < 0.05$, *) than the animated VA. Additionally, its movements are viewed as more fluid ($p < 0.001$, ***), semantically aligned ($p < 0.001$, ***), and synchronized ($p < 0.01$, **).**

[69] as in the elicitation study, while the animated VA's movements were programmed as a solid white circle that expands and contracts in response to the volume magnitude of the audio, similar to the volume indicators found in commercial voice assistants like Amazon Alexa and Google Home [37].

4.2.3 Procedure.

After consenting to participate in the study, participants engaged in 17 trials, each corresponding to a pair of video stimuli. In each trial, participants viewed the video pair and completed a survey comparing the two VAs on six characteristics: animacy, likability, perceived intelligence, fluidity, semanticity, and synchronization. The survey used a 7-point Likert scale, with questions phrased as, "Please assess to what extent the following characteristic applies to the voice assistant/movement of the voice assistant." The response scale ranged from "Voice Assistant A is significantly better than Voice Assistant B" to "Voice Assistant B is significantly better than Voice Assistant A."

The order of VA presentation in the video pairs was balanced with randomization: in 9 pairs, the swarm-based VA was presented as Voice Assistant A and the animated VA as Voice Assistant B, while in 8 pairs, the roles were reversed. The data was processed so that a score of -3 indicates "Animated VA is significantly better than swarm-based VA," 0 indicates "Both VAs are equal," and 3 indicates "Swarm-based VA is significantly better than animated VA."

4.3 Analysis

We used Shapiro-Wilk tests to assess data normality for all six metrics, which revealed that none of the data were normally distributed (all $p^{***} < 0.001$). Consequently, we applied the Wilcoxon signed-rank test to compare participants' ratings on each metric against the neutral midpoint score of 0.

4.4 Results

As shown in Figure 8, we found that the swarm-based VA was perceived to be more animated ($M = 0.39 \pm 0.12$, $p^{***} < 0.001$), likable ($M = 0.31 \pm 0.10$, $p^{**} = 0.0011$), and intelligent ($M = 0.17 \pm 0.10$, $p^{*} = 0.027$) than the animated VA, supporting H1. In addition, we found that the gestures of the swarm-based VA were viewed as more fluid ($M = 0.62 \pm 0.15$, $p^{***} < 0.001$), semantically aligned ($M = 0.55 \pm 0.14$, $p^{***} < 0.001$), and synchronized ($M = 0.45 \pm 0.14$, $p^{**} = 0.0016$) than the animated VA, supporting H2.

5 DISCUSSION

Drawing from the evaluation study's results, we explore potential applications of co-speech gestures in swarm robots and address the current challenges in implementing these applications. Next, we use the elicitation study's findings to propose design insights for the hardware and software requirements of future swarm-based co-speech gesture systems.

5.1 Applications of Co-speech Gestures with Swarm Robots

The evaluation study's findings indicate that the co-speech gesture set developed through the elicitation study enhances human-robot interaction in a voice assistant (VA) context, making the VA appear more animate, likable, and intelligent, with its movements perceived as more fluid, semantically aligned, and synchronized. These results support the use of swarm robots for co-speech gestures and demonstrate a possible real-world application of the technology: making multimodal and interactive voice assistants.

Swarm robots also have potential applications in interactive storytelling and narration for artistic and educational purposes. For example, they could generate co-speech gestures in real-time during storytelling, creating dynamic visual interactions that engage audiences and enhance learning in classrooms or performances. Building on prior research that explored the use of drones for artistic applications [2] and tabletop swarm robots for narration [56] and storytelling [53], a promising research direction is the development of software for tabletop swarm robots to generate co-speech gesture sequences for extended speech. This approach involves leveraging natural language processing techniques to analyze semantic content and rhythm of speech, then create synchronized swarm-based gestures. Advancing these research efforts could unlock innovative applications in immersive storytelling, performative arts, and multimodal interaction.

While the potential applications of co-speech gestures in swarm robots are promising, current hardware and software limit their implementation. For example, programming collective movements on existing swarm robots, such as the Toio [69] used in our study, remains cumbersome, and its hardware cannot perform certain complex movements.

5.2 Hardware & Software Requirements of Co-speech Gesture Systems

Here, we discuss how the results from the elicitation study can guide the development of hardware and software components for future swarm robotic systems capable of rendering co-speech gestures. First, our study reveals that for each swarm-based co-speech gesture, participants typically organize 2 to 3 groups of robots to perform synchronized movements. These movements are relatively simple, as a preference for simple visual representations of words was expressed in the post-study survey. These robot group formations include single robots, lines, simple shapes, or symbols, with collective movements often involving positional transforms, back-and-forth motions, or shape-tracing. Nevertheless, current software for tabletop swarm robots, such as the Toio platform used in this study, primarily supports individual robot movements and lacks capabilities for managing collective formations and synchronized actions. Although research on larger swarm robot systems has explored these capabilities [4], such developments have not yet been applied to smaller tabletop robots, which are better suited for conversational interfaces. Therefore, enhancing software for tabletop swarm robots to support collective formations and synchronized movements is crucial for advancing swarm-based conversational agents. We recommend that future software algorithms for tabletop swarm robots include the following features:

- The ability to arrange a group of robots into a cohesive formation, particularly simple lines and shapes.
- The capability to move the entire robot group from one location to another.
- The option to modify the formation of the robot group.
- The functionality to perform collective movements, such as expanding, contracting, or shape-tracing.

Secondly, participants' preferred speeds for swarm robot gestures vary by semantic type. Therefore, when designing swarm robots for co-speech gestures, it's crucial to equip them with hardware that supports a wide range of speeds. We recommend a maximum speed of 450 mm/s, as Toio's maximum speed of 350 mm/s [69] was insufficient for capturing some complex motions.

We also observed that the number of robots used in each gesture is generally consistent across different semantic types, except for nouns, which require more robots to support iconic gestures. While some participants also noted that a limited number of robots poses significant challenges when creating more complex gestures, this also encourages the generated gestures to be more creative, simpler, and faster. Therefore, we recommend equipping swarm-based gesture systems with at least 10 robots, and the system should be able to move robots in and out of the gesture area to accommodate the necessary number of robots for dynamic co-speech gestures.

Furthermore, the hardware of swarm robots needs improvement to support small internal movements, such as jittering, oscillating, or spinning in place. Although these movements are not required for every gesture, they are helpful for certain cases, particularly when indicating locations, as an alternative to pointing with dedicated hands or fingers in humanoid robots. No tabletop swarm robot systems, including the Toio robots used in our study, can perform all of these subtle movements effectively. Enhancing hardware to support such precise movements would greatly improve the capability of swarm robots to generate more nuanced and effective co-speech gestures.

6 LIMITATIONS & FUTURE WORKS

In the elicitation study, to simplify the task for participants, we limited the referent speech to single-word units rather than longer phrases or sentences. However, longer phrases or sentences might provide additional context that influences gesture generation. Future studies could investigate how participants create co-speech gestures for more complex speech contexts by incorporating longer phrases or sentences.

Another limitation of the elicitation study was the use of white, square-shaped swarm robots, which restricted the generalizability of the findings. In the post-study survey, participants suggested that alternative shapes, such as circles, could inspire more flexible and diverse gesture designs. Participants also proposed incorporating color, either through LED lights or painted surfaces, to enhance the robots' expressive capabilities. Future studies could address these suggestions by experimenting with various robot shapes, sizes, and colors, potentially enabling more complex and varied gestures and leading to richer, more nuanced co-speech interactions.

Finally, the evaluation study is limited by its comparison of swarm-based voice assistants (VAs) solely to a simple animated VA, rather than to a humanoid VA with co-speech gestures. While this choice reflects the current reality that most popular and commercial VAs are represented by simple audio indicators rather than fully embodied humanoid forms [8, 37], we acknowledge this limitation and propose it as a direction for future research, focusing on the effectiveness of swarm-based versus humanoid co-speech gestures across various task contexts.

7 CONCLUSION

The current research presents a user-defined gesture set designed for co-speech interactions with swarm robots and offers valuable insights into the hardware and software advancements needed for seamless real-time gesture generation. As swarm robots become more dynamic and versatile, they have the potential to serve as effective alternatives to humanoid robots in roles such as embodied conversational agents and compact personal assistants. This potential underscores the importance of enhancing swarm robot capabilities to improve user interaction.

ACKNOWLEDGMENTS

We thank the participants for their time and participation. This work is supported by the Canada Foundation for Innovation (CFI) John R. Evans Leaders Fund (JELF) grant no. 44170, and Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant no. RGPIN-2023-04148.

REFERENCES

[1] Parastoo Abtahi, David Y. Zhao, Jane L. E., and James A. Landay. 2017. Drone Near Me: Exploring Touch-Based Human-Drone Interaction. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 3, Article 34 (sep 2017), 8 pages. https://doi.org/10.1145/3130899

[2] Javier Alonso-Mora, Andreas Breitenmoser, Martin Rufli, Roland Siegwart, and Paul Beardsley. 2011. Multi-robot system for artistic pattern formation. In *2011 IEEE International Conference on Robotics and Automation*. 4512–4517. https://doi.org/10.1109/ICRA.2011.5980269

[3] Amir Aly and Adriana Tapus. 2016. Towards an intelligent system for generating an adapted verbal and nonverbal combined behavior in human–robot interaction. *Autonomous Robots* 40, 2 (01 Feb 2016), 193–209. https://doi.org/10.1007/s10514-015-9444-1

[4] Hong an Yang, Yuhua Li, Xin Duan, Gaopan Shen, and Shaohua Zhang. 2022. A parallel shape formation method for swarm robotics. *Robotics and Autonomous Systems* 151 (2022), 104043. https://doi.org/10.1016/j.robot.2022.104043

[5] Lucy Anderson-Bashan, Benny Megidish, Hadas Erel, Iddo Wald, Guy Hoffman, Oren Zuckerman, and Andrey Grishko. 2018. The Greeting Machine: An Abstract Robotic Object for Opening Encounters. In *2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. 595–602. https://doi.org/10.1109/ROMAN.2018.8525516

[6] Christoph Bartneck, Dana Kulić, Elizabeth Croft, and Susana Zoghbi. 2009. Measurement Instruments for the Anthropomorphism, Animacy, Likeability, Perceived Intelligence, and Perceived Safety of Robots. *International Journal of Social Robotics* 1, 1 (01 Jan 2009), 71–81. https://doi.org/10.1007/s12369-008-0001-3

[7] Jaden Berger, Alexandra Bacula, and Heather Knight. 2021. Exploring Communicatory Gestures for Simple Multi-robot Systems. In *Social Robotics*, Haizhou Li, Shuzhi Sam Ge, Yan Wu, Agnieszka Wykowska, Hongsheng He, Xiaorui Liu, Dongyu Li, and Jairo Perez-Osorio (Eds.). Springer International Publishing, Cham, 819–823.

[8] Michael Bonfert, Nima Zargham, Florian Saade, Robert Porzel, and Rainer Malaka. 2021. An Evaluation of Visual Embodiment for Voice Assistants on Smart Displays. In *Proceedings of the 3rd Conference on Conversational User Interfaces* (Bilbao (online), Spain) *(CUI '21)*. Association for Computing Machinery, New York, NY, USA, Article 16, 11 pages. https://doi.org/10.1145/3469595.3469611

[9] Paul Bremner, Anthony G. Pipe, Chris Melhuish, Mike Fraser, and Sriram Subramanian. 2011. The effects of robot-performed co-verbal gesture on listener behaviour. In *2011 11th IEEE-RAS International Conference on Humanoid Robots*. 458–465. https://doi.org/10.1109/Humanoids.2011.6100810

[10] Adi Feiner Guy Hoffman Danielle Rifinski, Hadas Erel and Oren Zuckerman. 2021. Human-human-robot interaction: robotic object's responsive gestures improve interpersonal evaluation in human interaction. *Human–Computer Interaction* 36, 4 (2021), 333–359. https://doi.org/10.1080/07370024.2020.1719839

[11] Mark Davies. 2010. The Corpus of Contemporary American English as the first reliable monitor corpus of English. *Literary and Linguistic Computing* 25, 4 (10 2010), 447–464. https://doi.org/10.1093/llc/fqq018

[12] Joshua R. de Leeuw. 2015. jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior Research Methods* 47, 1 (01 Mar 2015), 1–12. https://doi.org/10.3758/s13428-014-0458-y

[13] Anthony Steven Dick, Susan Goldin-Meadow, Uri Hasson, Jeremy I. Skipper, and Steven L. Small. 2009. Co-speech gestures influence neural activity in brain regions associated with processing semantic information. *Human Brain Mapping* 30, 11 (2009), 3509–3526. https://doi.org/10.1002/hbm.20774 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/hbm.20774

[14] Marco Dorigo, Guy Theraulaz, and Vito Trianni. 2021. Swarm Robotics: Past, Present, and Future [Point of View]. *Proc. IEEE* 109, 7 (2021), 1152–1165. https://doi.org/10.1109/JPROC.2021.3072740

[15] Starkey Duncan Jr. 1969. Nonverbal communication. *Psychological bulletin* 72, 2 (1969), 118.

[16] Hadas Erel, Tzachi Shem Tov, Yoav Kessler, and Oren Zuckerman. 2019. Robots are Always Social: Robotic Movements are Automatically Interpreted as Social Cues. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI EA '19)*. Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/3290607.3312758

[17] Susan Goldin-Meadow. 1999. The role of gesture in communication and thinking. *Trends in Cognitive Sciences* 3, 11 (1999), 419–429. https://doi.org/10.1016/S1364-6613(99)01397-2

[18] Susan Goldin-Meadow, Martha Wagner Alibali, and R. Breckinridge Church. 1993. Transitions in concept acquisition: Using the hand to read the mind. *Psychological Review* 100, 2 (1993), 279 – 297.

[19] Susan Goldin-Meadow, Howard Nusbaum, Spencer D. Kelly, and Susan Wagner. 2001. Explaining Math: Gesturing Lightens the Load. *Psychological Science* 12, 6 (2001), 516–522. https://doi.org/10.1111/1467-9280.00395 arXiv:https://doi.org/10.1111/1467-9280.00395 PMID: 11760141.

[20] Google Cloud. 2024. *Google Cloud Text-to-Speech*. https://cloud.google.com/text-to-speech.

[21] Kasper Grispino, Damian Lyons, and Truong-Huy D. Nguyen. 2020. Evaluating the Potential of Drone Swarms in Nonverbal HRI Communication. In *2020 IEEE International Conference on Human-Machine Systems (ICHMS)*. 1–6. https://doi.org/10.1109/ICHMS49158.2020.9209467

[22] Fritz Heider and Marianne Simmel. 1944. An Experimental Study of Apparent Behavior. *The American Journal of Psychology* 57, 2 (1944), 243–259. http://www.jstor.org/stable/1416950

[23] Guy Hoffman and Wendy Ju. 2014. Designing robots with movement in mind. *J. Hum.-Robot Interact.* 3, 1 (Feb. 2014), 91–122. https://doi.org/10.5898/JHRI.3.1.Hoffman

[24] Sosuke Ichihashi, So Kuroki, Mai Nishimura, Kazumi Kasaura, Takefumi Hiraki, Kazutoshi Tanaka, and Shigeo Yoshida. 2024. Swarm Body: Embodied Swarm Robots. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 267, 19 pages. https://doi.org/10.1145/3613904.

3642870

[25] Chen Ji, Lawrence Kim, Amori Mikami, Elizabeth Reid, Raquel Robinson, Rebecca Todd, Vasileia Karasavva, Karon Maclean, and Katherine Isbister. 2023. Sharing Feelings via Mini Robot Gestures. In *Proceedings of the 2nd Empathy-Centric Design Workshop* (Hamburg, Germany) *(EMPATHICH '23)*. Association for Computing Machinery, New York, NY, USA, Article 1, 5 pages. https://doi.org/10.1145/3588967.3588968

[26] Inchan Jung, Hankyung Kim, and Youn-kyung Lim. 2021. Understanding How Users Experience the Physiological Expression of Non-humanoid Voice-based Conversational Agent in Healthcare Services. In *Proceedings of the 2021 ACM Designing Interactive Systems Conference* (Virtual Event, USA) *(DIS '21)*. Association for Computing Machinery, New York, NY, USA, 1433–1446. https://doi.org/10.1145/3461778.3462082

[27] Adam Kendon. 1972. Some relationships between body motion and speech. *Studies in dyadic communication* 7, 177 (1972), 90.

[28] Byoungjae Kim, Jiwoo Hong, and Woohun Lee. 2022. Poly: Shape-changing Conversational Agent Helps Identify Multiple Characters in Storytelling. In *Proceedings of the Sixteenth International Conference on Tangible, Embedded, and Embodied Interaction* (Daejeon, Republic of Korea) *(TEI '22)*. Association for Computing Machinery, New York, NY, USA, Article 69, 7 pages. https://doi.org/10.1145/3490149.3505573

[29] Lawrence H Kim. 2020. *Designing In Situ Interaction with Ubiquitous Robots*. Stanford University.

[30] Lawrence H Kim, Veronika Domova, Yuqi Yao, Chien-Ming Huang, Sean Follmer, and Pablo E Paredes. 2022. Robotic presence: the effects of anthropomorphism and robot state on task performance and emotion. *IEEE Robotics and Automation Letters* 7, 3 (2022), 7399–7406.

[31] Lawrence H Kim, Veronika Domova, Yuqi Yao, and Pablo E Paredes. 2022. Effects of a Co-Located Robot and Anthropomorphism on Human Motivation and Emotion across Personality and Gender. In *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 155–162.

[32] Lawrence H. Kim, Daniel S. Drew, Veronika Domova, and Sean Follmer. 2020. User-defined Swarm Robot Control. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376814

[33] Lawrence H. Kim and Sean Follmer. 2017. UbiSwarm: Ubiquitous Robotic Interfaces and Investigation of Abstract Motion as a Display. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 3, Article 66 (sep 2017), 20 pages. https://doi.org/10.1145/3130931

[34] Lawrence H Kim and Sean Follmer. 2019. Swarmhaptics: Haptic display with swarm robots. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–13.

[35] Lawrence H Kim and Sean Follmer. 2021. Generating legible and glanceable swarm robot motion through trajectory, collective behavior, and pre-attentive processing features. *ACM Transactions on Human-Robot Interaction (THRI)* 10, 3 (2021), 1–25.

[36] Taras Kucherenko, Rajmund Nagy, Youngwoo Yoon, Jieyeon Woo, Teodor Nikolov, Mihail Tsakov, and Gustav Eje Henter. 2023. The GENEA Challenge 2023: A large-scale evaluation of gesture generation models in monadic and dyadic settings. In *Proceedings of the 25th International Conference on Multimodal Interaction* (Paris, France) *(ICMI '23)*. Association for Computing Machinery, New York, NY, USA, 792–801. https://doi.org/10.1145/3577190.3616120

[37] Veton Këpuska and Gamal Bohouta. 2018. Next-generation of virtual personal assistants (Microsoft Cortana, Apple Siri, Amazon Alexa and Google Home). In *2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC)*. 99–103. https://doi.org/10.1109/CCWC.2018.8301638

[38] Mathieu Le Goc, Lawrence H. Kim, Ali Parsaei, Jean-Daniel Fekete, Pierre Dragicevic, and Sean Follmer. 2016. Zooids: Building Blocks for Swarm User Interfaces. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology* (Tokyo, Japan) *(UIST '16)*. Association for Computing Machinery, New York, NY, USA, 97–109. https://doi.org/10.1145/2984511.2984547

[39] Mathieu Le Goc, Charles Perin, Sean Follmer, Jean-Daniel Fekete, and Pierre Dragicevic. 2019. Dynamic Composite Data Physicalization Using Wheeled Micro-Robots. *IEEE Transactions on Visualization and Computer Graphics* 25, 1 (jan 2019), 737–747. https://doi.org/10.1109/TVCG.2018.2865159

[40] Jiatong Li, Ryo Suzuki, and Ken Nakagaki. 2023. Physica: Interactive Tangible Physics Simulation based on Tabletop Mobile Robots Towards Explorable Physics Education. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference*. 1485–1499.

[41] Ting-Han Lin, Willa Yunqi Yang, and Ken Nakagaki. 2023. ThrowIO: Actuated TUIs that Facilitate "Throwing and Catching" Spatial Interaction with Overhanging Mobile Wheeled Robots. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–17.

[42] Yu Liu, Gelareh Mohammadi, Yang Song, and Wafa Johal. 2021. Speech-based Gesture Generation for Robots and Embodied Agents: A Scoping Review. In *Proceedings of the 9th International Conference on Human-Agent Interaction* (Virtual Event, Japan) *(HAI '21)*. Association for Computing Machinery, New York, NY, USA, 31–38. https://doi.org/10.1145/3472307.3484167

[43] Jiadi Luo, Veronika Domova, and Lawrence H Kim. 2024. Impact of Multi-Robot Presence and Anthropomorphism on Human Cognition and Emotion. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–15.

[44] Michal Luria, Samantha Reig, Xiang Zhi Tan, Aaron Steinfeld, Jodi Forlizzi, and John Zimmerman. 2019. Re-Embodiment and Co-Embodiment: Exploration of social presence for robots and conversational agents. In *Proceedings of the 2019 on Designing Interactive Systems Conference* (San Diego, CA, USA) *(DIS '19)*. Association for Computing Machinery, New York, NY, USA, 633–644. https://doi.org/10.1145/3322276.3322340

[45] Hirotaka Osawa Mahisorn Wongphati and Michita Imai. 2015. User-defined gestures for controlling primitive motions of an end effector. *Advanced Robotics* 29, 4 (2015), 225–238. https://doi.org/10.1080/01691864.2014.978371 arXiv:https://doi.org/10.1080/01691864.2014.978371

[46] N. Matsumoto, H. Fujii, M. Goan, and M. Okada. 2005. Minimal design strategy for embodied communication agents. In *ROMAN 2005. IEEE International Workshop on Robot and Human Interactive Communication, 2005.* 335–340. https://doi.org/10.1109/ROMAN.2005.1513801

[47] David McNeill. 1995. *Hand and Mind.* De Gruyter Mouton, Berlin, Boston, 351–374. https://doi.org/doi:10.1515/9783110874259.351

[48] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Neural and Information Processing System (NIPS)*. https://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf

[49] Masahiro Mori, Karl F. MacDorman, and Norri Kageki. 2012. The Uncanny Valley [From the Field]. *IEEE Robotics & Automation Magazine* 19, 2 (2012), 98–100. https://doi.org/10.1109/MRA.2012.2192811

[50] S. Nyatsanga, T. Kucherenko, C. Ahuja, G. E. Henter, and M. Neff. 2023. A Comprehensive Review of Data-Driven Co-Speech Gesture Generation. *Computer Graphics Forum* 42, 2 (2023), 569–596. https://doi.org/10.1111/cgf.14776 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/cgf.14776

[51] Mohammad Obaid, Markus Häring, Felix Kistler, René Bühling, and Elisabeth André. 2012. User-Defined Body Gestures for Navigational Control of a Humanoid Robot. In *Social Robotics*, Shuzhi Sam Ge, Oussama Khatib, John-John Cabibihan, Reid Simmons, and Mary-Anne Williams (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 367–377.

[52] Stefan Palan and Christian Schitter. 2018. Prolific.ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance* 17 (2018), 22–27. https://doi.org/10.1016/j.jbef.2017.12.004

[53] Yu PENG, Nan WANG, Yuting DIAO, and Haipeng MI. 2019. Swarm robotics platform for intelligent interaction. *Virtual Reality & Intelligent Hardware* 1, 3 (2019), 316–329. https://doi.org/10.3724/SP.J.2096-5796.2019.0019 Human-computer interactions for virtual reality.

[54] Ekaterina Peshkova and Martin Hitz. 2017. Exploring user-defined gestures to control a group of four UAVs. In *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. 169–174. https://doi.org/10.1109/ROMAN.2017.8172297

[55] Gaëtan Podevijn, Rehan O'Grady, Nithin Mathews, Audrey Gilles, Carole Fantini-Hauwel, and Marco Dorigo. 2016. Investigating the effect of increasing robot group sizes on the human psychophysiological state in the context of human–swarm interaction. *Swarm Intelligence* 10, 3 (01 Sep 2016), 193–210. https://doi.org/10.1007/s11721-016-0124-3

[56] David J. Porfirio, Laura Stegner, Maya Cakmak, Allison Sauppé, Aws Albarghouthi, and Bilge Mutlu. 2021. Figaro: A Tabletop Authoring Environment for Human-Robot Interaction. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 414, 15 pages. https://doi.org/10.1145/3411764.3446864

[57] Viva Sarah Press and Hadas Erel. 2023. Humorous Robotic Behavior as a New Approach to Mitigating Social Awkwardness. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 427, 16 pages. https://doi.org/10.1145/3544548.3580821

[58] Samira Pulatova and Lawrence H Kim. 2024. Co-Designing Programmable Fidgeting Experience with Swarm Robots for Adults with ADHD. In *Proceedings of the 26th International ACM SIGACCESS Conference on Computers and Accessibility*. 1–14.

[59] Jaime Ruiz, Yang Li, and Edward Lank. 2011. User-defined motion gestures for mobile interaction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vancouver, BC, Canada) *(CHI '11)*. Association for Computing Machinery, New York, NY, USA, 197–206. https://doi.org/10.1145/1978942.1978971

[60] Maha Salem, Friederike Eyssel, Katharina Rohlfing, Stefan Kopp, and Frank Joublin. 2011. Effects of Gesture on the Perception of Psychological Anthropomorphism: A Case Study with a Humanoid Robot. In *Social Robotics*, Bilge Mutlu, Christoph Bartneck, Jaap Ham, Vanessa Evers, and Takayuki Kanda (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 31–41.

[61] Maha Salem, Stefan Kopp, Ipke Wachsmuth, Katharina Rohlfing, and Frank Joublin. 2012. Generation and Evaluation of Communicative Robot Gesture. *International Journal of Social Robotics* 4, 2 (01 Apr 2012), 201–217. https://doi.org/10.1007/s12369-011-0124-9

[62] Maha Salem, Katharina Rohlfing, Stefan Kopp, and Frank Joublin. 2011. A friendly gesture: Investigating the effect of multimodal robot behavior in human-robot interaction. In *2011 RO-MAN*. 247–252. https://doi.org/10.1109/ROMAN.2011.6005285

[63] María Santos and Magnus Egerstedt. 2021. From Motions to Emotions: Can the Fundamental Emotions be Expressed in a Robot Swarm? *International Journal of Social Robotics* 13, 4 (01 Jul 2021), 751–764. https://doi.org/10.1007/s12369-020-00665-6

[64] Shane Saunderson and Goldie Nejat. 2019. How Robots Influence Humans: A Survey of Nonverbal Communication in Social Human–Robot Interaction. *International Journal of Social Robotics* 11, 4 (01 Aug 2019), 575–608. https://doi.org/10.1007/s12369-019-00523-0

[65] Eike Schneiders, Eleftherios Papachristos, and Niels van Berkel. 2022. The Effect of Embodied Anthropomorphism of Personal Assistants on User Perceptions. In *Proceedings of the 33rd Australian Conference on Human-Computer Interaction* (Melbourne, VIC, Australia) *(OzCHI '21)*. Association for Computing Machinery, New York, NY, USA, 231–241. https://doi.org/10.1145/3520495.3520503

[66] Syamimi Shamsuddin, Luthffi Idzhar Ismail, Hanafiah Yussof, Nur Ismarrubie Zahari, Saiful Bahari, Hafizan Hashim, and Ahmed Jaffar. 2011. Humanoid robot NAO: Review of control and motion exploration. In *2011 IEEE International Conference on Control System, Computing and Engineering*. 511–516. https://doi.org/10.1109/ICCSCE.2011.6190579

[67] Qiming Shen, Kerstin Dautenhahn, Joe Saunders, and Hatice Kose. 2015. Can Real-Time, Adaptive Human–Robot Motor Coordination Improve Humans' Overall Perception of a Robot? *IEEE Transactions on Autonomous Mental Development* 7, 1 (2015), 52–64. https://doi.org/10.1109/TAMD.2015.2398451

[68] Satoshi Shigemi, Ambarish Goswami, and Prahlad Vadakkepat. 2018. ASIMO and humanoid robot research at Honda. *Humanoid robotics: A reference* 55 (2018), 90.

[69] Sony. 2019. Toio. Robot Toy. https://toio.io

[70] David St-Onge, Florent Levillain, Elisabetta Zibetti, and Giovanni Beltrame. 2019. Collective expression: how robotic swarms convey information with group motion. *Paladyn, Journal of Behavioral Robotics* 10, 1 (2019), 418–435. https://doi.org/doi:10.1515/pjbr-2019-0033

[71] Ryo Suzuki, Clement Zheng, Yasuaki Kakehi, Tom Yeh, Ellen Yi-Luen Do, Mark D. Gross, and Daniel Leithinger. 2019. ShapeBots: Shape-changing Swarm Robots. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology* (New Orleans, LA, USA) *(UIST '19)*. Association for Computing Machinery, New York, NY, USA, 493–505. https://doi.org/10.1145/3332165.3347911

[72] Leila Takayama, Doug Dooley, and Wendy Ju. 2011. Expressing thought: improving robot readability with animation principles. In *Proceedings of the 6th International Conference on Human-Robot Interaction* (Lausanne, Switzerland) *(HRI '11)*. Association for Computing Machinery, New York, NY, USA, 69–76. https://doi.org/10.1145/1957656.1957674

[73] Markus Thaler, Stephan Schlögl, and Aleksander Groth. 2020. Agent vs. Avatar: Comparing Embodied Conversational Agents Concerning Characteristics of the Uncanny Valley. In *2020 IEEE International Conference on Human-Machine Systems (ICHMS)*. 1–6. https://doi.org/10.1109/ICHMS49158.2020.9209539

[74] Elisabeth T. van Dijk, Elena Torta, and Raymond H. Cuijpers. 2013. Effects of Eye Contact and Iconic Gestures on Message Retention in Human-Robot Interaction. *International Journal of Social Robotics* 5, 4 (01 Nov 2013), 491–501. https://doi.org/10.1007/s12369-013-0214-y

[75] Petra Wagner, Zofia Malisz, and Stefan Kopp. 2014. Gesture and speech in interaction: An overview. *Speech Communication* 57 (2014), 209–232. https://doi.org/10.1016/j.specom.2013.09.008

[76] Michael L. Walters, Dag S. Syrdal, Kerstin Dautenhahn, René teBoekhorst, and Kheng Lee Koay. 2008. Avoiding the uncanny valley: robot appearance, personality and consistency of behavior in an attention-seeking home scenario for a robot companion. *Autonomous Robots* 24, 2 (01 Feb 2008), 159–178. https://doi.org/10.1007/s10514-007-9058-3

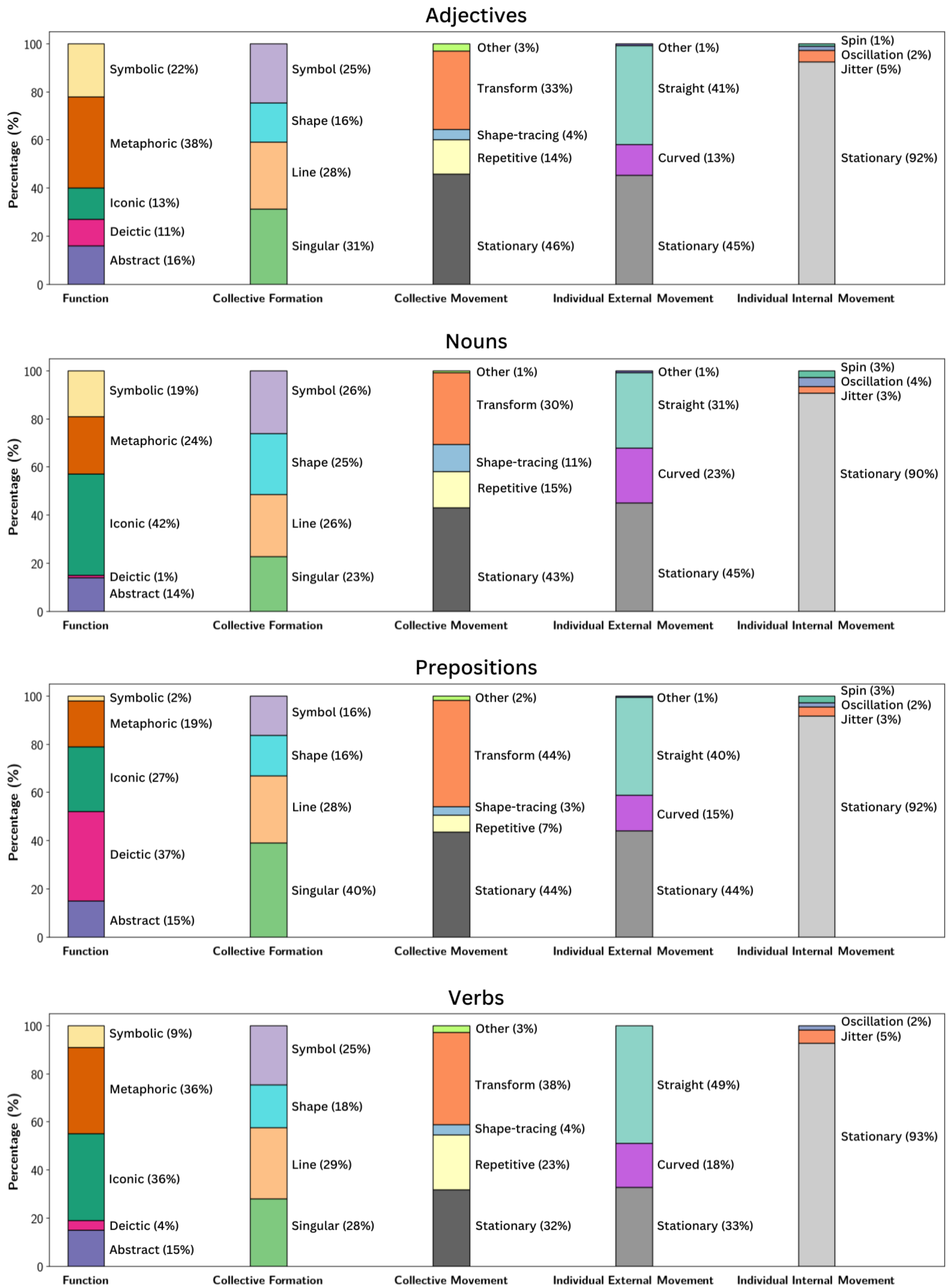
[77] Jacob O. Wobbrock, Meredith Ringel Morris, and Andrew D. Wilson. 2009. User-defined gestures for surface computing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Boston, MA, USA) *(CHI '09)*. Association for Computing Machinery, New York, NY, USA, 1083–1092. https://doi.org/10.1145/1518701.1518866

[78] Lilith Yu, Chenfeng Gao, David Wu, and Ken Nakagaki. 2023. AeroRigUI: Actuated TUIs for Spatial Interaction using Rigging Swarm Robots on Ceilings in Everyday Space. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–18.

[79] Nungduk Yun and Seiji Yamada. 2022. Physical embodiment vs. smartphone: which influences presence and anthropomorphism most in telecommunication?. In *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. 1065–1070. https://doi.org/10.1109/RO-MAN53752.2022.9900758

[80] Ayberk Özgür, Séverin Lemaignan, Wafa Johal, Maria Beltran, Manon Briod, Léa Pereyre, Francesco Mondada, and Pierre Dillenbourg. 2017. Cellulo: Versatile Handheld Robots for Education. In *2017 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI*. 119–127.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009



Appendix I. Taxonometric breakdown of the generated co-speech gestures for different referent word categories.