2025 IEEE World Haptics Conference (WHC)
Suwon Convention Center, Suwon, Korea
July 8 ~ 11, 2025

Simulation-Guided Subset Aggregation for Large-Scale Tacton Similarity Ratings

Chungman Lim
AI Graduate School
Gwangju Institute of Science and Technology
Gwangju, South Korea
chungman.lim@gm.gist.ac.kr

Hasti Seifi
School of Computing and Augmented Intelligence
Arizona State University
Tempe, USA
hasti.seifi@asu.edu

Gunhyuk Park
School of AI Convergence
Gwangju Institute of Science and Technology
Gwangju, South Korea
maharaga@gist.ac.kr

***Abstract*—Exploring perceptual dissimilarity spaces of large-scale Tactons (i.e., Tactile icons) can inform the design of distinguishable haptic feedback. Yet, collecting pairwise similarity ratings for entire Tacton sets becomes costly as set size increases, prompting the need for alternative methods like subset aggregation. Despite previous efforts, little systematic investigation exists on efficient subset size or participant number needed to estimate large-scale Tacton perceptual spaces within a bounded error threshold. We address this gap by introducing a model that simulates between-subject variability in similarity perception. The model explores various distributions under different conditions, including total Tacton numbers and subset-to-total ratios, to guide user studies. Guided by these simulations, we evaluated subset aggregation with three small-scale Tacton sets (12 or 14 patterns) and one large-scale set (48 patterns). Study 1 revealed that initial simulations underestimated real-world variability. We refined the model, ran simulations for larger-scale conditions, and validated them in subsequent studies. The updated model closely matched reality, showing that designers can use our subset aggregation method to prototype perceptual spaces for large-scale Tacton sets. Notably, 4–7 observations were sufficient to achieve $\rho \geq 0.6$, compared to the typical 12 required for generalization. We discuss the efficacy of subset aggregation and future research directions.**

***Index Terms*—Perceptual Dissimilarity Space, Tactile Icon, Distinguishability, Subset Aggregation**

I. Introduction

With the growth in haptic devices over the last decades, Tactons (i.e., Tactile icons) have helped convey information in various applications, such as alerting users to events or system states [1]–[3]. Tactons have also been used with a wide range of haptic devices, such as wearable devices [4]–[6], VR controllers [7]–[9], and smartphones [10], [11], enabling the creation of immersive user interactions and enhancing overall user experiences.

To ensure that Tactons convey clear meanings to users, Tactons must be easily distinguishable. Designers can achieve this goal by creating vibrations through variations in signal parameters, such as amplitude, carrier frequency [12]–[16], or abstract parameters like note length and the evenness of vibration signals (i.e., parameter-based design) [17], [18]. Alternatively, they may use vibration libraries featuring vibrations associated with metaphors, such as a heartbeat (i.e., metaphor-based design) [19]–[22]. After designing an initial set of Tactons, the designers conduct user studies to identify the most distinguishable (i.e., dissimilar) Tactons in the set. The most common approach is pairwise similarity ratings, where users rate all possible pairs of Tactons in the set. Complementary approaches include cluster-sorted ratings, where users group Tactons based on similarity into a predefined number of groups [23]. After user studies are complete, designers calculate a dissimilarity matrix and derive the perceptual dissimilarity space of the Tactons using dimension reduction algorithms such as multi-dimensional scaling (MDS) [24]. Finally, they select the most distinguishable Tactons from the set, ensuring that users can easily differentiate the Tactons and associate them with their corresponding meanings, enabling effective use in target applications.

Despite the usefulness of these approaches, their scalability for large-scale Tactons and the validity of their methodologies present challenges in exploring myriad design parameters within a perceptual space. The time complexity of pairwise similarity ratings is $O(n^2)$, which limits scalability. Similarly, cluster-sorted ratings face issues such as skewed distributions and remain limited in scalability for large-scale Tactons [23]. To address these limitations, several subset aggregation methods have been proposed, using either cluster-sorted ratings or probabilistic models to integrate user ratings from subsets of the entire set [25], [26]. However, these methods lack sufficient validation against gold-standard data, raising concerns about their reliability in real-world scenarios. Recent work proposed a computational model for predicting the perceptual spaces of Tactons [15], but it requires further validation to ensure scalability for large-scale Tactons. While supplementing these models with extensive similarity rating data could enhance their accuracy, the availability of such data in this field remains limited and warrants further collection.

To address these gaps, we proposed a new subset aggregation method guided by computational simulation to derive valid perceptual spaces for large-scale Tacton sets within a bounded error threshold. Building on prior literature [16], [18], we introduced a simulation model for between-subject consistency based on standard deviations in user ratings. We then conducted computational simulations to explore perceptual dissimilarity spaces by addressing the following questions: (1) Can the subset aggregation method for pairwise ratings achieve strong correlations with the ground truth similarity space

under different Tacton distributions, varying the number of Tactons, and subset-to-total ratios? (2) How many observations are needed to achieve a strong correspondence between the aggregated space and the ground truth space?

We ran three experiments to validate the simulation results against the real-world user similarity ratings in online and lab-based experiments. We selected pairwise similarity ratings to evaluate subsets of the entire Tacton set due to their methodological fidelity and suitability for crowdsourced settings [16], [18]. In Study 1, we crowdsourced studies using three small-scale Tacton sets (12 or 14 patterns) by varying Tacton design approaches with subset ratios around 40% to assess the correspondence between simulation results from the proposed simulation model and real-world aggregated ratings for subsets. The study revealed that our model underestimated real-world variability in similarity ratings, likely because it was built using standard deviations in user ratings for the entire set. To address this issue, we introduced additional parameters for the simulation and fitted them to the results of Study 1. Using the revised simulation, we ran Study 2 and 3 on a large-scale Tacton set (48 patterns) and tested smaller subset ratios (10% and 21%). Study 2 established the gold standard for the large-scale set in a controlled lab setting, while Study 3 tested the proposed subset aggregation methods based on updated simulations through online studies. The updated simulations aligned closely with real-world results across subset ratios, achieving a Spearman's rank correlation of $\rho \simeq 0.7$ compared to the gold standard using the number of participants predicted by the updated simulations. Furthermore, the results demonstrated that a strong correspondence could be achieved with fewer observations: four observations were sufficient to reach $\rho = 0.6$, and seven observations yielded $\rho = 0.65$–0.7, compared to the typical requirement of 12 observations for generalization. Overall, the proposed subset aggregation method effectively captured the main structure of the perceptual space and enabled scalable evaluation of the large-scale set, though with some acceptable loss of fine detail. Based on our simulations and study results, we discuss effective subset aggregation methods using pairwise ratings and online studies, and outline future research directions. Our contributions include:

- A method for aggregating pairwise similarity evaluations of subsets in crowdsourced settings, guided by computational simulations, which requires fewer observations than typical approaches and achieves strong correspondence with the gold standard collected in controlled lab settings (i.e., $\rho \geq 0.6$ and the Alienation Coefficient $K \leq 0.2$).
- New pairwise similarity rating data collected from 782 users for 5 Tactons, 250 users for 10 Tactons, and 12 users for 48 Tactons, designed using parameter-based and metaphor-based approaches, as well as their combinations.

II. Related Work

We review past research on design approaches and similarity perception for vibrotactile Tactons.

A. Design Approaches for Tactons

Prior haptic research has proposed various vibration parameters and constructed vibration libraries to convey clear sensations through Tactons, enabling users to learn and associate their meanings effectively [2]. Several studies have focused on sinusoid parameters, or low-level signal parameters, such as amplitude [13], [27], carrier frequency [12], [13], [28], envelope frequency [14], [15], and duration [29]. Other studies explored more complex signal parameters, such as the superposition of two sinusoids [15], [30], [31]. In addition, prior research has investigated abstract parameters, or high-level signal parameters, including rhythmic structures [17], [18], [32], intervals between vibrations [33], and sound waveforms or timbre [34], [35]. To evaluate the effectiveness of the proposed subset aggregation method, we test two Tacton sets designed using this parameter-based approach in Study 1 and create 24 Tactons for Studies 2 and 3. These Tactons use a range of parameters, including amplitude, carrier frequency, envelope frequency, duration, and rhythmic structure, to ensure diverse designs.

In addition to varying vibration parameters, designers can create Tactons by modifying templates from existing vibration libraries to suit specific applications [36]. Many studies have proposed vibration libraries by transforming existing libraries from other sensory modalities into vibration-based libraries or by exploring the semantic and taxonomy spaces of Tactons (i.e., metaphor-based design) [19]–[22]. The metaphor-based Tactons typically feature more complex waveforms than parameter-based Tactons. These include intricate rhythmic structures in the temporal domain, varying frequency spectra over time, and diverse durations. We test one metaphor-based Tacton set in Study 1. Additionally, to enhance the complexity of vibration design parameters in our large-scale Tacton set and ensure compatibility with iPhones for online studies, we create 24 metaphor-based Tactons for Study 2 and 3 by modifying templates from the only open-source vibration library VibViz [22].

B. Similarity Perception for Tactons

Identifying and selecting distinguishable Tactons requires a sequential design process: (1) conducting user studies to collect similarity ratings and construct a dissimilarity matrix, (2) visualizing the dissimilarity matrix as a perceptual space, and (3) analyzing a single perceptual space or comparing multiple perceptual spaces.

One common approach for investigating similarity perception of Tactons involves having users rate the perceived similarity for all possible pairs of Tactons in a set [13], [15], [16], [18], [30]. While this method provides high-fidelity rating data, as users experience all pairs, its time complexity is $O(n^2)$, making it impractical for larger Tacton sets. Cluster-sorted ratings offer an alternative, where users group perceptually similar Tactons into a specific number of groups [23]. Similarity scores are then calculated by summing the weighted occurrences of Tactons being grouped together, which are used to derive a dissimilarity matrix. This approach allows designers

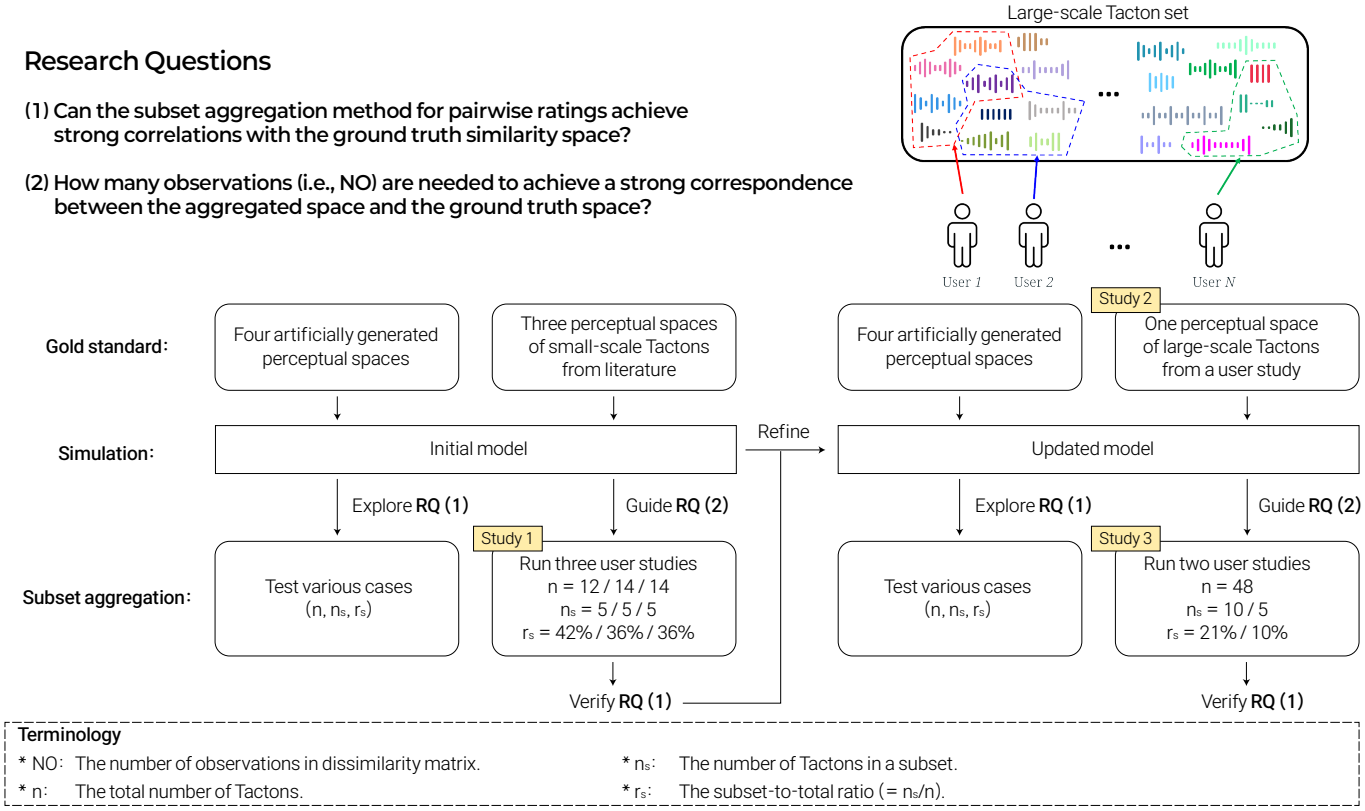


Fig. 1: An overview of our subset aggregation method utilizing pairwise similarity ratings in crowdsourced settings, guided by the proposed simulations.

to investigate perceptual spaces of larger-scale Tactons more easily than the pairwise rating method. However, the resulting dissimilarity scores are often skewed toward "totally different", and its time complexity still converges to $O(n^2)$.

Due to the scalability challenges and time complexity limitations of these methods, prior research proposed a subset aggregation method using cluster-sorted ratings to constrain time complexity to $O(n_s^2)$, where n_s is the number of Tactons in a subset [25]. In this study, each participant evaluated 50 Tactons using cluster-sorted ratings from a set of 84 Tactons, and the experimenters aggregated the dissimilarity matrices from 17 participants. While this approach improved scalability, it exhibited a low correlation between the dissimilarity matrix and the gold standard ($\rho = 0.15$), despite successfully deriving a statistically similar perceptual space ($K = 0.45$). Here, Spearman's rank correlation (ρ) [16], [18], [37], [38] and the Alienation Coefficient (K) [25], [39] are the most commonly used metrics to compare corresponding dissimilarity matrices or perceptual spaces. Recent work introduced an active sampling strategy as a complementary method to traditional approaches [26]. In this method, users cluster Tactons into any number of groups (at least two), and a probabilistic model is applied to derive a dissimilarity matrix. By actively presenting subsets to 129 users, this approach achieved an empirical time complexity of $O(n \log n)$ and enabled the construction of perceptual spaces for up to 252 Tactons. However, the design space in this study was limited to rhythmic structures, and its validity was uncertain due to the absence of comparison with gold standard data. To overcome these limitations, we propose a subset aggregation method using pairwise ratings collected in crowdsourced settings, guided by simulations, to achieve strong correlation with ground truth and bounded time complexity ($O(n_s^2)$). Our approach produces dissimilarity matrices and perceptual spaces that exhibit high correspondence with the gold standard (i.e., $\rho \geq 0.6$ and $K \leq 0.2$), offering a scalable and valid solution to collect similarity perception for large-scale Tacton sets.

Recent studies validated the feasibility of crowdsourcing pairwise similarity ratings by comparing Spearman's ρ values between corresponding dissimilarity matrices [16], [18]. Other research employed both ρ and the Alienation Coefficient (K) to compare multiple perceptual spaces [15]. Following these approaches, we use both metrics to assess the similarity ratings derived from the proposed subset aggregation method against those obtained from the traditional evaluations of entire Tacton sets. Additionally, we crowdsource pairwise similarity rating studies to take advantage of the parallel processing capabilities and cost efficiency of online platforms.

III. Initial Simulation

To explore the potential of our method for aggregating pairwise similarity ratings of subsets, we propose a model for simulating between-subject consistency based on the standard deviations of user ratings (Figure 1). We first artificially

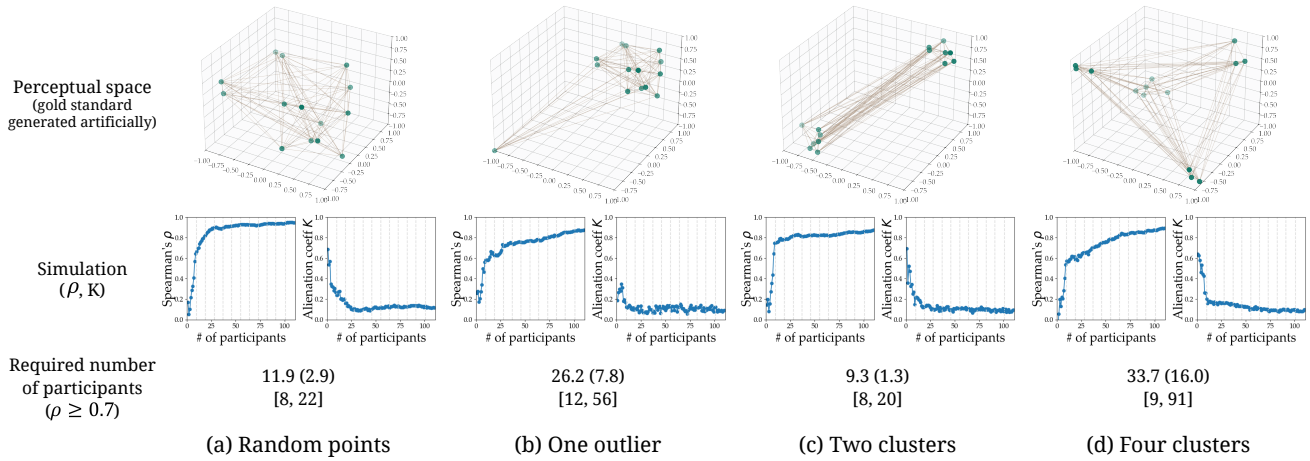


Fig. 2: The four artificially generated perceptual dissimilarity spaces for Tactons, along with Spearman's ρ and the Alienation Coefficient K, across 100 simulations when $n = 14$ and $n_s = 5$ ($NO = 12$). Gray vertical lines (middle row) represent each NO. The number of participants (bottom row) indicates the point at which $\rho \geq 0.7$, comparing the ground truth dissimilarity matrix to the aggregated dissimilarity matrix generated by simulations with the proposed models. The Kruskal's stress values for all perceptual spaces are below 0.1, suggesting a good fit.

generated four distributions of Tactons in the perceptual space and tested various the total number of Tactons (n), the number of Tactons in a subset (n_s), and r_s ($= \frac{n_s}{n}$) to evaluate whether our proposed method—aggregating pairwise similarity ratings for subsets—works consistently across different distributions (Figure 2). Next, we selected three small-scale gold standards from the literature [16], [18], including dissimilarity matrices and standard deviations of user ratings (Figure 3). Using these gold standards, we applied our simulation model to further assess the robustness of our subset aggregation approach on real-world data.

A. Simulation Model

Human distinguishability for vibrations varies due to individual differences in tactile sensitivities, cognitive judgment of similarity, or environmental noise, leading to variability even when rating identical Tacton pairs. This variability necessitates recruiting a sufficient number of participants to ensure a robust sample size for collecting reliable data. We modeled this rating variability (R) using a Gaussian distribution with the maximum standard deviation of user ratings for Tacton pairs observed in each gold standard across the three sets (e.g., Figure 3 (b): $\sigma_{max} = 32.22$ on a scale of 0–100), as it is uncertain which pair will exhibit the maximum standard deviation for unseen Tacton sets:

$$R \sim N(0, \sigma_{max}^2) \tag{1}$$

With this noise formulation, we simulated the dissimilarity ratings by individual users and aggregated them into a dissimilarity matrix. To simulate user ratings, we applied R to the dissimilarity matrix derived from the gold standard. The gold standards were either inversely calculated from distances between points in artificially generated perceptual dissimilarity spaces or sourced from prior literature [16], [18]. In pairwise ratings, designers must determine the number of observations (NO) to ensure the generalizability of the results. Also, if designers assign n_s Tactons to a user, the user compares $n_p = \binom{n_s}{2}$ pairs. To assign these comparisons, we used a non-overlapping sequential sampling method to distribute distinct sets of n_p Tacton pairs across participants as much as possible, using n_s Tactons per participant, with pair overlaps introduced only when necessary to complete the matrix. In the early stages of aggregation, no pair is rated more than once across users at each NO level. However, as the simulation proceeds and some pairs remain unrated, pair overlaps may occur to complete the matrix. The aggregation continues by incrementally increasing NO (i.e., $NO = 1, 2, 3, \dots$). In later stages, more users may be required to cover the remaining unrated pairs, even if some overlap with previously assigned Tactons. The sampled n_p scores were normalized to a scale of 100, representing the maximum perceptual distance for a user. Next, we added R, sampled from the rating distribution defined in Equation 1. The resulting values were clipped within the range [0, 100], where 0 indicates total similarity and 100 indicates total dissimilarity. These noisy ratings were aggregated with those from other users (each rating a different subset) until the dissimilarity matrix reached the desired number of observations (NO) per pair. For our initial simulation, we used $NO = 12$, as this is typically sufficient for generalizing the results of similarity ratings.

B. Test Cases

We tested various Tacton set sizes n (12, 14, 30, 60, 120, and 240) and subset sizes n_s (5, 10, $\lceil \frac{n}{4} \rceil$, $\lceil \frac{n}{2} \rceil$, and n) using four artificially generated distributions of Tactons and three gold standards from literature (Figure 1). In other words, we aimed to explore various subset-to-total ratio cases ($r_s = \frac{n_s}{n}$). To evaluate these test cases, we used Spearman's Rank Correlation (ρ) to compare dissimilarity matrices, and the Alienation Coefficient (K) to compare perceptual spaces. Each simulation was repeated 100 times for these comparisons.

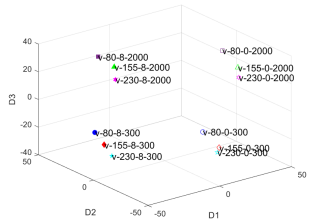
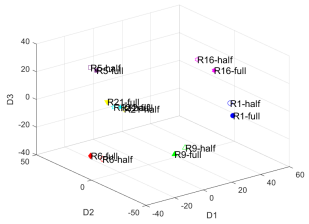
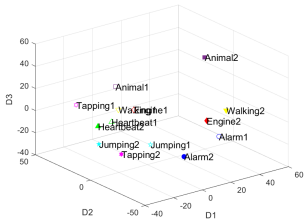
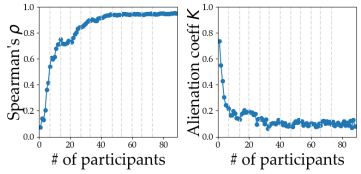
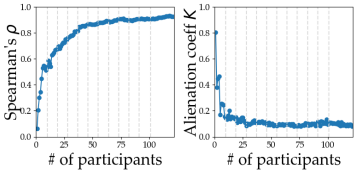
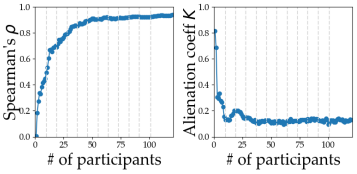
Reference	(a) Kwon et al. 2023 {1}	(b) Abou Chahine et al. 2022	(c) Kwon et al. 2023 {2}
The number of Tactons (n)	12	14	14
σ: 0-100 scale (mean; [min, max])	20.67 [7.64, 32.11]	21.56 [6.80, 32.22]	20.05 [8.69, 27.72]
Design approach	Parameter-based design	Parameter-based design	Metaphor-based design
Vibration parameter	Carrier frequency Envelope frequency Duration	Rhythm Amplitude	Complex waveform
Perceptual space (gold standard from literature)			
Simulation (ρ, K)			
Required number of participants ($\rho \geq 0.7$) (mean (std); [min, max])	11.8 (2.5) [7, 22]	23.0 (5.2) [12, 37]	18.9 (3.2) [13, 35]

Fig. 3: Details of the three gold standards for Tacton sets from literature used in the initial simulation and Study 1. The sources of (a) and (c) are [16], while (b) is from [18]. The Kruskal's stress values for all perceptual spaces are below 0.1, suggesting a good fit. The simulation results include tests for dissimilarity matrices and perceptual spaces, along with Spearman's ρ and the Alienation Coefficient (K), conducted across 100 simulations ($NO = 12$).

Four Distinct Tacton Distributions Generated Artificially: Based on typical Tacton distributions reported in the literature [13], [15], [16], [30] and considering extreme cases, we generated perceptual spaces with 14, 30, 60, 120, and 240 points (i.e., n) across four different distributions of points in the space: (1) random points, (2) one cluster and one outlier, (3) two clusters, and (4) four clusters. Note that while we tested additional configurations, such as one cluster with two outliers or circular layouts, we present only these four representative distributions, as the simulation results for the others fell within the range of these four. For each distribution, we calculated the distances between points to derive dissimilarity matrices and tested the five n_s cases described above. Since these distributions were artificially generated and lacked information on rating variability, we used $R \sim N(0, 32.22^2)$ for simulations, where 32.22 corresponds to the largest σ observed among the three gold standards.

Three Existing Perceptual Spaces of Tacton Sets From Literature: We selected three Tacton sets that provided dissimilarity matrices and standard deviations of user ratings from [16], [18] (Figure 3). The first set consisted of 12 Tactons designed using a parameter-based approach by varying three carrier frequencies (80, 150, and 230 Hz), two envelope frequencies (0 and 8 Hz), and two durations (300 and 2000 ms). The second set with 14 Tactons was also designed using a parameter-based approach and varied on seven rhythmic structures and two amplitude levels on iPhones. The last set consisted of 14 Tactons designed using a metaphor-based approach, varying across seven metaphors with complex waveforms. We used the model (Section III-A) for simulations corresponding to each gold standard.

C. Simulation Results

Simulations on aggregating pairwise similarity ratings of subsets revealed their potential for the four selected distributions of Tactons in perceptual spaces, as well as for the three existing Tacton sets, which varied in design parameters.

Four Distinct Tacton Distributions Generated Artificially: We tested various numbers of points (n= 14, 30, 60, 120, and 240 points) for Tactons in perceptual spaces and five n_s values (5, 10, $\lceil \frac{n}{4} \rceil$, $\lceil \frac{n}{2} \rceil$, and n) for the four distributions. Figure 2 shows the results for 14 points with $n_s = 5$. Across all n and n_s values, correlations between the distributions and subset aggregation methods quickly reached $\rho = 0.6$ with $NO = 2$ for all four cases. For $\rho = 0.7$, the required NO varied depending on the distributions. Specifically, cases with one outlier and four clusters required more participants to achieve $\rho \geq 0.7$. Nevertheless, all distributions consistently achieved $\rho \geq 0.7$ well before reaching $NO = 6$ across the 100 simulations. The Alienation Coefficient (K) reached approximately 0.2 at $NO = 2$ in all cases and maintained a value below 0.2 during the subsequent stages of the subset

aggregation process. A K value below 0.2 indicates that the two perceptual spaces are statistically similar at a 95% confidence interval, regardless of the number of points (n) or dimensions in the perceptual spaces [39].

Three Existing Perceptual Spaces of Tacton Sets From Literature: We tested the proposed simulation model using the three dissimilarity matrices while varying three n_s values (Figure 3). Across all n_s values, correlations between the distributions and the subset aggregation methods quickly reached $\rho = 0.6$ with $NO = 2$ for all cases and achieved $\rho = 0.7$ by approximately $NO = 6$, even in the worst-case scenario. The Alienation Coefficient (K) also dropped to around 0.2 with $NO = 2$ for all cases and remained stable (below 0.2) throughout the rest of the subset aggregation process.

Brief Discussion: The simulations with the proposed models for user rating variability allowed for assessing the potential of aggregating pairwise similarity ratings for subsets of the three gold standards before running user studies. Additionally, by testing various distributions in perceptual spaces, we confirmed that our aggregation method consistently achieves correlations of $\rho \geq 0.7$ and generalizes across diverse distributions and values of n, n_s, and r_s, regardless of the variability in user ratings for vibration perception.

IV. Study 1

Next, we ran Study 1 with three Tacton sets on the Amazon Mechanical Turk (mTurk) online study platform to compare the simulation results with human similarity ratings using the subset aggregation method in practice.

A. Subset design

We used the same Tacton sets as those employed for the three gold standards (Section III-B). These sets consisted of 12, 14, and 14 Tactons, respectively. For all sets, we selected $n_s = 5$, resulting in 10 comparison tasks per user (i.e., $n_p = 10$). The subset-to-total ratios (r_s) for the three sets were 42%, 36%, and 36%, respectively. Although simulation results suggested that this study could achieve $\rho = 0.7$ with 12, 23, and 19 participants (i.e., subsets) on average across 100 simulations, we generated 46, 64, and 65 subsets to obtain $NO = 6$, as the minimum number of observations used in perceptual studies and to achieve $\rho \geq 0.7$ based on simulations. While $\rho \geq 0.6$ indicates strong correspondence [40], we set the target ρ at 0.7 to ensure a stronger relationship between the dissimilarity matrices and better capture the variability in human perception and the complexity of perceptual dissimilarity data in real-world contexts [41].

B. Participants

We recruited 175 mTurk workers, all of whom had completed 10,000 Human Intelligence Tasks (HITs) on the mTurk platform with a success rate greater than 98%. The participants were 20–64 years old (mean age: 36.7 years). The sample included 141 Americans, 19 Indians, 8 Brazilians, 2 Italians, 2 Sri Lankans, 1 Canadian, 1 Dominican, and 1 British. Participants used 25 different smartphone models,

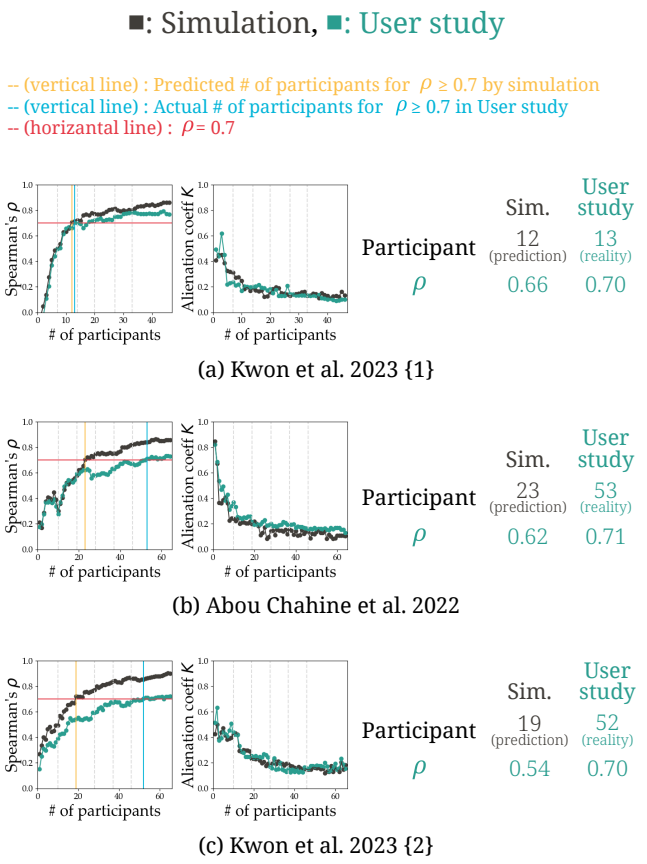


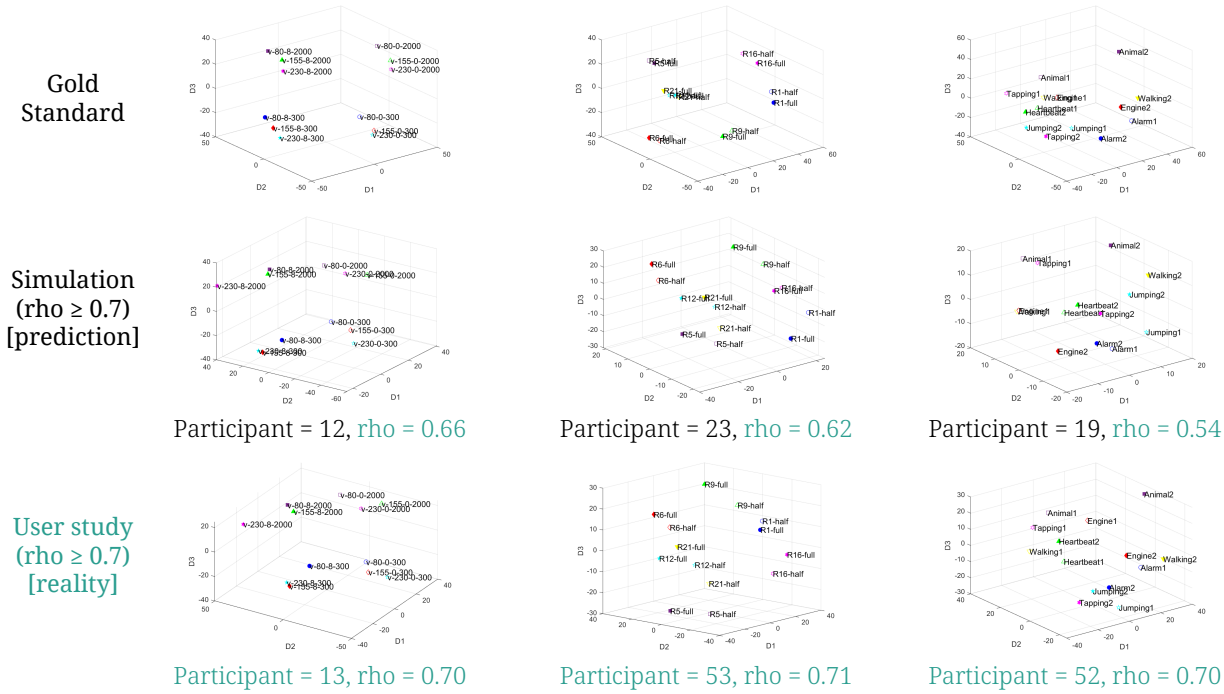
Fig. 4: Graphs showing how the dissimilarity matrices were aggregated as the number of participants increased, with ρ (left) and K (right) plotted for each gold standard. Participant (Sim.) denotes the number of participants at which $\rho \geq 0.7$ is achieved between the simulated predictions and the gold standards. Participant (User study) indicates the number of users at which $\rho \geq 0.7$ is achieved between the real dissimilarity values derived from user studies and the gold standards.

ranging from the iPhone 8 (oldest) to the iPhone 15 Pro (newest). Note that prior research has shown that similarity ratings remain consistent across different iPhone models [16], [18]. Participants completed the rating tasks on the iPhone application in 3–15 minutes (mean: 4.8 minutes) and received $3 USD as compensation.

C. Experiment Procedure

We developed an iPhone application consisting of four sessions: consent, training, main, and feedback. The consent session provided information on the informed consent and instructions for the study procedures. After obtaining consent, the app collected demographics, such as age, biological sex, nationality, and experience with haptic technology. Following prior work on haptic crowdsourcing [16], we specifically instructed participants to remove any phone case, hold the smartphone in their left hand, and interact with the application using their right index finger.

In the training session, the application assigned participants a random subset from one of the 46, 64, or 65 subsets for



(a) Kwon et al. 2023 {1} (b) Abou Chahine et al. 2022 (c) Kwon et al. 2023 {2}

Fig. 5: Perceptual spaces of the gold standards (top row) for three Tacton sets and perceptual spaces derived using the subset aggregation method (middle and bottom rows). The middle row shows perceptual spaces generated using the predicted number of participants from simulations, where $\rho \geq 0.7$ was expected. However, the corresponding ρ values observed in user studies were 0.66, 0.62, and 0.54, respectively, all falling below 0.7. The bottom row shows those generated using the actual number of participants required in user studies to achieve $\rho \geq 0.7$.

the three Tacton sets, respectively. Participants interacted with a set of buttons, each corresponding to a randomly assigned Tacton from their subset, ensuring they experienced all Tactons in the set. In the main session, participants provided similarity ratings for all possible Tacton pairs in their assigned set, using a scale of 0 (totally different) to 100 (totally similar). The application presented Tacton pairs in random order and included one identical Tacton pair as an attention check. During this session, participants could play the Tactons multiple times but were unable to modify previous ratings. Finally, the feedback session collected participants' comments on the experiment. We connected our application to the Google Firebase Realtime Database to store all user responses, including demographics, ratings, and comments.

D. Study 1 Results

For Tacton sets (a) and (b) in Section III-B (Figure 3), both the simulations and user studies showed a similar trend, reaching $\rho \geq 0.6$ and $K \leq 0.2$ with approximately $NO = 2$ (Figure 4). In contrast, Tacton set (c) required more participants compared to sets (a) and (b) to achieve $\rho \geq 0.6$ (approximately $NO = 3$). Notably, the simulations matched well with all three user studies for K. However, the number of participants required to achieve $\rho \geq 0.7$ varied across the Tacton sets (Figure 4). For Tacton set (a), predictions from the simulation closely matched the user study results, with only a one-participant difference (12 participants predicted vs. 13 participants observed) to achieve $\rho \geq 0.7$. For Tacton sets (b) and (c), the simulations predicted $\rho \geq 0.7$ with 23 and 19 participants, respectively, whereas the user studies required 53 and 52 participants.

The perceptual dissimilarity spaces derived using the subset aggregation method closely reflected the main trends of the gold standards (Figure 5). For Tacton Sets (a) and (b), the main structures of the ground truth perceptual spaces were captured when $\rho \geq 0.7$ was achieved (Figure 5, top vs. bottom row). Specifically, for Tacton Set (a), the perceptual space was divided primarily by envelope frequency and duration, forming four clusters. However, the lowest dominant parameter, carrier frequency, did not show clear patterns through subset aggregation, unlike the gold standard, which exhibited consistent patterns across all clusters. For Tacton Set (b), the perceptual space revealed that rhythm dominated over amplitude, although the detailed perceptual distances between rhythms differed from the gold standard. Additionally, the effects of amplitude were slightly overestimated compared to the gold standard. Overall, when $\rho \geq 0.7$ was achieved, the subset aggregation method effectively captured the main structures of the gold standards' perceptual spaces, though some fine details were lost. For complex vibrations in Tacton Set (c), the perceptual space generated by the subset aggregation method closely matched the gold standard, despite minor differences in

detailed distances between Tactons. In all cases, the Alienation Coefficient (K) remained below 0.2, indicating that the subset aggregation method consistently produced perceptual spaces statistically similar to those of the gold standards.

Brief Discussion: For Tacton sets with r_s values of 36% or 42% and n values of 12 or 14, the subset aggregation method quickly converged to a ρ of 0.6 with $NO = 2$. However, the number of participants required to achieve $\rho \geq 0.7$ varied depending on the Tacton sets. We conjecture that Set (a), designed using low-level signal parameters such as carrier frequency and duration, had relatively clear perceptual effects for users. This clarity allowed users to rate these Tactons effectively, even when ratings were divided and aggregated. In contrast, Sets (b) and (c) used high-level parameters, such as rhythmic structure and complex waveform, which are more abstract compared to frequency or duration. This abstraction likely required more participants to achieve $\rho \geq 0.7$. As a result, although the simulation predicted a similar alienation coefficient (K), the initial simulation underestimated the variability in individual ratings observed in real-world settings. This finding suggests the need to improve the initial simulation model for user rating variability, particularly by considering the design approaches used in the Tacton sets. Finally, while aggregating pairwise evaluations for subsets successfully identified the main structure of the perceptual spaces, it did so at the cost of detail. In other words, the method effectively captured dominant design parameters but showed reduced sensitivity to parameters with lower perceptual effects on human perception.

V. Updated Simulation

To improve the initial simulation that used Equation 1 for modeling user rating variability (R), we introduce a new model parameter, w, which adjusts Equation 1 to account for r_s values and the complexity of vibration parameters.

We updated Equation 1 as follows:

$$R \sim N(0, (w \cdot \sigma_{mean})^2) \tag{2}$$

Here, we used $\sigma_{mean} = 21.65$, the mean standard deviation derived from six reference datasets in [16], [18]. To determine w, we fitted it such that the mean number of participants required to achieve $\rho \geq 0.7$ in 100 simulations matched the number of participants observed in Study 1 for each Tacton set. As in the initial simulation, R was sampled from the distribution defined in Equation 2 and clipped to the range [0, 100]. The fitted w values for the three Tacton sets used in Study 1 were 1.6, 2.5, and 2.6, respectively.

We tested the updated simulation against four artificially generated perceptual spaces and the three perceptual spaces from the literature (Section III-B), consistent with the methodology described for the initial simulation (Section III-C). The subset aggregation process for pairwise ratings in the updated simulation closely matched the aggregation of real ratings observed in Study 1 (Figure 6). Correlations between the gold standard and subset aggregation methods quickly reached

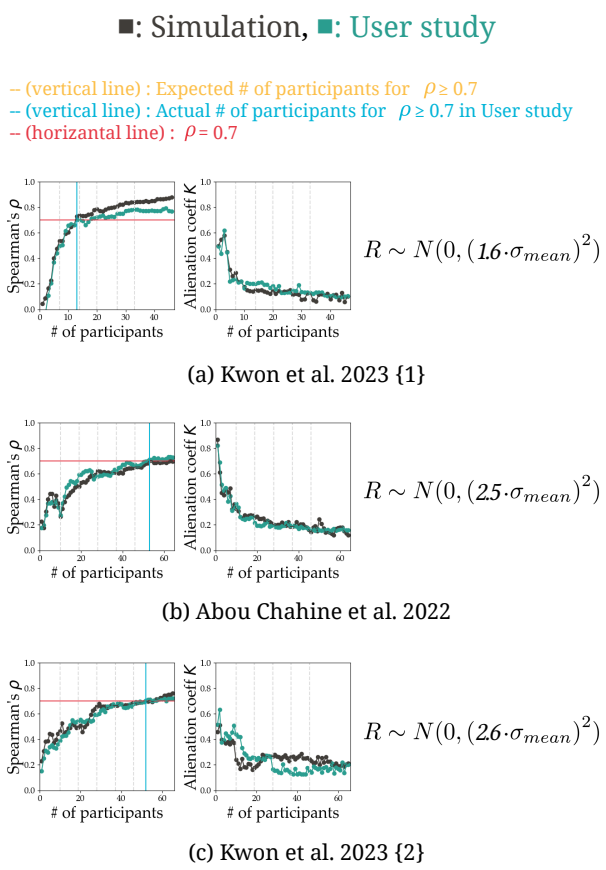


Fig. 6: Updated simulation results based on User Study 1 results for the three Tacton sets.

$\rho = 0.6$ with $NO = 2$ for all three cases, similar to the initial simulation model. However, across different numbers of points in perceptual spaces (14, 30, 60, 120, and 240 points) and n_s values (5, 10, $\lceil \frac{n}{4} \rceil$, $\lceil \frac{n}{2} \rceil$, and n), correlations required slightly more participants to reach $\rho = 0.6$, with $NO = 3$–4 depending on the distribution. Also, achieving $\rho = 0.7$ typically required more participants, with NO values around 6–7 depending on the distributions. Nonetheless, in all cases, the proposed method for aggregating pairwise subjective evaluations consistently reached $\rho = 0.7$ without exceptions. Similar to the initial simulation, the Alienation Coefficient (K) reached approximately 0.2 at $NO = 2$ in all cases and remained below 0.2 throughout the subsequent stages of the subset aggregation process.

Brief Discussion: We interpret Equation 2 such that σ_{mean} represents the inherent similarity rating variability among users (i.e., the subjective nature of subjective similarity evaluations), while w serves as a parameter for aggregating subjective evaluations for subsets of Tactons. At this point, it remains unclear whether w is influenced by the subset-to-total ratio (r_s) or the complexity of the vibration parameters. For instance, in Set (a), $w = 1.6$ with $r_s = 42\%$ and sinusoidal parameters. In Set (b), $w = 2.5$ with $r_s = 36\%$ and abstract parameters, while in Set (c), $w = 2.6$ with $r_s = 36\%$ and complex waveforms. To investigate this further, we fix $w = 2.6$ in Study

3, where Tactons consist of complex waveforms as vibration parameters, and test two lower r_s values (21% and 10%). If the updated simulation predictions align closely with the gold standard results, this would suggest that w is primarily related to the complexity of the vibration parameters. Conversely, if the results vary with the two r_s values, it would indicate that w is influenced by r_s.

VI. Study 2

Study 2 constructs the ground truth dissimilarity space for 48 Tactons using pairwise ratings. This gold standard will be used in Study 3 to compare against the subset aggregation methods.

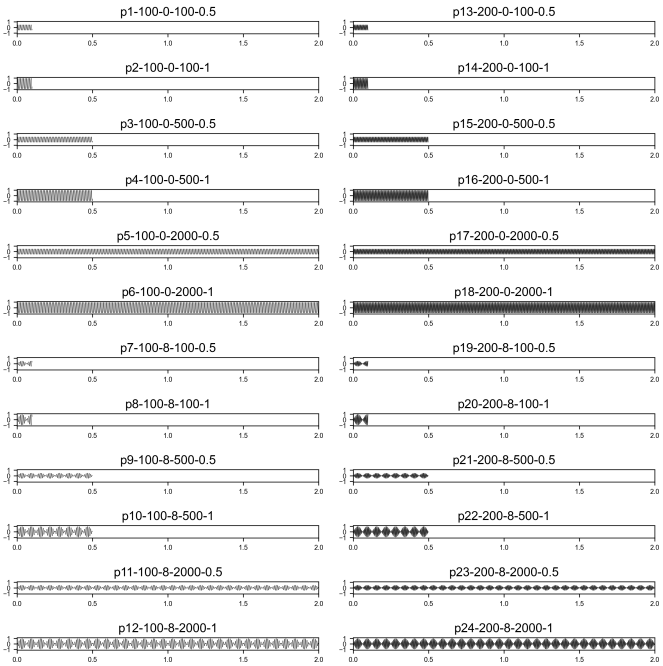
A. Tacton design

We created 48 Tactons using two design approaches based on existing literature [22], [29]: parameter-based and metaphor-based approaches (Figure 7). We chose these methods to enhance the diversity of Tactons within the set. 24 Tactons were designed using the parameter-based approach, varying on two carrier frequencies (100 Hz and 200 Hz), two envelope frequencies (0 Hz and 8 Hz), three durations (100 ms, 500 ms, and 2000 ms), and two amplitudes (*half* and *full*). We selected these parameters to ensure compatibility with iPhones, particularly for the carrier frequency. The vibrations were programmed using Apple's Haptic and Audio Pattern (AHAP) format, a JSON-like file required for defining vibration patterns on iPhones. The AHAP format supports temporal envelopes, which only allow positive values, and carrier frequencies between 80 Hz and 230 Hz. The temporal envelopes for the 24 Tactons were generated using the mathematical formula $E(t) = A \cdot |sin(2\pi f_e t)|$. In this formula, A denotes the amplitude, with *half* and *full* corresponding to 0.5 and 1 in the AHAP format, respectively, while f_e refers to the envelope frequency. When $f_e = 0$ Hz, the envelope becomes constant, simplifying the formula to $E(t) = A$. The carrier frequencies for all Tactons were kept constant at either 100 Hz or 200 Hz.

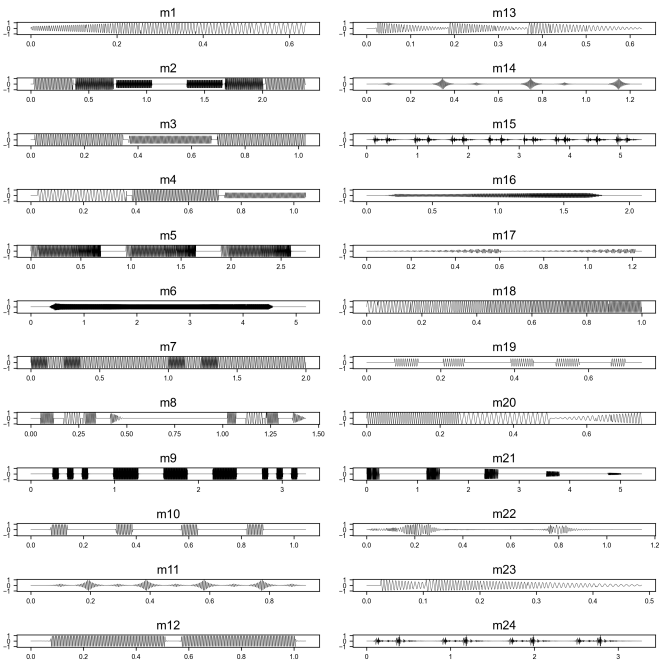
The remaining 24 Tactons were from the VibViz library, which contains metaphor-based Tactons [22]. We chose these Tactons based on their distribution in the original sensory space of the library and adjusted them for compatibility with iPhones. These 24 Tactons featured more complex waveforms compared to the earlier 24 Tactons, which were designed using sinusoidal parameters. The duration of the metaphor-based Tactons ranged from 0.49 to 5.42 seconds. None of the 48 Tactons used in this study overlapped with the Tactons from Study 1.

B. Participants

We recruited 12 participants (10 women and two men; 19–28 years old). All participants were right-handed and reported no impairments in their hands. Rating all possible pairs for 48 Tactons (i.e., 1,128 comparison tasks) required significant time. To manage this, participants completed 551 comparison ratings on the first day and 579 ratings on a subsequent day, with each session including one attention test. Each experiment



(a) Twenty-four vibration Tactons designed using a parameter-based approach for AHAP format. We named Tactons as "p{index}-{carrier frequency}-{envelope frequency}-{total duration}-{amplitude}".



(b) Twenty-four vibration Tactons designed using a metaphor-based approach for AHAP format. We named Tactons as "m{index}".

Fig. 7: Two plots displaying the 48 Tactons used in Studies 2 and 3 (Section VI-A). The x-axis represents time (in seconds), and the y-axis denotes amplitude in the AHAP format. A value of 1.0 corresponds to the maximum acceleration on iPhones.

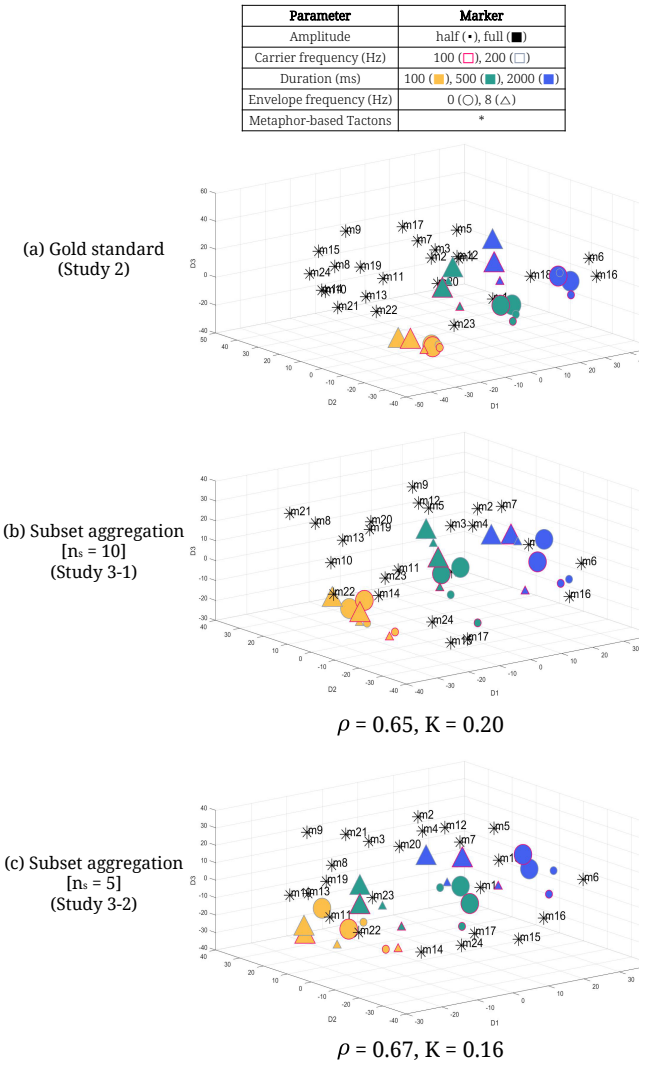


Fig. 8: (a) Perceptual space of the gold standard (Study 2). (b) and (c) Perceptual spaces derived using the subset aggregation methods (Studies 3-1 and 3-2).

session lasted approximately three hours per day (i.e., total six hours). Participants received $76 USD as compensation.

C. Experiment Procedure

After obtaining informed consent, we explained the study details to the participants. The participants completed the experiment using an iPhone application, similar to the one used in Study 1 for collecting pairwise similarity ratings, on one of four types of iPhones. To block any sound from experimental noises, participants wore noise-canceling headphones playing white noise. Participants could take breaks at any time during the study, in addition to a mandatory break (three minutes) after every 200 comparison tasks. We maintained the room temperature between 20–23 degrees Celsius.

D. Results

The derived perceptual spaces were primarily divided into parameter-based and metaphor-based Tactons, with some exceptions for metaphor-based Tactons that featured relatively simple waveforms or a continuous envelope with a single pulse (Figure 8 (a)). Within the cluster of parameter-based Tactons, two envelope frequencies (0 Hz and 8 Hz) and three durations (100 ms, 500 ms, and 2000 ms) served as the primary parameters. Notably, the effects of envelope frequency became more pronounced as duration increased. These two parameters—envelope frequency and duration—further divided the cluster of parameter-based Tactons into six sub-clusters. In contrast, amplitude and carrier frequency were the least dominant parameters, exhibiting no discernible patterns within the sub-clusters.

VII. Study 3

Study 3 evaluates the potential of the proposed method for aggregating pairwise similarity evaluations based on the updated simulations for a large-scale set ($n = 48$). The study explores two n_s values (5 and 10) to test lower r_s values (10% and 21%) compared to those used in Study 1 (36% and 42%).

A. Subset design

We used the same 48 Tactons from Study 2 for both Study 3-1 and Study 3-2. In Study 3-1, we used $n_s = 10$, resulting in 45 comparison tasks per participant. Each participant experienced 21% of the total 48 Tactons in the set. While achieving $NO = 12$ would require 362 participants, we conducted the study with 250 participants ($NO = 6$–7), as predicted by the updated simulations (mean value from 100 simulations) to achieve $\rho \geq 0.7$ with the gold standards. In Study 3-2, we used $n_s = 5$, resulting in 10 comparison tasks per participant. Each participant experienced 10% of the total 48 Tactons in the set. Achieving $NO = 12$ would require 1,547 participants. However, we conducted the study with 607 participants ($NO = 5$–6), based on predictions from the updated simulations to achieve $\rho \geq 0.7$ with the gold standards.

B. Participants

For Study 3-1, we recruited 250 mTurk workers, all of whom had completed 10,000 HITs on the mTurk platform with a success rate greater than 98%. The participants were 20–78 years old (mean age: 36.6 years). The sample consisted of 183 Americans, 41 Indians, 13 Brazilians, 8 Canadians, 2 Colombians, 1 Italian, 1 Sri Lankan, and 1 Venezuelan. Participants used 24 different smartphone models, ranging from the iPhone 8 (oldest) to the iPhone 15 Pro (newest). They completed the rating tasks on the iPhone application in 5–54 minutes (mean: 10.5 minutes) and received $5 USD as compensation.

For Study 3-2, we recruited 607 mTurk workers under the same requirements as Study 3-1. The participants were 20–76 years old (mean age: 35.7 years). The sample included 397 Americans, 93 Indians, 41 Brazilians, 28 Canadians, 23 American Samoans, 7 Colombians, 6 Sri Lankans, 6 Venezuelans, 2 Italians, 1 Armenian, 1 Australian, 1 Austrian, and 1 Turkish. Participants used 25 different smartphone models, ranging from the iPhone 8 Plus (oldest) to the iPhone 15

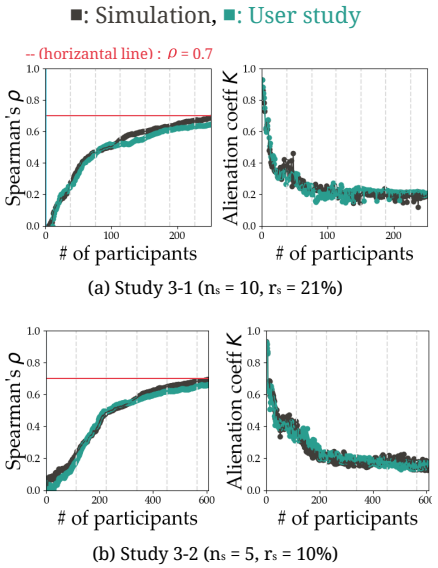


Fig. 9: Simulations of the subset aggregation methods (Studies 3-1 and 3-2).

Pro (newest). They completed the rating tasks on the iPhone application in 3–76 minutes (mean: 4.2 minutes) and received $3 USD as compensation.

C. Experiment Procedure

We used the same iPhone application as in Study 1, which consisted of consent, training, main, and feedback sessions. One key difference was that Study 3-1 and Study 3-2 were conducted independently due to differences in compensation, as participants completed different comparison tasks. In the training session, Study 3-1 participants experienced 10 Tactons, while Study 3-2 participants experienced 5 Tactons. Similarly, in the main session, Study 3-1 participants rated the similarity of 45 pairs, whereas Study 3-2 participants rated 10 pairs.

D. Results

The updated simulation (Section V) closely matched the aggregation of pairwise ratings for subsets in terms of both ρ and K (Figure 9). At the point where the simulation predicted $\rho = 0.7$, Study 3-1 achieved $\rho = 0.65$ with around $NO = 7$, and Study 3-2 achieved $\rho = 0.67$ with around $NO = 6$. Although there were slight differences, these ρ values (0.65 and 0.67) are sufficient to demonstrate strong correspondence between the gold standard (Study 2) and the subset aggregation methods across n_s and r_s values. Furthermore, the updated simulations provided predictions that closely aligned with reality for the Alienation Coefficient (K), achieving $K \leq 0.2$ with $NO = 2$, as predicted by the simulations.

The perceptual spaces derived from Study 3-1 and Study 3-2 also captured the main structure of the perceptual space of the gold standard (Figure 8). Similar to the gold standard, their perceptual spaces were primarily divided into parameter-based and metaphor-based Tactons, with a few exceptions where metaphor-based Tactons featured simple or continuous waveforms (e.g., m1, m6, m16, and m23). Within the cluster of parameter-based Tactons, the primary parameters—envelope frequency and duration—identified in Study 2 were also observed in Study 3-1 and Study 3-2. Secondary effects of amplitude and carrier frequency were similarly noted, aligning with the findings from Study 2. Although some fine details in the perceptual spaces of metaphor-based Tactons differed slightly from the gold standard, these differences do not undermine the effectiveness of the subset aggregation method in capturing the main structure of the gold standard's perceptual space.

VIII. Discussion

In this paper, we proposed a method for aggregating pairwise similarity evaluations for subsets of Tactons to construct robust perceptual spaces for large-scale Tacton sets, comparable to those of the gold standard. To improve the efficiency of collecting ratings, we utilized an online crowdsourcing platform and developed a simulation model to guide user studies in advance. The updated simulation model refined through Study 1 guided the construction of perceptual spaces using subset aggregation methods by providing the number of participants required to achieve $\rho \geq 0.6$ and $K \leq 0.2$, representing state-of-the-art correspondence with the gold standard. Based on the findings, we discuss the user study results and highlight implications for future research.

A. Reflection on User Studies

For both r_s values of 21% and 10%, our simulation closely matched the real user similarity ratings, achieving higher correspondence with the gold standard ($\rho = 0.65$–0.7 and $K \leq 0.2$) compared to previous work ($\rho = 0.15$, $K = 0.45$) [25]. Given that $w = 2.6$ showed consistent performance in both Studies 3-1 and 3-2 across various n_s and r_s values, the results suggest that the added model parameter w in the updated simulation represents the complexity of the vibration parameters. In simpler terms, when the effects of vibration parameters are perceptually clear, such as with sinusoidal parameters, the simulation requires a lower multiplying parameter, around $w = 1.6$, and fewer participants to achieve $\rho \geq 0.6$. In contrast, for vibrations designed using abstract parameters or complex waveforms, where interpreting the effects of vibration parameters demands higher cognitive effort, the simulation requires a higher w, around 2.6, and more participants to achieve the same correspondence with the gold standard. In other words, when Tacton sets were designed using sinusoidal parameters (i.e., when $w = 1.6$), an NO of around 2 is sufficient to achieve $\rho \geq 0.6$ (Study 1 (a)). However, for Tacton sets designed using abstract parameters or complex waveforms (Study 1 (b) and (c), Studies 3-1 and 3-2), $NO = 4$ was needed to achieve $\rho \geq 0.6$, and at least $NO = 7$ was required to achieve $\rho = 0.65$–0.7 across n and n_s. Overall, our proposed subset aggregation method, which uses pairwise ratings in crowdsourced settings, successfully generated perceptual spaces for entire Tacton sets with strong

correspondence to the ground truth across various n, n_s, r_s, and design approaches, guided by NO values from our simulation model. We attribute the improved performance to the combination of pairwise ratings from a diverse participant pool recruited via crowdsourcing and the effective aggregation of well-distributed subsets using a non-overlapping sequential sampling method. Together, these factors led to substantially higher correspondence with the gold standard compared to prior methods [25].

B. Limitations

While our proposed method for aggregating pairwise similarity ratings for subsets of Tactons in crowdsourced environments based on simulations achieved state-of-the-art correspondence with the gold standard, it has several limitations. First, due to the time complexity of pairwise ratings ($O(n^2)$), we limited the construction of the ground truth perceptual space to 48 Tactons. Completing pairwise ratings for 48 Tactons required six hours spread over two days. Increasing the number of Tactons would significantly amplify participant fatigue and task completion time, making the process impractical. Thus, we consider 48 Tactons an appropriate upper limit for generating high-quality gold standard perceptual spaces using pairwise ratings. Second, while our updated simulations indicated that the proposed method can work for larger Tacton sets (e.g., 120 or 240 Tactons), its performance at such scales remains untested with real users. Future research should focus on validating the method for these larger sets to determine its practical limits and ensure scalability. Lastly, our simulation framework relies on the availability of a gold-standard dissimilarity matrix. Nevertheless, it still provides guidance on how participant burden can be reduced for unseen Tacton sets across diverse design approaches while maintaining strong correspondence with ground-truth perceptual spaces. Future work could explore gold-standard-free simulations that optimize the required number of participants using w, which we currently conjecture to represent a complexity indicator for Tacton design.

C. Implications for Future Work

We outline how our results can support the investigation of perceptual dissimilarity spaces of Tactons and inform future research directions.

Designers can prototype the perceptual effects of large-scale Tactons using our approach. Our findings empirically established that $NO = 4$ was sufficient to achieve $\rho = 0.6$ with the dissimilarity matrix of the gold standard, while $NO = 7$ was required to reach $\rho = 0.65$–0.7. These results held across various Tacton design approaches and values of n (12–48), n_s (5 and 10), and r_s (10%–42%). For the Alienation Coefficient (K), $NO = 2$ was sufficient to construct a statistically similar perceptual space to that of the gold standard. These results remained stable across various distributions in perceptual spaces, as demonstrated by the proposed simulations, as well as across different n, n_s, and r_s values. This consistency suggests the potential for early stopping when exploring the perceptual effects of Tactons, requiring fewer participants than the conventional sample sizes ($NO = 12$) typically needed for generalization, whether using subset aggregation methods ($n_s < n$) or evaluating the entire set of Tactons ($n_s = n$).

Our results can inform the development of future computational models to predict similarity perception for vibrotactile Tactons. Developing prediction models for Tacton similarity perception poses significant challenges, particularly in collecting high-quality rating data for various Tactons created using diverse design approaches. Pairwise ratings are especially demanding, requiring $O(n^2)$ comparison tasks, and similarity perception can be influenced by the specific Tactons included in the set. To address these challenges, a recent study proposed a computational model inspired by the neural processes involved in transmitting vibrations from the skin to the brain [15]. Enhancing such models would benefit from access to extensive, high-quality similarity datasets, particularly for training deep learning-based models that rely on numerous parameters to capture complex relationships. Our study contributes to this effort by providing similarity data collected from 782 users for 5 Tactons, 250 users for 10 Tactons, and 12 users for 48 Tactons, designed using parameter-based and metaphor-based approaches, as well as their combinations. This dataset serves as a new resource for training advanced computational models, offering the potential to improve predictions of similarity perception for vibrotactile Tactons.

IX. Conclusion

Aggregating subjective similarity ratings for subsets of Tactons offers a valuable opportunity to explore the perceptual space of large-scale Tacton sets. In this study, we conducted multiple user studies to collect pairwise ratings in crowdsourced settings, guided by computational simulations refined with user study data. The results from our studies and simulations highlight the number of participants needed to prototype the similarity perception of the gold standard with high correspondence ($\rho = 0.65$–0.7 and $K \leq 0.2$). We hope that our findings assist designers in accelerating the design process for creating distinguishable Tactons for a variety of applications and inspire the development of future computational models for predicting similarity perception of Tactons.

Acknowledgment

This work was supported by research grants from VILLUM FONDEN (VIL50296, 25%), the National Science Foundation (#2339707, 25%), the Institute of Information & communications Technology Planning & Evaluation (IITP) funded by the Korea government (MSIT) (Artificial Intelligence Graduate School Program (GIST), No.2019-0-01842, 25%), and the Korea Medical Device Development Fund grant funded by the Korean government (RS-2021-KD000009, 25%).

REFERENCES

[1] Y. Gaffary and A. Lécuyer, "The use of haptic and tactile information in the car to improve driving safety: A review of current technologies," *Frontiers in ICT*, vol. 5, p. 5, 2018.

[2] J. Ryu, J. Chun, G. Park, S. Choi, and S. H. Han, "Vibrotactile feedback for information delivery in the vehicle," *IEEE Transactions on Haptics*, vol. 3, no. 2, pp. 138–149, 2010.

[3] R. K. Katzschmann, B. Araki, and D. Rus, "Safe local navigation for visually impaired users with a time-of-flight and haptic feedback device," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 26, no. 3, pp. 583–593, 2018.

[4] R. F. Friesen and Y. Vardar, "Perceived realism of virtual textures rendered by a vibrotactile wearable ring display," *IEEE Transactions on Haptics*, 2023.

[5] K. Jung, S. Kim, S. Oh, and S. H. Yoon, "Hapmotion: motion-to-tactile framework with wearable haptic devices for immersive vr performance experience," *Virtual Reality*, vol. 28, no. 1, p. 13, 2024.

[6] M. Schirmer, J. Hartmann, S. Bertel, and F. Echtler, "Shoe me the way: A shoe-based tactile interface for eyes-free urban navigation," in *Proceedings of the 17th International Conference on Human-Computer Interaction with Mobile Devices and Services*, 2015, pp. 327–336.

[7] M. Lee, G. Bruder, and G. F. Welch, "Exploring the effect of vibrotactile feedback through the floor on social presence in an immersive virtual environment," in *2017 IEEE Virtual Reality (VR)*. IEEE, 2017, pp. 105–111.

[8] J. Seo, S. Mun, J. Lee, and S. Choi, "Substituting motion effects with vibrotactile effects for 4d experiences," in *Proceedings of the 2018 CHI conference on human factors in computing systems*, 2018, pp. 1–6.

[9] E. Pezent, A. Israr, M. Samad, S. Robinson, P. Agarwal, H. Benko, and N. Colonnese, "Tasbi: Multisensory squeeze and vibrotactile wrist haptics for augmented and virtual reality," in *2019 IEEE World Haptics Conference (WHC)*. IEEE, 2019, pp. 1–6.

[10] J. Seo and S. Choi, "Perceptual analysis of vibrotactile flows on a mobile device," *IEEE transactions on haptics*, vol. 6, no. 4, pp. 522–527, 2013.

[11] T. Singhal and O. Schneider, "Juicy haptic design: Vibrotactile embellishments can improve player experience in games," in *Proceedings of the 2021 chi conference on human factors in computing systems*, 2021, pp. 1–11.

[12] A. Israr, H. Z. Tan, and C. M. Reed, "Frequency and amplitude discrimination along the kinesthetic-cutaneous continuum in the presence of masking stimuli," *The Journal of the Acoustical society of America*, vol. 120, no. 5, pp. 2789–2800, 2006.

[13] I. Hwang and S. Choi, "Perceptual space and adjective rating of sinusoidal vibrations perceived via mobile device," in *2010 IEEE Haptics Symposium*. IEEE, 2010, pp. 1–8.

[14] G. Park and S. Choi, "Perceptual space of amplitude-modulated vibrotactile stimuli," in *2011 IEEE world haptics conference*. IEEE, 2011, pp. 59–64.

[15] C. Lim and G. Park, "Can a computer tell differences between vibrations?: Physiology-based computational model for perceptual dissimilarity prediction," in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 2023, pp. 1–15.

[16] D. Kwon, R. Abou Chahine, C. Lim, H. Seifi, and G. Park, "Can we crowdsource tacton similarity perception and metaphor ratings?" in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 2023, pp. 1–13.

[17] D. Ternes and K. E. MacLean, "Designing large sets of haptic icons with rhythm," in *Haptics: Perception, Devices and Scenarios: 6th International Conference, EuroHaptics 2008 Madrid, Spain, June 10-13, 2008 Proceedings 6*. Springer, 2008, pp. 199–208.

[18] R. Abou Chahine, D. Kwon, C. Lim, G. Park, and H. Seifi, "Vibrotactile similarity perception in crowdsourced and lab studies," in *International Conference on Human Haptic Sensing and Touch Enabled Computer Applications*. Springer, 2022, pp. 255–263.

[19] Ultraleap, "Haptic muse," 2013, [Online; accessed 27-August-2023]. [Online]. Available: https://ir.immersion.com/news-releases/news-release-details/immersion-releases-haptic-muse-effect-preview-app-android-game

[20] J. B. van Erp, M. M. Spapé *et al.*, "Distilling the underlying dimensions of tactile melodies," in *Proceedings of Eurohaptics*, vol. 2003, 2003, pp. 111–120.

[21] A. Israr, S. Zhao, K. Schwalje, R. Klatzky, and J. Lehman, "Feel effects: enriching storytelling with haptic feedback," *ACM Transactions on Applied Perception (TAP)*, vol. 11, no. 3, pp. 1–17, 2014.

[22] H. Seifi, K. Zhang, and K. E. MacLean, "Vibviz: Organizing, visualizing and navigating vibration libraries," in *2015 IEEE World Haptics Conference (WHC)*. IEEE, 2015, pp. 254–259.

[23] J. Pasquero, J. Luk, S. Little, and K. MacLean, "Perceptual analysis of haptic icons: an investigation into the validity of cluster sorted mds," in *2006 14th Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems*. IEEE, 2006, pp. 437–444.

[24] M. L. Davison and S. G. Sireci, "Multidimensional scaling," in *Handbook of applied multivariate statistics and mathematical modeling*. Elsevier, 2000, pp. 323–352.

[25] D. R. Ternes, "Building large sets of haptic icons: rhythm as a design parameter, and between-subjects mds for evaluation," Ph.D. dissertation, University of British Columbia, 2007.

[26] M. Demers, P. E. Fortin, A. Weill, Y. Yoo, J. R. Cooperstock *et al.*, "Active sampling for efficient subjective evaluation of tactons at scale," in *2021 IEEE World Haptics Conference (WHC)*. IEEE, 2021, pp. 1–6.

[27] J. Ryu, J. Jung, G. Park, and S. Choi, "Psychophysical model for vibrotactile rendering in mobile devices," *Presence*, vol. 19, no. 4, pp. 364–387, 2010.

[28] H. Z. Tan, N. I. Durlach, C. M. Reed, and W. M. Rabinowitz, "Information transmission with a multifinger tactual display," *Perception & psychophysics*, vol. 61, no. 6, pp. 993–1008, 1999.

[29] Y. Yoo, T. Yoo, J. Kong, and S. Choi, "Emotional responses of tactile icons: Effects of amplitude, frequency, duration, and envelope," in *2015 ieee world haptics conference (whc)*. IEEE, 2015, pp. 235–240.

[30] I. Hwang, J. Seo, and S. Choi, "Perceptual space of superimposed dual-frequency vibrations in the hands," *PloS one*, vol. 12, no. 1, p. e0169570, 2017.

[31] Y. Yoo, I. Hwang, and S. Choi, "Perceived intensity model of dual-frequency superimposed vibration: Pythagorean sum," *IEEE Transactions on Haptics*, vol. 15, no. 2, pp. 405–415, 2022.

[32] L. M. Brown, S. A. Brewster, and H. C. Purchase, "Multidimensional tactons for non-visual information presentation in mobile devices," in *Proceedings of the 8th conference on Human-computer interaction with mobile devices and services*, 2006, pp. 231–238.

[33] J. Tan, Y. Ge, X. Sun, Y. Zhang, and Y. Liu, "User experience of tactile feedback on a smartphone: Effects of vibration intensity, times and interval," in *Cross-Cultural Design. Methods, Tools and User Experience: 11th International Conference, CCD 2019, Held as Part of the 21st HCI International Conference, HCII 2019, Orlando, FL, USA, July 26–31, 2019, Proceedings, Part I 21*. Springer, 2019, pp. 397–406.

[34] L. M. Brown, *Tactons: structured vibrotactile messages for non-visual information display*. University of Glasgow (United Kingdom), 2007.

[35] L. M. Brown, S. A. Brewster, and H. C. Purchase, "A first investigation into the effectiveness of tactons," in *First joint eurohaptics conference and symposium on haptic interfaces for virtual environment and teleoperator systems. world haptics conference*. IEEE, 2005, pp. 167–176.

[36] O. S. Schneider and K. E. MacLean, "Studying design process and example use with macaron, a web-based vibrotactile effect editor," in *2016 IEEE Haptics Symposium (Haptics)*. IEEE, 2016, pp. 52–58.

[37] Y. Vardar, C. Wallraven, and K. J. Kuchenbecker, "Fingertip interaction metrics correlate with visual and haptic perception of real surfaces," in *2019 IEEE World Haptics Conference (WHC)*. IEEE, 2019, pp. 395–400.

[38] C. Lim, G. Park, and H. Seifi, "Designing distinguishable mid-air ultrasound tactons with temporal parameters," in *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 2024, pp. 1–18.

[39] I. Borg and D. Leutner, "Measuring the similarity of mds configurations," *Multivariate Behavioral Research*, vol. 20, no. 3, pp. 325–334, 1985.

[40] Y. Chan, "Biostatistics 104: correlational analysis," *Singapore Med J*, vol. 44, no. 12, pp. 614–619, 2003.

[41] C. P. Dancey and J. Reidy, *Statistics without maths for psychology*. Pearson education, 2007.