A Data-Driven Strategy for the Subjective Evaluation of Tacton Perceptual Similarity

Marc Demers

Department of Electrical & Computer Engineering

McGill University

Montréal, Québec, Canada

May 2021

A thesis submitted to McGill University in partial fulfillment of the requirements for the degree of Master of Science.

2021/05/20

Abstract

Traditional tacton evaluation studies often rely on aggregating the results of a small number of participants to a predefined number of stimuli. This approach does not scale well to a large number of tactons. In this work, we propose to flip this traditional framework upside-down: instead of designing the tactons *a priori*, we randomly generate tactons on-the-fly and send them for evaluation to the user. In addition, we develop a scalable haptic smartphone data collection method that can deal with concurrent haptic ratings to gather robust and relevant data, and deploy it remotely on the Amazon Mechanical Turk platform. We measure vibrotactile perceptual similarity between tactons via a probabilistic model, and further develop an active sampling strategy grounded in probability and information theory to efficiently sample the space of possible tactons, thereby reducing the amount of data required by 6.5 times. We conduct an experiment with over 200 participants, from which we extract key information about tactile perceptual similarity such as communities of perceptually similar tactons. Furthermore, we find evidence of "personas," or groups of people that share perceptual similitude, and report on the characteristics and possible origins of these personas. We also show an approach to perform machine learning on a graph representation of the similarity ratings, allowing us to successfully predict the out-of-sample similarity scores. The personas are shown to help with this out-of-sample prediction, proving to a greater extent their relevance and utility. All in all, experimental results indicate that this high-data regimen is a promising new take at conducting user studies in haptics.

Résumé Scientifique

Traditionnellement, les études d'évaluation de tactons se fient sur l'aggrégation de résultats d'un faible nombre de participants à une quantité prédéfinie de stimuli. Cette approche ne s'étend pas aisément à un grand nombre de tactons. Nous proposons ici de générer les tactons en temps réel et de les envoyer à l'utilisateur pour évaluation directement, au lieu de les concevoir *a priori*. De plus, nous développons une approche de collection de données haptiques via smartphone qui puisse récupérer des annotations robustes et pertinentes de manière concurrente, et la déployons sur Amazon Mechanical Turk. Nous mesurons la similarité perceptuelle entre les tactons via un modèle probabiliste, et développons une stratégie d'apprentissage active basée sur la théorie de l'information pour échantilloner l'espace de tactons de manière efficace, réduisant de ce fait la quantité de donnée requise par un facteur de 6.5. Nous conduisons une experience avec plus de 200 participants, de laquelle nous tirons de l'information clé sur la perception tactile, telle que des communautés de tactons similaires. En outre, nous montrons une preuve de l'existence de "personas," ou des groupes de personnes qui partagent des caractéristiques similaires sur la perception haptique, et discutons de l'origine possible de ces personas. Finalement, nous montrons une approche pour effectuer de l'apprentissage machine sur une représentation graphique d'annotations de similarité, ce qui nous permet de prédire les scores de similarité des tactons. L'intégration des personas améliore cette prédiction, ce qui prove leur pertinence et leur utilité. En somme, les résultats empiriques de nos expériences montrent que ce régime haut en données est une approche prometteuse pour conduire des études d'évaluation haptique.

Acknowledgements

First and foremost, I would like to thank the laboratory members at the Shared Reality Lab, who have given feedback on this work and helped structure the ideas. Particularly, I would like to thank Jeffrey Blum and Yongjae Yoo, who always actively participated in this project, and, more importantly, Pascal Fortin and Antoine Weill-Duflos, who always believed in the craziness of the ideas and persevered with me throughout the multiple iterations and backlashes. I want to thank my advisor, Jeremy Cooperstock, for believing in me, for his help in preparing this work, and for providing me with the academic freedom to pursue my own interests and ideas.

I express my gratitude to Ilja Frissen, the internal examiner of this thesis, for their comments and valuable feedback that enhanced the quality of the work.

To my friends and family, no words could describe how grateful I am that you are in my life. Particularly, to my parents: thank you for providing me with the tools to become who I wanted to be; to my brother Éric, thank you for being the supportive friend you have always been; to my partner Léa-Frédéricke, thank you for supporting me through the highs and the lows of graduate school life.

Finally, I would like to thank all those who, throughout my life, have given me support and fueled my desire to push my limits: *je vous en suis éternellement reconnaissant.*

This work was supported by a Natural Sciences and Engineering Research Council (NSERC Canada) Discovery Grant RGPIN-2017-05013.

Contents

1 Introduction **1**
1.1 Perceptual Similarity . . . 2
1.1.1 Modeling Users . . . 3
1.2 Graph Approach . . . 4
1.3 Author's Contribution . . . 4

2 Background **6**
2.1 Tacton Perception . . . 6
2.1.1 Interpreting Tactons . . . 6
2.1.2 Tacton Parameters . . . 7
2.1.3 Tacton Libraries . . . 10
2.1.4 Personalization & Customization . . . 11
2.2 Evaluating Perceptual Similarity . . . 12
2.2.1 Measuring Similarity . . . 12
2.2.2 Similarity in Music . . . 14
2.2.3 Similarity in VT Haptics . . . 14

3 A Bottom-Up Approach to Tacton Evaluation **16**
3.1 Exploring the Tacton Space . . . 18
3.1.1 Introducing the Bottom-Up Approach . . . 18
3.1.2 Tacton Rendering . . . 19
3.1.3 Tacton Characteristics . . . 20
3.2 Discrete Choice Analysis . . . 21

3.3 Outsourcing to the Crowd . . . 22
3.4 Data Flow . . . 24
3.5 Probabilistic Modeling for Similarity Evaluation . . . 26
3.5.1 Definitions . . . 27
3.5.2 Probabilistic Model . . . 28
3.5.3 Maximal Information Gain . . . 30
3.5.4 Graph Representation . . . 33
3.5.5 Machine Learning on the Graph . . . 34

4 Experiments **36**
4.1 Methodology . . . 38
4.1.1 Participants and Remuneration . . . 38
4.1.2 Concurrency . . . 38
4.1.3 Pilot Studies . . . 38
4.1.4 Description of the Experiments . . . 41
4.2 Global Similarity Assessment . . . 42
4.2.1 Attention Tests . . . 43
4.2.2 Active Sampling . . . 43
4.2.3 Perceptual Similarity Aggregation . . . 44
4.2.4 Discussion . . . 50
4.3 Personas . . . 54
4.3.1 Feature Saliency across Personas . . . 57
4.3.2 Demographic Information and Personas . . . 57
4.3.3 Discussion . . . 59
4.4 Predicting Similarity Ratings . . . 60
4.4.1 Graph Representation Learning . . . 60
4.4.2 Predicting Similarity on the Global Experiment Data . . . 61
4.4.3 Extending Similarity Prediction to Personas . . . 61
4.4.4 Discussion . . . 62

5 Conclusion **65**
5.1 Summary . . . 65

5.1.1 Shortcomings 67
5.1.2 Potential Impact and Future Work 68

References **70**

Appendices **81**

Appendix A Instruction Form **82**

Appendix B Post-experiment Questionnaire **83**

Appendix C Persona Groups **85**

Appendix D Description of the Features **87**

Appendix E Learning Experiments Details **88**
E.1 Persona Clustering 88
E.1.1 Hyperparameters 88
E.2 Graph Representation Learning 89
E.2.1 Hyperparameters 89
E.2.2 Train/Test Loss Curves 89
E.3 Regressor Hyperparameters 90

List of Figures

3.1 The proposed bottom-up approach to designing tactons, as opposed to the traditional top-down approach. 19
3.2 One round of grouping the tactons into similarity clusters. The user iteratively evaluates the similarity between the tactons and places them into clusters. Once they are satisfied with it, they submit the grouping. This process repeats for R rounds. 25
3.3 Overview of the software architecture behind the haptic rating system. . . . 26
3.4 Graphical explanation of the weighing scheme for the PCM matrix A. Note the annotator performance weight η^r is not depicted here. 29
3.5 EIG as a function of the mean rating between two tactons and its associated variance. As expected, pairs that have as many similarity ratings as dissimilarity ratings exhibit the highest EIG. 32
3.6 Tacton similarity as edge weights in a graph. 34

4.1 Performance of participants on the "gold standard" attention test. 41
4.2 Performance of active sampling vs. a baseline that randomly selects the same number of pairs at each round. A lower EIG signifies greater certainty in the groupings. 43
4.3 Well-connectedness in the communities found by the Leiden algorithm (left). The right cluster becomes disconnected because node 0 was sent to another community, thereby dismantling the previously formed red community. Inspired by Traag *et al.* [1]. 44

4.4 Community detection using the Leiden algorithm in the similarity network. Edges representing similarity are depicted in green, and those representing dissimilarity edges are depicted in red. 45
4.5 Pairwise similarity ratings across tactons and participants. 46
4.6 Feature distribution across all clusters. The distributions were normalized so as to compare them on a similar scale. The asterix represents statistical significance across all groups for that feature according to a Kruskal-Wallis H-test with 95% confidence. 47
4.7 Characterizing individual differences by decomposing the ratings for each feature, and looking at the uncertainty (measured by the EIG) of the probabilistic model with respect to the difference in feature values in the tacton pairs. 49
4.8 Flattening the persona ratings. 55
4.9 HDBSCAN clustering linkage tree for grouping the personas. A smaller distance indicates a finer clustering. We settled on a splitting distance of 4.4 because it exhibits the highest agreement between the personas and the tacton features. 56
4.10 Feature saliency for each persona group and tacton feature. 57
4.11 Demographic information distribution for each persona group. 58
4.12 The triplet loss maximizes the distance between the anchor embedding $\psi(A)$ and the negative embedding $\psi(N)$ whilst minimizing the distance between the anchor embedding and the positive embedding $\psi(P)$. 61
4.13 Detailed R^2 results for all persona groups for the gradient boosting regressor. 63

A.1 Instruction form presented to participants before completing the experiment. 82

B.1 Post-experiment questionnaire. 84

E.1 Triplet loss training and testing loss curves, plotted for 1 seed and a single fold for ease of view. 89

List of Tables

2.1 Common tacton parameters found in the literature. 9
2.2 Classification of similarity psychometric models. Adapted from [2]. 14

4.1 Iterative process behind the final experiment. We ensured that no participant did the experiment more than once across all iterations. 40
4.2 Average error on the testing set for the similarity prediction task. All results are averaged over five different seeds (for the gradient boosting regressor), and cross-validated in a five-fold scheme. 62

C.1 Persona groups composition. 86

E.1 Hyperparameters for Persona Clustering using UMAP and HDBSCAN. . . 88
E.2 Hyperparameters for graph representation learning. 89
E.3 Hyperparameters for gradient boosting. 90

Chapter 1

Introduction

The sense of touch, or haptics, has long been a source of intringue. The first occurrence of synthetically generated haptics being used for communication goes back to the 1960s [3]. More recently however, there has been a need to develop novel means of communication for situations in which other prevalent senses are overwhelmed with information [4], or to contribute to the User Experience (UX) [5].

Tactile icons, largely referred to in the literature as tactons or *haptic icons*, depict structured abstract tactile messages that encode information that can be carried non-visually [6], usually via tactile encoding or via force. These tactons are typically individually designed, but can also be hierarchically designed, meaning that one tacton can build upon previous tactons so as to combine their perceptual properties. By far, the most common way to convey tactons is through vibrations. This work focuses primarily on vibrotactile (VT) tactons, or tactons delivered to the user via vibrations rendered on the surface of the skin, through an actuator more or less tightly coupled to the body.

While investigative work has looked into the effectiveness of tactons, there remain three main challenges that push back on the adoption of tactons in life outside the lab. First, very little work has been done to aggregate the results of tacton evaluation of several user studies into one coherent, organized database [7, 8]. This decreases the ability of designers to readily make use of the results of previous experiments, which in turn can force designers to reinvent the wheel on every iteration [7, 9]. However, most if not all the data that constitute these databases are aggregated from participants in user studies

that were done in-the-lab, in a known and isolated setting, whereas in real life, tactons are typically perceived in various uncontrolled environments. The data that constitute the databases can thus be far from the reality of the perception of tactons.

Second, experiments involving tacton evaluation and characterization also tend to exhibit large variability across participants, known as individual differences (ID) [10, 11]. These individual differences affect the takeaways of these types of studies: the results are typically only applicable to a particular setting, and they are hardly ever reproducible as the user base that one might encounter in real life could be completely different from that of the experiment. One crucial component of the aforementioned databases could thus include the characterization and the quantification of the variability in perception across several individuals or groups thereof due to IDs.

Third, experiments typically explore a limited subset of the tacton space, where the experimenter designs the tactons and evaluates them in an iterative fashion across experiments in one or multiple user studies, thereby ignoring potentially meaningful but seemingly unrelated tactons. This introduces a bias in the methodology in that the prior knowledge of the experimenter can affect the outcome of the findings.

Last, negative results or results that do not depict an agreement across the population are rarely reported and/or emphasized, although in the context of IDs they may not necessarily be invalid. One hypothesis is that they may be characteristics of a given segment of the participant population.

1.1 Perceptual Similarity

Obtaining a consensus on perception is often an arduous task, as it closely relates to the senses, which in turn are very dependent on the individual who is sensing and the surrounding context. Several studies have attempted to assess the thresholds for which human beings perceive music to be similar, especially in the new context of music recommendation on online platforms, so much so that a whole book has been dedicated to the topic [12]. However, despite its use in everyday life, VT haptics have yet to be investigated to the same extent. This can be partly attributed to the fact that haptics are less expressive as a channel of communication, but also because of the emergent nature of

the subfield of haptic perception. Additionally, the abundance of contexts and the lack of tailoring methods to those contexts make it very hard to generate *meaningful* haptics. For instance, an alarming tacton may be perceived as an disturbance when presented during a meeting, whereas it may be welcomed when notifying the user of an emergency. Although the same stimulus was sent to the user, the context surrounding the percept was different: its interpretation, therefore, is strongly linked to the context.

There has been a recent interest in the haptics community to *automatically* generate haptic stimuli. Contrarily to music where the content generation process typically takes place on the artist's side, this haptic generation is expected to stem from both the designer (the haptician [13]) and the end-user (the recipient) of the stimuli (through customization or personalization). Before taking this step, we believe that the nature of haptics must be analyzed from a perceptual standpoint. We make the case that without prior knowledge of the perceptual similarity of haptics among several groups of users, customizing or personalizing tactons to suit a user in particular or a task in particular will prove difficult, perhaps impossible. Hence, this work treats perceptual similarity as a stepping-stone toward automatic generation of haptic content with the intent of personalizing the stimuli.

To accomplish this goal, our strategy removes inductive biases in the methodology mainly by flipping the traditional approach upside-down: instead of evaluating a known set of tactons, we randomly generate tactons and attempt to converge to an agreement (in our case on similarity) through an iterative and data-intensive process.

1.1.1 Modeling Users

Our perception of touch is altered by our personality, our past experiences, our ethnic background, and culture. As such, it is difficult to obtain agreement on perception of haptic stimuli through independent user studies. Seifi et al. [14] has tackled end-user stimulus customization in an attempt to make up for the IDs. Because end-user customization aims to fit the user's individual needs effectively, it is sometimes referred to as *stimulus personalization*.

However, the literature still does not agree on a method for clustering users into groups of people who perceive vibrations similarly. Such a grouping would allow to reach an agreement across a number of people [15], such that stimuli could be tailored

to that group of people only and thus be more effective in conveying meaning or intent. This would let us *quantify* the differences and variations by which external factors influence our perception. In addition, finding groups of people who perceive tactons similarly can have a major impact in domains such as affective haptics — which is increasingly personal and subject to a different interpretation by each individual — or machine-generated haptic stimuli. In the case of machine-generated stimuli, any information about the user can provide information about the perception of the intent of the stimulus; this intuition is something that expert designers have but that machines lack.

1.2 Graph Approach

VT tactons are frequently classified by some characteristics derived from signal processing or from the music world: energy, frequency spectrum, rhythm or tempo. Attempting to extract meaning solely from these characteristics does not tell the whole story of the *impact* of those characteristics on our perception, because they are noisy, punctual measures in the tacton space. It remains unclear how modifying a single tacton's characteristics impacts our perception of it. As such, we wish to represent not only the tactons themselves but their relationship to one another, in an attempt to characterize the *linkage* between the different characteristics. Hence, there is a need for an approach that encompasses both individual tacton features as well as the topology of tactons.

To accomplish this goal, and central to our methodology, we develop in Chapter 3 an approach that leverages tactons as objects in a graph: the nodes represent the tactons, the edges the relationships between them. This approach is both scalable to the hundreds or thousands of tactons, and can naturally improve our understanding of the perception of the tacton space despite its complexity and abstractness.

1.3 Author's Contribution

In this work, we address the challenges reported in these previous section.

Our contributions are the following:

1. We develop an Android application for smartphones that can render VT stimuli and gather feedback about them in a crowdsourced setting, via Amazon Mechanical Turk.[1] (AMT)

2. We outline the lessons learned from crowdsourcing haptic tacton evaluation.

3. We apply our graph theoretic methodology that combines both tacton characteristics and topological information to analyzing the data and show its relevance in haptic perceptual studies.

4. We show strong evidence of the presence of users who share haptic perceptual characteristics.

5. We use our graph approach to predict similarity between never-seen-before tacton pairs in the graph, and show that the predictive power increases if we predict similarity inside a particular group of users, thereby validating our groupings.

This is, to our knowledge, the first work to tackle both haptic perceptual similarity, as well as quantifying the perceptual differences of tactons among several groups across the population.

In Chapter 2, we review the existing literature about haptics, crowdsourcing, and perceptual similarity. Next, we present our approach in Chapter 3. We evaluate the approach and discuss the experimental results in Chapter 4. We conclude with the implications of the current work and our expectations for future work.

[1] https://www.mturk.com/

Chapter 2

Background

This thesis focuses on perceptual similarity in VT tactons. We divide this chapter into two parts: haptic tacton perception and perceptual similarity evaluation. In the first part, we overview the cognitive mechanisms that underlie how we perceive tactons, along with the signal parameters by which we make sense of and attribute meaning to them. Then, we survey how personalization, customization and error characterization help develop more robust models of tacton perception. In the second part, we discuss perceptual similarity in both music and haptics. Because of the proximity of the two domains, we include relevant work from the musical literature; challenges that have been overcome in one could translate to the other. Afterwards, we overview existing work in the domain of haptic tactons similarity, from what was done to what needs improvement.

2.1 Tacton Perception

Tactile icons have been explored and analyzed from a variety of different angles. In this section, we overview the development of tacton perception analysis through time.

2.1.1 Interpreting Tactons

The first occurrence of the term "tacton" to define short abstract tactile signals conveying meaning was carried out by Brewster *et al.* [6]. Prior to this, drawing influence from the

music literature, vibrations that were destined to be interpreted by the end user were known as "tactile melodies" [16].

Subsequent research has focused on evaluating tactons for their usefulness to and perception by users [17]. The investigation was further extended to mobile devices [18, 19] and mobile phone alerts [20].

Tacton analysis typically refers to the analysis of the *interpretation* of tactons from a perceptual standpoint. In the case of VT tactons, this interpretation implies a form of *communication* between the generating device (i.e. the actuator) and the recipient (i.e. the user). This communication channel is typically exploited in a number of contexts. In some instances, it is used to carry meaning [21, 22], or affect [23, 24, 25, 26]. It can also be used to convey spatial cues [27, 28, 29], and even transmit linguistic information [30, 31, 32, 33, 34].

These pieces of work highlight the importance of considering the cognitive mechanisms (the mappings) underlying the interpretation of haptic tactons as metaphoric entities (the meanings).

2.1.2 Tacton Parameters

At the physical level, tactons are time series signals, the two main physical characteristics of which are frequency and amplitude. Just like any signal, there are, however, a multitude of physical parameters that can be used to further describe them, some examples of which are: energy, rhythm, duration, roughness, spectral bandwidth, entropy, ... Due to the breadth of this parameter space, explaining and classifying tactons only through frequency and amplitude proves difficult. Designers have thus conducted experiments to identify the most important characteristics that humans use to distinguish and thus interpret tactons.

Table 2.1 summarizes the parameters studied in the tacton perception literature through time. The reader should notice the lack of trend in the table. This lack of trend justifies the need to identify what physical parameters influence our *perception* of tactons, in the hope that this will accelerate the development of meaningful tools to help design them. In the more recent years, Jones *et al.* pointed out that "a better understanding of which

dimensions of vibrotactile stimuli are perceptually important will require a larger set of stimuli and a wider range of actuators.” [35] In the present work, we tackle just that.

Table 2.1: Common tacton parameters found in the literature.

Author	Year	Spectral parameters (low-level concepts)					Spatio-temporal parameters (high-level concepts)			
		Amplitude/ Intensity	Frequency/ Pitch	Waveform/ Roughness	Duration	Tempo/ Rate	Rhythm	Texture	Direction	Spatial Location
Brewster *et al.* [6]	2004	x	x	x	x		x			x
Brown *et al.* [17]	2005	x	x	x	x		x			x
Brown *et al.* [20]	2006	x	x	x			x			x
Brown *et al.* [18]	2006	x		x						
Hoggan *et al.* [36]	2006	x		x			x	x		x
Luk *et al.* [19]	2006	x		x	x					
Hoggan *et al.* [37]	2007			x						
Lin *et al.* [27]	2008					x	x			
Ternes *et al.* [38]	2008	x	x				x			
Hoggan *et al.* [39]	2009				x	x	x	x		x
Brewster *et al.* [21]	2010			x			x			
Azadi *et al.* [40]	2013	x	x		x					
Qian *et al.* [41]	2013	x	x		x					
Tam *et al.* [10]	2013			x						
Osman *et al.* [24]	2014	x			x		x			
Pakkanen *et al.* [29]	2014						x			
Barber *et al.* [28]	2015								x	
Seifi *et al.* [8]	2015	x	x	x	x	x				
Ernst *et al.* [42]	2016	x		x	x					
Schneider *et al.* [43]	2016	x	x							
Stein *et al.* [44]	2017	x	x	x	x		x			
Egloff *et al.* [45]	2018		x				x			
Ferguson *et al.* [26]	2018		x	x	x	x				
Jones *et al.* [35]	2018	x		x						
Seifi *et al.* [46]	2018	x	x	x	x	x	x			

2.1.3 Tacton Libraries

Tacton libraries are structured groupings of stimuli that can give insight into existing haptic tacton perception knowledge. While not all libraries categorize tactons according to their physical parameters, the effects that are in the libraries can improve our understanding of the cognitive processes behind the meaning-mapping done when receiving the stimuli.

Guest *et al.* presented a comprehensive language that describes the whole experience of touch, thereby establishing a touch lexicon [47]. They conducted a user study to see how well the adjectives described sensory and emotional aspects of touch, and presented the sensory attributes in an organized fashion. In 2013, Immersion Corporation released Haptic Muse,[1] an application part of their Haptic SDK that "invites developers into a haptic museum with galleries built around common gaming use cases, like sports, transportation, combat and casinos."

In 2014, Israr *et al.* released a library of haptic feedback called FeelEffects [7]. They defined a "Feel Effect" as an explicit pairing between a meaningful linguistic phrase and a rendered haptic pattern. This is the first iteration of a library that illustrates a systematic approach to "designing a vocabulary of haptic sensations that are related in both the semantic and parametric spaces." [7] This set the way forward for VibViz [8], which categorized 120 VT effects into 5 distinct categories (physical, sensory, emotional, usage examples, metaphoric). The library has an interactive tool for end-user library navigation,[2] as well as support for open-ended questions such as "Find a vibration that feels like...". This is a step towards more robust understanding behind the internal cognitive schemas that we use to attribute meaning to the abstract stimuli. More recent work from the same author has focused on improving VibViz, and concluded that the challenges inherent to haptic evaluation can be "approached through the development of new, haptic-specific methodologies and evaluation instruments." [14]

All things considered, it seems that approaches involving scalable data collection to mapping the users' comprehension of large sets of haptic tactons are beneficial to further

[1] https://ir.immersion.com/news-releases/news-release-details/immersion-releases-haptic-muse-effect-preview-app-android-game

[2] https://www.cs.ubc.ca/~seifi/VibViz/main.html

improve our understanding of the VT space. We conclude by noting that there is, to our knowledge, no existing library on tacton perceptual similarity.

2.1.4 Personalization & Customization

Previous sections assume that there is a global agreement on perception, that we perceive tactons in a consistent manner across individuals. Historically, the design of haptic effects for the general audience has been based on the aggregated perceptual characteristics of an assumed "average user." [15] The reality, however, is that people use several cognitive schemas to make sense of, and describe, qualitative attributes of vibrations, typically relying on past *personal* experiences [22, 48]. Therefore, one can assume that tacton perception varies from one individual to the next; this gives rise to the need for mechanisms that can model this variation.

One representational approach to express alternate perceptions of the same vibration is that of *facets* [14]. Akin to taxonomies, facets are alternative perceptions of the same vibration. Because facets naturally incorporate the multiple schemas that people typically make use of, personalizing stimuli can be made easier; however, they do not constitute a distinct instrument for the development of and/or access to personalized haptics.

Another approach is to ask the end-user to tune the haptics patterns themselves, according to their own preferences [49, 9]. Notably, Seifi *et al.* [49] found that participants were not interested in "building" their own haptic signals by combining them, but that there was a significant interest in customizing smartphone notifications in a meaningful way. When asked to customize the vibrations, they were mostly focused on higher-level, higher-impact, coarser changes rather than fine-grained tuning of haptic signals, an indication that users might not be interested in precise personalization but rather more crude modifications.

Yet another approach to end-user customization on the designer side is to combine predesigned tactile building blocks to create a personalized vibration or a set of VT stimuli [46]. For example, to assign haptic alerts to specific events on smartphones, users may choose from a small repertoire of integrated vibrotactile patterns, or tap their own vibra-

tions into the interface.[3] Another example is to morph VT patterns (morphing is defined as constructing a "child" pattern from its parent(s) by deforming it). Clark *et al.* [50] found that Dynamic Time Warping (DTW) was useful and showed promise in constructing new tactons that have predictable features from both parents, but that are also distinguishable in that they are perceptually different from their parents.

To sum up, personalization and customization are important topics when talking about tacton perception, because touch is a very intimate and personal sense: we all perceive, make sense of, and describe haptics based on our own life experiences and our cultural upbringing. Consequently, characterizing this difference is a necessity towards a greater understanding behind our reasoning mechanisms for attributing meaning to VT stimuli [14].

2.2 Evaluating Perceptual Similarity

Similarity plays a major theoretical role in the study of human cognition, building the foundation both for the theory of inductive reasoning [51] and categorization [52, 53].

The degree to which we perceive similarity among a number of things fundamentally affects our rational thought and behavior. As such, similarity is a core element in achieving an understanding of variables that motivate behavior and mediate affect — something that haptics needs to gain access to a broader audience. This is especially true in theories of the recognition, identification, and categorization of objects, where a common assumption is that the greater the similarity between a pair of objects, the more likely one will be confused with the other [2].

2.2.1 Measuring Similarity

Measuring similarity is a complex problem in itself that is still the subject of ongoing discussion in the literature. There are two types of models for measuring similarity: deterministic and probabilistic models. The line between these two alternate views can be blurry: both the percept and the decision process can be probabilistic or deterministic.

[3] https://www.pcworld.com/article/242238/how_to_use_custom_vibrations_in_ios_5.html

The decision process is deterministic if the same information always yields the same response; it is probabilistic if a response is random sampled from a probability distribution at each informational query.

One popular distance-based deterministic technique that leverages similarity judgments is Multidimensional Scaling (MDS), built on the assumption that similarity is inversely correlated to percetual distance [54]: stimuli that are judged by subjects to be similar are close in the perceptual space. This exposes the flaw of these purely deterministic techniques: they provide worthwhile information about the *aggregate* behavior, but they cannot account for variability in the performance of subjects over time or, perhaps, across individuals — many theorists thus argue that percepts are more probabilistic in nature than deterministic [55]. To address this concern, a later statistical procedure called INDSCAL (INdividual Differences SCALing) [56] extended MDS to account for Individual Differences and hypothesized that people give different similarity judgments because of how they weight the various stimulus dimensions. Moreover, a number of machine learning algorithms have been proposed that can learn the similarity metric either via Support Vector Machines [57] or via clustering [58].

Conversely, probabilistic models mostly stem from two assumptions: (1) the percept varies probabilistically over repeated exposures to the stimulus, and (2) there is a well-defined rule that describes how a response is selected for any momentary value of the percept [59]. Because they represent the perceptual space by naturally leveraging IDs, probabilistic models can quantify the uncertainty and the variance that takes places inter- and intra-participant in haptic perceptual studies. This makes them suitable to characterize the uncertainty around tacton similarity perception.

A summary of the different theoretical approaches for measuring similarity is presented in Table 2.2. In this work, we design a model that is purely probabilistic in both percept and decision process, i.e. we develop a Type III model inspired from a mixture between Type I and Type II models (see Section 3.5).

Table 2.2: Classification of similarity psychometric models. Adapted from [2].

		Decision Process	
		Deterministic	Probabilistic
Percept	Deterministic	Type 0 MDS	Type II Logistic
	Probabilistic	Type I Classical Thurstonian psychophysics [59]	Type III Probabilistic extensions of Type II models

2.2.2 Similarity in Music

While the notion of haptic similarity remains somewhat unexplored to its full extent, similarity in music has been explored to great lengths. Indeed, the music literature is a profound inspiration for the analysis of VT stimuli [16].

Several studies have attempted to assess the thresholds at which human beings perceive music to be similar, especially in the context of music recommendation on online platforms, so much so that a whole book has recently been dedicated to the subject [12]. Jehan [60] presented a bottom-up approach to music analysis: by combining several short pieces of musical beats from a sample library (approx. 300 ms), they gathered samples to produce an "audio DNA" of the tracks.[4] Silva *et al.* [61] measured similarity between various *subsequences* of music, thereby improving the cover song recognition problem. Other notable achievements include the works of Casey *et al.* [62], who first used machine learning for sound classification and similarity using Hidden Markov Models, and Cooper *et al.* [63], who developed a method to summarize music by averaging the similarity of segments of entire tracks.

The work done in music similarity serves us as a guide to follow in the haptic domain.

2.2.3 Similarity in VT Haptics

Closest to the present investigation are a number of perceptual haptic similarity studies. Pasquero *et al.* [64] were the first to use MDS in measuring perceptual distances between

[4]This article led to the founding of "The Echo Nest", a company specialized in audio feature design. The company is now owned by Spotify. As a second side note, the idea to represent similarity ratings using a graph also appears in Jehan's "infinite jukebox" (see `http://infinitejukebox.playlistmachinery.com/faq.html`).

tactons. The authors gathered the results from the similarity comparison of a number of simple waveforms rendered as tactons, and concluded that MDS was a suitable algorithm to cluster VT stimuli dissimilarity. However, this approach cannot model annotator error or characterize individual differences of the perception of the tacton. Hwang *et al.* [65] extended the work of Pasquero *et al.* by adding adjective comparison on top of the dissimilarity comparison. Park *et al.* [66] extended the dissimilarities between amplitude-modulated waveforms. They found that the most distinguishable feature to perceptual similarity was the stimulus envelope. Hwang *et al.* [67] extended the framework by comparing superimposed VT stimuli, i.e. two or more sinusoidal signals with different frequencies.

As described in Section 2.2.1, these perceptual studies use MDS as a model for dissimilarity, which does not account for Individual Differences and noise, yet requires a large amount of data for each tacton. Contrarily, our method tackles the problem of haptic similarity from a probabilistic perspective, allowing for statistical error characterization and statistical learning. In addition, this enables us to derive active sampling strategies, thereby increasing the amount of comparisons that can be done given the same time budget (see Sections 3.5.2 and 3.5.3).

Chapter 3

A Bottom-Up Approach to Tacton Evaluation

Preface

This chapter reports on the data-driven approach taken to evaluate the perceptual similarity of VT tactons, and summarizes the technical background necessary for this work. Section 3.1 describes the flipped approach for sampling tactons and the limitations that were placed on their generation.

Next, in Section 3.2, we detail the processes by which human judgement evaluate stimuli. Afterwards, we detail the crowdsourcing (Section 3.3) and data flow (Section 3.4) processes through which participants rate the stimuli. Then, we formalize the probabilistic model used for perceptual similarity in Section 3.5, and detail how this model facilitates actively sampling the tacton space for the most informative tactons.

While our setup would technically be suitable for grouping tactons according to any qualifier (e.g., aggressiveness, urgency, ...), we choose to cluster similarity because (1) in the literature, it has not yet been fully observed at this scale and (2) it is a necessary step towards identifying more complex qualifiers. Knowing that a group of tactons is perceived similarly will accelerate the exploration of the tacton space for other qualifiers.

Author's Contribution

Marc Demers suggested the research direction with the help of Antoine Weill-Duflos and Pascal Fortin. He designed the probabilistic framework around which the research is done, and conducted and supervised all experiments, both in person and through the AMT platform. Initial pilot studies were done with the collaboration of Antoine Weill-Duflos and Pascal Fortin. Hasti Seifi and Oliver Schneider helped design and provide early feedback on the crowdsourcing methodologies as well as the feasibility. Yongjae Yoo further provided feedback on the design of the experiment. Prof. Cooperstock financed the multiple crowdsourced experiments, supervised the research, and proofread the work.

3.1 Exploring the Tacton Space

3.1.1 Introducing the Bottom-Up Approach

In natural language, humans typically use two reading paradigms to infer meaning and process language. Top-down processing of language refers to using background information to predict the meaning of language. On the other hand, bottom-up processing of language occurs when affixation is used to guess the meaning. In language, this is typically used in conjunction with context to understand a text.

In haptics however, practitioners commonly design tactons and then conduct user studies to evaluate their hypotheses about their perception by the users. This process is cumbersome, does not scale well to a large number of tactons, and makes the extrapolation of the influence of the tacton characteristics on our perception difficult to evaluate. This hinders the ability to scale haptics to a level where it could have positive practical influence on perceptibility [68], to customize tactons in social chat applications [69], or to encode smartphone interaction parameters [18, 70]. Because they touch on multiple constraints such as device diversity and human perception, these use cases are very broad in nature and require extensive studies with large amounts of data, and may require novel methods for gathering that data at scale.

Equally important, very little research has been done where small affixations or subtractions are added or removed from tactons to examine the influence of low-level features on our perception of VT tactons. Close to this idea however, there have been attempts at performing blending of multiple tactons, also called *morphing* [50].

In this context, we plan to examine the influence of low-level tacton characteristics on our perception of similarity. As depicted in Figure 3.1, we propose to flip the design approach of tactons. We browse the space of all possible tactons, evaluate them in an iterative yet efficient procedure, and analyze their perceptual characteristics *a posteriori* through data mining. This effectively bypasses the cumbersome "design" phase, and yields greater insight into the perceptual space. Consequently, there is no need for predefined sets of tactons before the experiment: tactons are randomly generated on-the-fly, alleviating the bias produced by the experimental methodology.

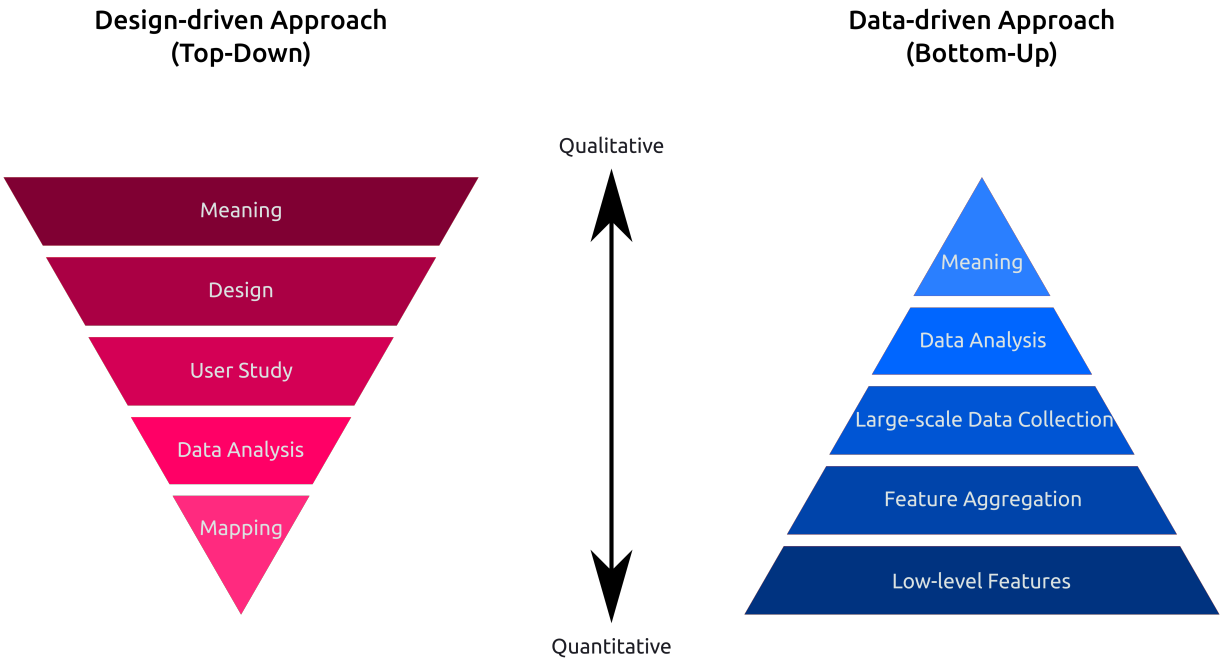


Fig. 3.1 The proposed bottom-up approach to designing tactons, as opposed to the traditional top-down approach.

Nonetheless, the proposed approach poses a problem, which is that VT icon ratings are typically very noisy [8]: the hope is that larger sample sizes may help reduce or simply characterize this error. This characterization would also give support to the "need to develop mechanisms for individual customatization" [11]. There could be groups of users who share preferences and interpretations in terms of haptic perception, but this has yet to be determined.

All in all, the bottom-up approach offers promise for helping solve the many challenges affecting the haptic world. In the rest of this thesis, we develop a haptic-specific evaluation tool that allows for a scalable data collection approach to identify hundreds of randomly generated tactons and report on the findings.

3.1.2 Tacton Rendering

Tactons are typically rendered on dedicated actuators. A number of years ago, the main limitation to overcome to successfully bring about large-scale haptic data collection was

the low availability of dedicated actuators that can render multiple patterns. Fortunately, the recent mass production of smartphones with haptic actuators is a convenient proxy for evaluating haptics on a larger scale and for improving our comprehension of our perception of VT icons.

Unfortunately, smartphones' actuators are not easily accessible for low-level control. For instance, at the time of writing, Apple does not yet provide an API to directly control the haptic actuator in their smartphones. On Android, Google provides an API to control the actuator, but some (older) smartphones have actuators that do not support custom vibrations other than binary control (on/off timings). On a large scale, this restricts the potential space of tactons that can be generated, and thus, evaluated. Nevertheless, as is the case with pulse-width modulation, we can simulate more complex waveforms with simple on-off triggering of vibrations. It has also been found that when asked to tweak or modify stimuli to fit a desired intent, users typically preferred coarser changes to VT icons than smaller, more subtle changes [49]. With the intent of designing our experiments for the broadest audience possible, we generate what we call *random binary* tactons, in the sense that they are random sequences of on/off vibrations rendered by the haptic actuator.

3.1.3 Tacton Characteristics

Due to the exploratory nature of study and to mitigate its complexity, we restrict the dimensionality of tactons in several aspects. First, Ternes *et al.* suggest that the greatest useful length of tactons is two seconds [38]. Second, we effectively only modify the frequency parameters of the stimuli, although both parameters (frequency and amplitude) are intermingled and altering the frequency of a signal also tends to change the perceived intensity [71]. It has also been suggested that designers vary only one of these two parameters when conducting experiments [72]. In addition, Nukarinen [73] notes that amplitude is a complicated parameter for encoding information. Given the novelty and the scale of the bottom-up approach, we elect to only vary the frequency of the stimuli.

Third, because we use smartphones as rendering proxies for the tactons, we expect all participants to feel the VT stimuli in the palm of their hands, thus constraining the spatial

location. Last, we use stimulus bins of 100 ms to vary the on/off status of the haptic actuator. This is long enough to resolve the temporal separation of 10 ms [74], yet short enough to carry meaningful content across two seconds of stimulus. Given one stimulus is 2 seconds in length, this gives twenty 100 ms bins per stimulus.

The aim of these constraints is to simplify the problem to fit the task of rendering tactons on a multitude of different smartphones, as well as simplifying the infinite space of tactons to something tractable and meaningful given previous literature's suggestions to designing haptic experiments.

3.2 Discrete Choice Analysis

Subjective data from humans is typically collected in one of two forms: *explicit/direct* and *implicit/indirect* labeling. Explicit labeling refers to collecting data from a data point itself (absolute data collection, single ratings), whereas implicit labeling refers to collecting data from a comparison of two or more elements (relative data collection, multiple ratings). The difference between score-based approaches and comparison-based approaches exemplifies a "global" vs a "local" view: a score is global, a comparison is local. In other words, in a multiple comparison scheme, a smaller number of alternatives N will lead to a more granular view of the space, while a greater number of alternatives will provide a coarser view of the space. The former is ideal in cases where the score can be defined naturally in terms of measurable utility. In most real-world scenarios however, an interpretable score (i.e. a Likert scale) can be difficult to define, or is simply nonexistent.

Alternatively, comparison-based approaches are usually done in adversarial contexts, for instance in online game matchmaking, in image/video quality assessment, in recommendation systems, or in chess (the ELO rating system is an example of a popular comparison-based approach). A well known phenomenon in the psychological study of human choice claims that human response to comparison questions is more stable in the sense that it is not easily affected by irrelevant alternatives [75]. Also, from an information theoretic point of view, you gain more information by asking which of two or more data points fits your desired label best, rather than placing them on a predefined scale: the latter does not allow to decide among points that would have been placed at the same

spot. This is not to say that implicit labelling is exempt of inconsistencies or is robust to noise: the annotator's expertise, their emotional state, or external factors such as the environment or their demographic background all play a role in judgement making.

On the other hand, the disadvantage of the multiple comparisons scheme is that as the amount of items to be compared, n, grows bigger, the amount of pairwise labels needed to infer full comparison setting grows exponentially to $\mathcal{O}(n^2)$. This makes settings in which n is large difficult to work with, especially considering that human annotations are expensive and time-consuming.

For the reasons described above, we elect a comparison scheme where we present between 6 to 8 tactons[1] at once to the annotator and asked them to group the tactons in *at least c* groups. This multiple comparison scheme also avoids biasing the annotators with regards to the task by presenting them with a richer sample of data points in the beginning of the experiment.

We provided a lower bound on the number of groups for a specific reason: in decision theory, the Independence of Irrelevant Alternatives (IIA) axiom implies that adding another option to two options does not affect the relative odds between the two options considered. We deem this axiom unrealistic in our scenario. We relaxed it by allowing annotators to produce singletons as groups, as long as the lower bound on the amount of groups is respected.

3.3 Outsourcing to the Crowd

Paid platforms for outsourcing experiments (i.e. deploying them on a large scale) such as Amazon Mechanical Turk[2] (AMT) remunerate participants (informally called *turkers*) for the work done. We designed a crowdsourced study to collect a large amount of data points through AMT.

[1] In the majority of the experiments that were conducted, we chose $n = 6$ taking inspiration of Miller's Law of 7 ± 2. Although we acknowledge that this law was later revised to lower amounts based on the cognitive load of the task [76], we experimented with several values and found that it led to consistent results among annotators (see Section 4.1.3). In addition, tactons can be replayed as many times as the annotator desires during the experiment, and we found that annotators typically solve the task by comparing the n stimuli in pairs and clustering them accordingly. Consequently, we expect these aids to lower the cognitive load of the task.

[2] `https://www.mturk.com/`

It is necessary to keep in mind that while the outreach of those types of studies is attractive because it allows to collect a considerable amount of data points in a short amount of time, one needs to be wary of several drawbacks. Compared to a quiet lab, the remote and asynchronous environment of AMT is not controlled, which leads to higher noise in the data. AMT participants have various profiles and reasons for providing data and participating in experiments, which makes them exhibit specific behaviors, especially in tasks of mapping subsets of items to categories [77, 78].

According to Eickhoff and de Vries [79], there exist primarily two type of turkers. The first are "entertainment-driven" workers, for whom the financial part of the task is not so important, and who work on tasks mainly for the challenge that they pose. The second are "money-driven" workers, for whom the monetary incentives are greater. For this category, their chances of cheating the task or providing noisier responses is greater. There is also evidence that there exists a third type of worker: the curious worker [80], who is incentivized by either curosity-inducing stimuli during a long task, or the task is designed in such a way to incentivize the worker not to quit and to complete more tasks (this resembles the idea of gamification of the task).

In order to fend off most "money-driven" workers, requesters must incorporate robust quality control techniques into their task on AMT to ensure that the data gathered is consistent and valid [81]. Quality control measures include but are not limited to: simplifying tasks to one single activity and using deterrents to prevent participants from cheating or not concentrating. Additional quality controls include design-time approaches (effective task preparation and active worker selection), runtime approaches (input agreement, ground truth labeling, majority consensus) [82], and ensuring that Turkers report problems, disabilities, and technical difficulties that might have impacted their performance during the experiment [83]. It is believed that this honest self-reporting allows for trust-building between the Requester and the worker pool.

Concretely, fend-off strategies to decipher cheating workers ordinarily consist of the use of gold-standard data (to discover arbitrarily picked answers) and the use of petty awareness questions (to catch annotators that are not paying attention) [82]. The aim behind these strategies is to ensure that data pollution is at its lowest, and to discourage malicious annotators or spammers to assign the wrong label to data. We elaborate about

the implementation of the deterrent strategies — which we refer to as *attention tests* — in Chapter 4.

In the field of haptics, to our knowledge,there is only one iteration of outsourcing haptic VT tactons to the crowd: HapTurk [11]. The authors designed a new method for proxying VT stimuli via the crowd, with the hope of empowering VT designers with the benefits of crowdsourcing and large-scale data collection. We consider this work a starting point to the present study and proof that outsourcing haptic VT stimuli to the crowd is a realistic framework for exploring haptic perception.

3.4 Data Flow

We designed a web architecture that allows for collecting ratings of haptic VT stimuli from smartphone devices. Each participant downloads the Android application from the Google Play Store,[3] which then connects to our central server and database (see Figure 3.3). Upon query, the tactons are generated on-the-fly through an API and sent to the user for annotation.[4] The concept of a "vibration API" has been recently explored in the context of web-based hapticons [84]. The annotation process takes place directly on the smartphone. Figure 3.2 shows the process by which annotators rate the similarity of a number of tactons.

Because portability and reusability of the system are a priority, the design of the application also takes into account version iteration and task refinement, as is recommended in crowdsourced user studies where iterating on your user base is considered a robust practice.

The application supports:

1. Version control: Android applications update automatically by default, so we provide participants with a password unique to that experiment to enter upon opening the application. Workers only have access to the password once they accept the Human Intelligence Task (HIT).

[3] `https://play.google.com/store`

[4] While in this context the data collection task is for similarity ratings, one could easily extend the process to preference ratings (i.e. this tacton feels more *urgent* than this other one) among a number of tactons.

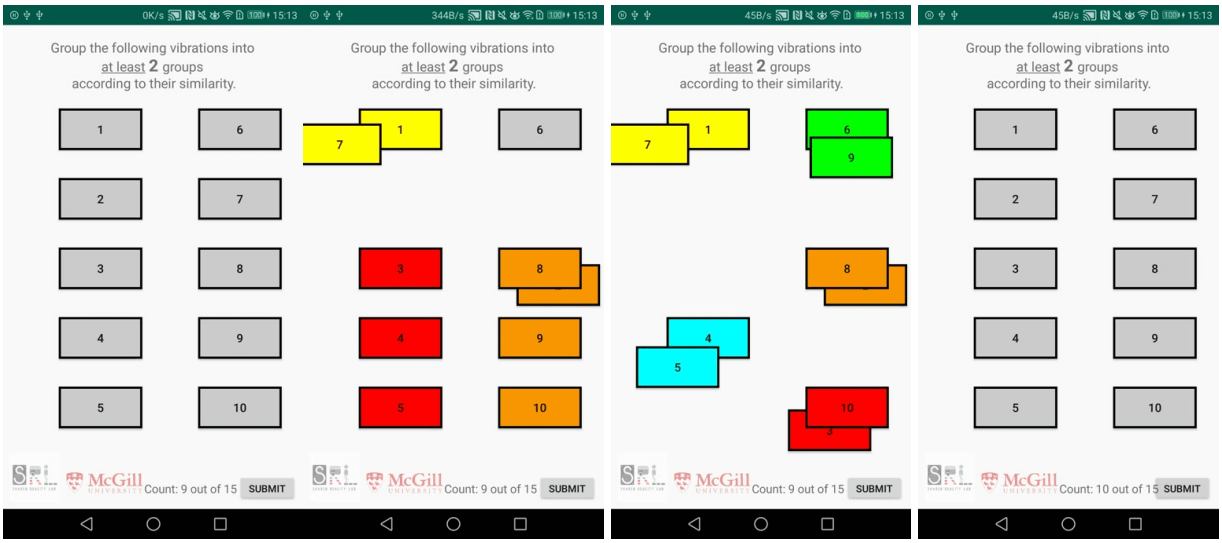


Fig. 3.2 One round of grouping the tactons into similarity clusters. The user iteratively evaluates the similarity between the tactons and places them into clusters. Once they are satisfied with it, they submit the grouping. This process repeats for R rounds.

2. Participant validation: The Unique User ID number (UUID) is only created once the user has entered the valid password for that particular experiment. This avoids populating the database with workers that are not active participants in the study.

3. Persistently anonymous UUIDs: The UUID changes for a same participant with the application versions. This means that the participant remains anonymous and cannot be retroactively discovered regardless of the application version.

Once the maximum amount of annotations has been reached, the application provides a secret key as a completion token. This completion token is logged into the database and is used to link the participant's e-mail (in the case of real participants), or their worker ID (in the case of AMT workers), to the completed HIT. The process can run in parallel for a number of participants, without interruption (batch-crowdsourcing). As indicated in Figure 3.3 and due to this parallelism, we make use of checksums between the server and the participants' smartphones in order to ensure that the right ratings are matched with the right haptic signals and the correct user.

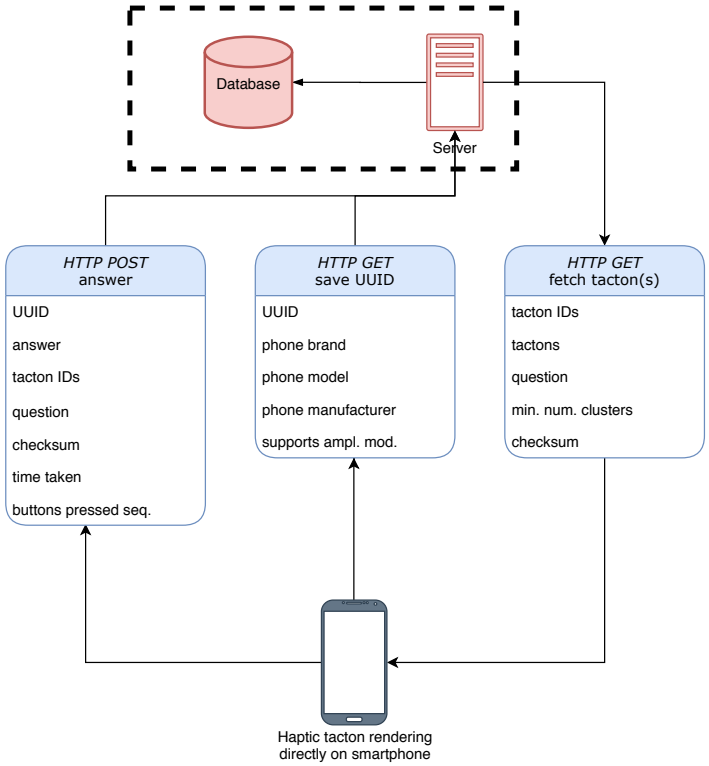


Fig. 3.3 Overview of the software architecture behind the haptic rating system.

3.5 Probabilistic Modeling for Similarity Evaluation

Because similarity rating implies an adversarial setting, we take inspiration in binary choice models such as the Bradley-Terry model for paired comparisons [85] to develop a probabilistic model of tacton similarity. Contrarily to the BT model however, we do not attempt to infer a global ranking between the test candidates; rather, our objective is merely to assess the similarity between the pairs of objects.[5]

[5]Because similarity measures are bidirectional, i.e. object A is similar to object B implies object B is similar to object A, there is no need to infer a global ranking of all the test candidates. It would, however, be possible to infer a global ranking on the similarity *pairs*, i.e. the similarity between objects A and B is greater than between objects A and C.

3.5.1 Definitions

We denote the whole tacton space T. We denote a candidate tacton $T_i \in T$, and pairs of candidates $\{(T_i, T_j) \in T, i \neq j\}$. We also extend this scheme to the multiple comparison of candidates by considering all the pairwise comparisons in set M, $\{(T_i, T_j)|i, j \in M, i \neq j\}$. On each round $r \in R$, the annotator produces c similarity groups of T, and the outcome y is decomposed into a series of pairwise comparisons of all $\{T_i, T_j\} \in T$, $i < j$.

At round r, the outcome of each pair, y_{ij}^r, follows a binomial random variable. The outcome can be that the pairs are similar (then $y_{ij}^r = 1$) or that the pairs are dissimilar (then $y_{ij}^r = 0$). In that sense, a rating of similarity represents a success while a rating of dissimilarity represents a failure. As one can imagine, in the multiple comparison scheme, in a round, we expect the amount of similar pairs will be much lower than the amount of dissimilar. The intuition behind this is that the mean of a Bernoulli variable is the relative frequency of the events in the data, and this constitutes the main information content of the data set [86]: this can lead to biases in the probability estimates move in the direction of the bias. To avoid having this imbalance in our data, we weighed the outcomes by a factor w_{ij}^r such that the sum of the outcomes is equal to the number of groupings (c) performed by the annotators, both in the similarity ratings and in the dissimilarity ratings.

$$w_{ij}^r = \begin{cases} \dfrac{c^r}{\tau_s^r}, & \text{if } T_i \sim T_j,\ i < j \\ \dfrac{c^r}{\tau_d^r}, & \text{if } T_i \nsim T_j,\ i > j \\ 0, & \text{otherwise} \end{cases} \tag{3.1}$$

$$\tau_s{}^r = \sum_i^m \sum_j^m \mathbb{1}\{T_i \sim T_j\}, i \neq j$$

$$\tau_d{}^r = \sum_i^m \sum_j^m \mathbb{1}\{T_i \nsim T_j\}, i \neq j$$

where $\sim$ represents similarity, $\mathbb{1}$ is the indicator function, τ_s^r is the amount of pairs of tactons that were rated as being similar, and τ_d^r is the amount of pairs that were rated as being dissimilar at round r.

We pose the similarity Pair Comparison Matrix (PCM) as a positive square matrix A in which each {row, column} ({i, j}) locus on the upper triangle ($i < j$) represents the similarity weight α_{ij} for objects T_i and T_j, and every locus on the lower triangle represents their dissimilarity weight α_{ji} ($i > j$). Figure 3.4 shows a graphical visualization of how we fill the PCM matrix A in a single round using the weighing scheme in Equation 3.2.

We parameterize the error on each annotation outcome y_{ij}^r between tactons T_i and T_j due to extrinsic factors (i.e. human error) and intrinsic factors (i.e. human judgment) by a Gaussian random variable ϵ_{ij} [87]. We model the performance of the annotator on the attention test at round r, η^r, as a surrogate for the reliability of the rating [88]. In practice, we use $\eta^r = 1$ for rounds where the attention test was a success, and $\eta^r = 0.5$ for rounds where the attention test was a failure. The pair similarity value α_{ij} is thus defined as the sum of weighted outcomes up until round R:

$$\alpha_{ij} = \sum_{r=0}^{R} \eta^r w_{ij}^r y_{ij}^r + \epsilon_{ij}, \qquad \epsilon_{ij} \sim \mathcal{N}(0, \sigma_{ij}^2) \tag{3.2}$$

3.5.2 Probabilistic Model

We find the probability of objects T_i and T_j to be rated as similar through logistic binomial regression,[6,7] such that

$$P(A_i \sim A_j) = \hat{s}_{ij} = \frac{1}{1 + e^{-(\alpha_{ij} - \alpha_{ji})}} = \frac{e^{\alpha_{ij}}}{e^{\alpha_{ij}} + e^{\alpha_{ji}}}, \; i < j \tag{3.3}$$

We fit the data to the model by maximizing the associated likelihood function:

[6]The logit scale is relevant in the context of similarity where one item is compared against one or multiple alternatives, and in which the outcome can be represented on a scale from 0 to 1 (the predicted odds). This representation also works well when the objects for which choices are expressed constitute a "small world", where attention is confined to a limited and predefined set of static choices [89]. Note that while we choose the logit link in the interest of simplicity, one could elect to link according to the probit function with very similar outcomes.

[7]Although the observed variable is a weighted function of a binomial random variable, the logistic function converts the output of the regression back into the predicted odds on a continuous scale.

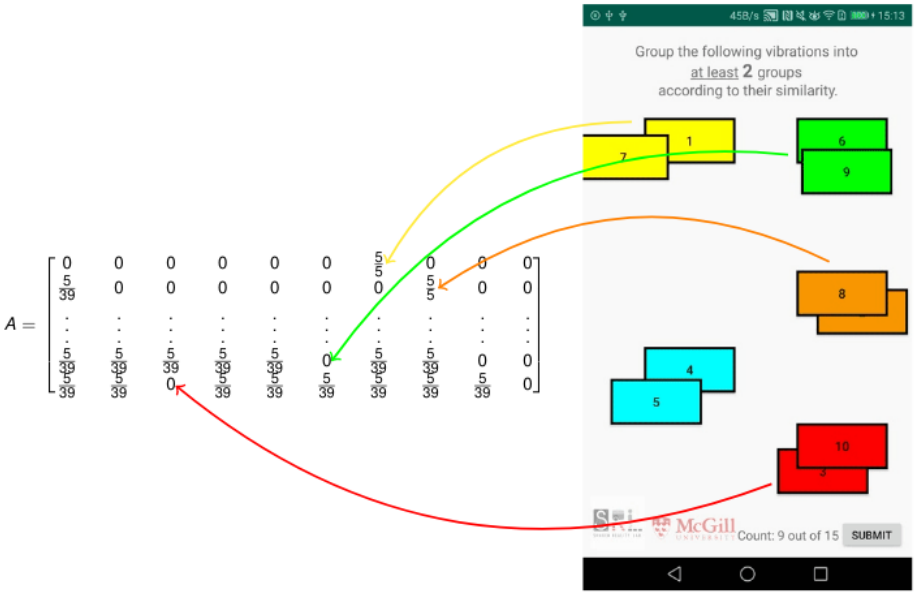


Fig. 3.4 Graphical explanation of the weighing scheme for the PCM matrix A. Note the annotator performance weight η^r is not depicted here.

$$\begin{aligned}\mathcal{L}(s \mid A) &= \prod_{i<j} s_{ij}^{\alpha_{ij}} (1-s_{ij})^{1-\alpha_{ij}},\ i<j \\ &= \prod_{i<j} s_{ij}^{\alpha_{ij}} (s_{ji})^{\alpha_{ji}},\ i<j\end{aligned} \tag{3.4}$$

There are two conditions that are required for this maximum likelihood to be valid. First, each partition of objects must be separable into two nonempty subsets such that some object in the second set has been preferred to at least one object in the first set [90]. For this very reason and because we do not know the number of annotators upfront, we grow the number of tactons (and thus the PCM) incrementally during the experiment. Second, the graph represented by the PCM must be strongly connected [88]. To ensure full connectivity across the graph, we initialize the PCM with virtual nodes with one success and one failure.

Assuming the outcomes y_{ij} are independent in probability across time and across annotators, we can define the variance σ_{ij}^2 of the binomial similarity probability s_{ij} as:

$$\hat{\sigma}_{ij}^2 = \hat{s}_{ij}(1-\hat{s}_{ij}) = \frac{\alpha_{ij}\alpha_{ji}}{(\alpha_{ij}+\alpha_{ji})^2} \tag{3.5}$$

By maximizing the log-likelihood in Equation 3.4, we can obtain the Maximum Likelihood Estimators (MLE) of the similarity pairs $\hat{s}_{ij}$ with their associated variance $\hat{\sigma}_{ij}$:

$$\zeta_{ij} \sim \mathcal{N}(\hat{s}_{ij}, \hat{\sigma}_{ij}) \tag{3.6}$$

Note that the σ parameter in Equation 3.6 does not depict the use of a multivariate normal distribution in the sense that it does not describe the joint interactions between T_i and T_j (due to their independence), rather, it characterizes the variance of the similarity ratings *between two tactons*. This notation is used to depict several *independent* normal distributions for notation purposes only.

3.5.3 Maximal Information Gain

One unique issue that arises when employing the bottom-up approach to sampling is the fact that the space of all tactons is unimaginably large and the estimated time to uncover similarity scores s_{ij} from those test candidates remains problematic. For instance, the time required to uncover the pairwise similarity scores of n test candidates is quadratic ($\mathcal{O}(n^2)$). While randomly sampling the space and hoping for randomness to give rise to useful information about our perception of tactons could work, we instead elect a sampling strategy that is rooted in information theory to sample data points in an intelligent and efficient manner so as to reduce this time complexity.

More formally, in the active learning case, we are interested in finding the pairs of annotations that would yield the highest Expected Information Gain (EIG). The pairwise comparison active learning literature typically focuses on using global information to efficiently sample the space of possible comparisons [88]. This strategy works efficiently for sampling pairs of items where there is an implicit correlation (i.e. covariance) between pairs of items. Similarity ratings are independent of one-another, so we need to look at local information to efficiently sample the tacton space. We use an approach similar to Li *et al.* [91] to sample pairs of items but adapt it to the case where we do not wish to infer a global ranking of test candidates.

We define the expected Kullback-Leibler Divergence (KLD) between the prior probability distribution $P(\zeta_{ij})$ and the posterior distribution given the current outcome $P(\zeta_{ij}|y_{ij})$ as a surrogate to the EIG to be the distance function D_{ij}:

$$D_{ij} = D_{KL}(P(\zeta_{ij} \mid y_{ij}) \parallel P(\zeta_{ij})), \quad i < j \tag{3.7}$$

$$D_{ij} = \int \sum_{y_{ij}} \log \frac{P(\zeta_{ij} \mid y_{ij})}{P(\zeta_{ij})} P(\zeta_{ij} \mid y_{ij}) P(y_{ij}) \partial \zeta_{ij}, \quad i < j \tag{3.8}$$

According to Bayes' theorem, we can rewrite Equation 3.8 as:

$$D_{ij} = \int \sum_{y_{ij}} \log \frac{P(y_{ij} \mid \zeta_{ij})}{P(y_{ij})} P(y_{ij} \mid \zeta_{ij}) P(\zeta_{ij}) \partial \zeta_{ij}, \quad i < j \tag{3.9}$$

where $P(y_{ij}|s_{ij})$ is the conditional probability of outcome y_{ij} when comparing T_i and T_j at round r.

We define $p_{ij} = P(y_{ij} = 1 \mid \zeta_{ij})$ and inversely $q_{ij} = P(y_{ij} = 0 \mid \zeta_{ij})$; it follows that $P(y_{ij}) = \mathbb{E}(p_{ij})$ and $P(y_{ji}) = \mathbb{E}(q_{ij})$.

Given that we only have two possible outcomes y_{ij}, one can simplify Equation 3.9:

$$D_{ij} = \int \left[\log \frac{p_{ij}}{\mathbb{E}(p_{ij})} p_{ij} + \log \frac{q_{ij}}{\mathbb{E}(q_{ij})} q_{ij} \right] P(\zeta_{ij}) \partial \zeta_{ij}, \quad i < j \tag{3.10}$$

According to Equation 3.6, the similarities follow a Gaussian distribution with mean $\hat{s}_{ij}$ and variance $\hat{\sigma}_{ij}$. The probability density function of the similarities is therefore:

$$P(\zeta_{ij}) = \frac{1}{\hat{\sigma}_{ij}\sqrt{2\pi}} e^{-\frac{(\zeta_{ij} - \Phi(\hat{s}_{ij}))^2}{2\hat{\sigma}_{ij}^2}}, i < j \tag{3.11}$$

where $\Phi(\cdot)$ is a function that maps the logistic output range $[0, 1]$ to a hyperbolic tangent range $[-1, 1]$ such that the mean of the normal distribution is not biased towards similarities. Setting $y^2 = \frac{(\zeta_{ij} - \Phi(\hat{s}_{ij}))^2}{2\hat{\sigma}_{ij}^2}$, we have $\zeta_{ij} = \sqrt{2}\hat{\sigma}_{ij} y + \Phi(\hat{s}_{ij})$ and $\partial \zeta_{ij} = \sqrt{2}\hat{\sigma}_{ij} \partial y$. Using Equations 3.3 and 3.11, Equation 3.10 can be rewritten as a sum of integrals:

$$D_{ij} = \int f_1(y)e^{-y^2}\partial y - \int f_2(y)\,log f_2(y)e^{-y^2}\partial y + \int f_3(y)e^{-y^2}\partial y - \int f_4(y)\,log f_4(y)e^{-y^2}\partial y \tag{3.12}$$

$$f_1(y) = \frac{1}{\sqrt{\pi}} \frac{1}{1+e^{-(\sqrt{2}\hat{\sigma}_{ij}y+\Phi(\hat{s}_{ij}))}} \log \frac{1}{1+e^{-(\sqrt{2}\hat{\sigma}_{ij}y+\Phi(\hat{s}_{ij}))}}, \qquad i<j \tag{3.13}$$

$$f_2(y) = \frac{1}{\sqrt{\pi}} \frac{1}{1+e^{-(\sqrt{2}\hat{\sigma}_{ij}y+\Phi(\hat{s}_{ij}))}}, \qquad i<j \tag{3.14}$$

$$f_3(y) = \frac{1}{\sqrt{\pi}} \frac{1}{1+e^{\sqrt{2}\hat{\sigma}_{ji}y+\Phi(\hat{s}_{ji})}} \log \frac{1}{1+e^{\sqrt{2}\hat{\sigma}_{ji}y+\Phi(\hat{s}_{ji})}}, \qquad i<j \tag{3.15}$$

$$f_4(y) = \frac{1}{\sqrt{\pi}} \frac{1}{1+e^{\sqrt{2}\hat{\sigma}_{ji}y+\Phi(\hat{s}_{ji})}}, \qquad i<j \tag{3.16}$$

Equations of the form $H(x) = \int_{-\infty}^{\infty} f(x)e^{-x^2}\partial x$ can be solved numerically using the Gauss-Hermite quadrature. We solve each term in Equation 3.12 individually to compute the EIG. In this study, we use 20 sample points for computing the approximation.

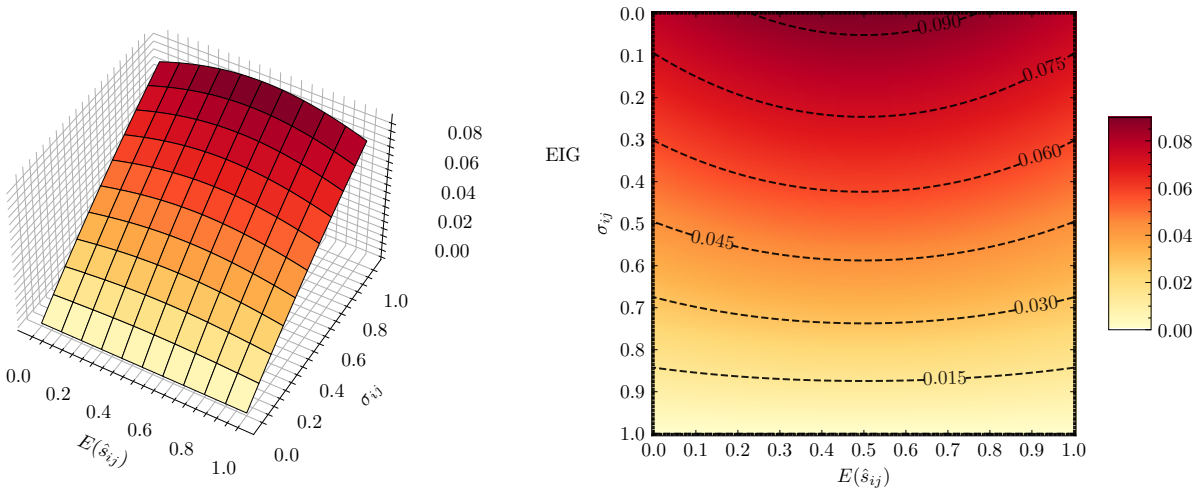


Fig. 3.5 EIG as a function of the mean rating between two tactons and its associated variance. As expected, pairs that have as many similarity ratings as dissimilarity ratings exhibit the highest EIG.

As seen in Figure 3.5, the EIG is maximal for pairs that have similar scores s_{ij} and s_{ji} and whose variance σ_{ij}^2 is large, and is minimal for pairs that have different scores and small σ_{ij}^2. This is consistent with the intuition that we should sample from pairs that have maximal *uncertainty*, meaning pairs that do not exhibit clear similarity or dissimilarity.

For active sampling, we select the batches of tacton pairs that yield the greatest EIG using the minimum spanning tree (MST) of $-D_{ij}$. We add a new tacton in the global mix when the batches from the MST are exhausted. Advantages of using the MST include a lower computational budget and the possibility of querying *batches* of pairs of tactons. For a more thorough explanation and in the interest of space, we refer the reader to Li *et al.* [91].

The tactons are then sent to the annotator for evaluation, and the outcomes are weighted according to the procedure defined in Equation 3.2. The process repeats for subsequent rounds.

3.5.4 Graph Representation

The PCM square matrix A can be viewed as the adjacency matrix of a directed acyclic graph. However, using logistic regression, we can represent the same information on an undirected acyclic weighted graph $G = (V, E, W)$ with edge-weight function set $W : E \rightarrow [0, 1]$, where the vertices (V) represent tactons, the presence of edges (E) represents whether the pair was compared, and the edge weights (W) represent the degree to which the objects in V are similar (a weight of 0 is extremely dissimilar, and 1 is extremely similar), as depicted in Figure 3.6. The edge weights W are obtained via binary logistic regression, as described in Equation 3.3.

This type of graph representation showing similarity has been used in several domains such as audio [92], social networks [93], and computer vision [94], among others. Graphs put emphasis on the structural relationships between items, and are typically used to represent information about the topology of the data.

Of notable mention is the work of Perraudin *et al.* [92], who constructed an audio similarity graph, where each vertex is a segment of musical content, and the edges represent the similarity between the segments. Note that they consider segments of musical tracks, which is consistent with the concept of tactons (*short* VT patterns).

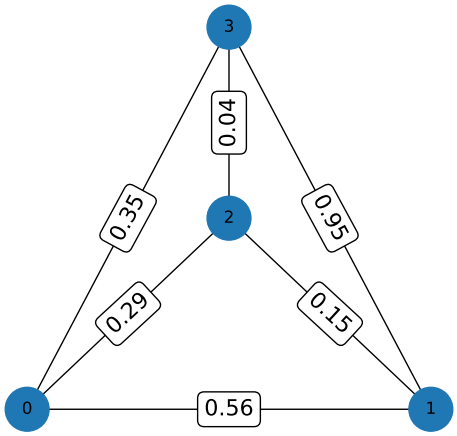


Fig. 3.6 Tacton similarity as edge weights in a graph.

3.5.5 Machine Learning on the Graph

We intend to go further than just establishing the similarity clusters and a hierarchy among them. We want to extrapolate the similarity to pairs of tactons that have not been evaluated during the experiment, that is, predict similarity between never-seen-before tacton pairs. Thus, we create a model that can predict the edge weights from the similarity graph.

The intuition behind edge weight prediction on a graph network is simple: suppose we have three tactons, the first two of which we know are similar, and we also know that the first is similar to a third, we can leverage that information to infer similarity between the second and third. The above example is one of first-order proximity, but it can easily be extended to the nth-order node proximities (i.e. their n-degree neighbors). The process of going from node to node along the edges of a graph is called "walking". In this study, we make use of Graph Convolutional Networks (GCN) [95] to perform these n-order walks on the graph. GCNs naturally integrate node features into the learning process, and high-order walks are made possible by stacking multiple layers.

In statistical inference and supervised learning, one typically splits the data into a training and a testing set, such that there is no overlap between the two: this is supervised learning, which is linked to inductive reasoning. A concern is that, in contrast,

representation learning on graphs typically assumes a fixed set of nodes to predict edge information, which is inherently transductive reasoning.

In more mathematical terms, an inductive algorithm aims to learn a function $f : X_{train} \to Y_{train}$, and inference will be made by evaluating $f(x_i)$ for all x_i in the test set. On the other hand, a transductive algorithm aims to learn a function $f : X_{train} \times Y_{train} \times X_{test} \to Y_{test}$, and the predictions follow from this function. In that sense, transductive learning is closer to semi-supervised learning.

However, in reality, many graphs are evolving: new nodes are added over time. To allow our graph model to extrapolate to never-seen-before pairs of tactons, i.e., create new similarity links, we must utilize techniques that do not rely on the whole graph to infer node embeddings; rather, we must rely on techniques that only consider a node's *neighborhood* when performing the walks. One such technique is that of GraphSAGE [96] convolution operators. It follows that there is no need to retrain the whole model when adding a new tacton: the prediction can be performed on-the-fly.

Chapter 4

Experiments

Preface

In the previous chapter, we introduced a data-driven methodology to gather efficient and reliable haptic ratings in a crowdsourced setting. This methodology is now applied to large-scale experiments on AMT.

We survey participants with the objective of mapping perceptual similarity in two parts:

1. Getting a consensus on what tactons are similar across a large number of tactons and of annotators.

2. Attempting to extract groups of people who share haptic perceptual similarity.

We refer to the former as the "global" experiment, and the latter as the "persona" experiment. Although we acknowledge that, in traditional Human-Computer Interaction, personas are intended to be descriptions of a specific aspect or area of focus of an archetype, here we will use the word in the broader context of "a group of people who share perceptual similarities."

Section 4.2 is dedicated to the global experiment, Section 4.3 to the persona experiment, and Section 4.4 links the two together.

All experiments were approved by McGill University's Research Ethics Board Office,[1] REB file #432-0416. To take part in the experiment, participants were required to acknowledge and accept a consent form.

Author's Contribution

Marc Demers proposed the research direction, and conducted all experiments. He received feedback from David Marino, Jeffrey R. Blum, Pascal Fortin, Yongjae Yoo and Antoine Weill-Duflos for the analysis of data.

[1] https://www.mcgill.ca/research/research/compliance/human

4.1 Methodology

4.1.1 Participants and Remuneration

All participants were recruited via AMT. We did not impose restrictions on geographical location or AMT status (Master workers, ...). We implemented qualifications to avoid participants submitting multiple HITs of the same experiment.

We compensated participants in accordance with the minimum hourly wage in Canada (converted to USD), proportionally to the estimated duration of the experiment, which we estimated to be 10 to 15 minutes. Accordingly, participants were compensated USD 0.40 for submitting a HIT, and were further awarded a bonus payment of USD 1.60 for obtaining an accuracy score of 80% or greater on the attention tests.

4.1.2 Concurrency

The objective of this study is twofold: find a global agreement on VT perceptual similarity, and extract individual characteristics of participants to obtain a better model of similarity perception. Given the intricate links between those two objectives, we decided to conduct both experiments concurrently. This concurrency is opaque to the participant; from their point-of-view, there was only a single experiment with multiple rounds of the same task.

To test the global agreement on VT similarity, we ran an experiment where we asked the participants to cluster tactons into a number of groups according to how "similar" they perceived them. In this setting, we evaluated similarity iteratively with the active sampling procedure described in Section 3.5.3. To extract personas, we ran the same procedure, but the tactons to group are the same across all iterations, across all participants. The hope is to identify groups of perceptual similarity among the participants from the repeated measures on the same task.

4.1.3 Pilot Studies

We conducted several pilot studies in order to assess the difficulty of the task for AMT workers, the summary of which are available in Table 4.1. We first attempted, without success, to evaluate ratings from a predefined clustering of tactons based on character-

istics that have previously been found to be meaningful in similarity evaluation in the literature. We later switched to the bottom-up approach in a simple pairwise comparison approach, in which the participants were only presented two stimuli at a time and were asked to rate them. We also set out to implement a calibration phase during which the participants were presented with toy examples of similar and dissimilar tactons. We noticed a heavy bias in the participant population after having implemented these measures. The bias stemmed from the fact that the number of samples of tactons pairs that were presented in the calibration phase was limited to six.

This made the subspace of tactons to which the participants were exposed non-representative of the global population of tactons. We did not wish to have the participant population simply replicate the similarity ratings in a calibration phase, as this would go against the objective of mapping *perceptual* similarity. Instead, we are looking for tacton similarity without inducing any *notion* of what is similar and what is not.

The final version of our experiment presented the users with a larger amount tactons at a time, without a calibration phase, and with a more detailed description of the high-level objectives of the experiment.

Iteration #	Exp. Type	# of HITs	Description	Takeaways
1	AMT	18	We clustered tactons according to their similarity in features, asked workers to rate the similarity and compared the agreement between the two clusterings. Participants were asked to report on 60 randomly selected tacton pairs.	The clusters were very seldom in agreement with our predefined clusters.
2	AMT	9	Given that the results were underwhelming, we gave the same task to a number of AMT workers that have the "Masters" qualification, meaning that they are known to pay attention to HITs.	Master workers could not identify the clusters either, indicating that the task was either badly designed or too hard.
3	AMT	30	In an attempt to clarify what was meant by "similarity" ratings, we implemented a calibration phase where the workers were shown toy examples of pairs of tactons and given the "true" answer before beginning with the experiment. We also implemented attention tests to see if the participants was trying to bypass the experiment for financial incentives.	Calibration introduced a bias in the experiment (see Section 4.1.3), but stabilized the results. Attention tests helped identify workers who did not pay attention during HITs.
4	live	13	We compared the above scheme with live participants, in an in-the-lab experiment.	The performance of live participants was found to be very similar to AMT workers on this type of task.
5	AMT	70	Due to poor performance, we discarded the idea of *a priori* clustering the tactons with respect to the similarity in their features, and used the bottom-up pairwise comparison scheme instead.	The pairwise comparison scheme allowed to extrapolate features from tactons and group them without setting any *a priori* clustering.
6	AMT	30	We attempted to have a "common" round for all participants in an effort to identify "personas". We also removed the calibration phase because of the bias that it would induce in personas.	Identification of "personas" or "groups of similar-minded people" in the data shows promise to have a clearer understanding of the tacton perception landscape.
7	AMT	10	The strategy that participants used was unclear to us, so we settled on a post-experiment questionnaire that asked workers about their strategy going into the experiment, and whether or not it changed during the course of the experiment.	Most users mapped their perception of tactons through rhythmic patterns that they perceived.
8	AMT	32	In order to avoid biasing participants with the first few iterations of pairs of tactons, we switch to querying the users with multiple tactons at a time and asked them to produce groupings (clusters) by themselves.	Asking the workers to work with more tactons and then reverting back to pairwise similarity ratings is more efficient and induces less fatigue in the workers.
9	AMT	35	We added multiple safeguards for submitting a rating, because some participants were suspected to rush through the experiment, highlighting the need for a disincentive to cheat is necessary in uncontrolled environments such as AMT; 12 ratings per round; 10 rounds.	12 ratings per round overwhelmed the participants as the agreement between participants decreased.
10	AMT	30	We reduced the cognitive load of the task to fewer rounds and less ratings per round.	6 ratings per round was a good number for participants to hold the tactons in memory without complexifying the task.
11	AMT	23	In order to improve the speed at which the probabilistic model converges, we implemented an active learning method that efficiently samples pairs from the tacton space given low information gain regions – simulations show the relevance of this method.	Active learning was successfully tested and shown to reduce overall uncertainty (as measured through the probabilistic variance) of tacton similarity.
12	AMT	310	We combine all of the previous pilots' learning experiences into one final experiment, the results of which are presented in this work.	

Table 4.1 Iterative process behind the final experiment. We ensured that no participant did the experiment more than once across all iterations.

4.1.4 Description of the Experiments

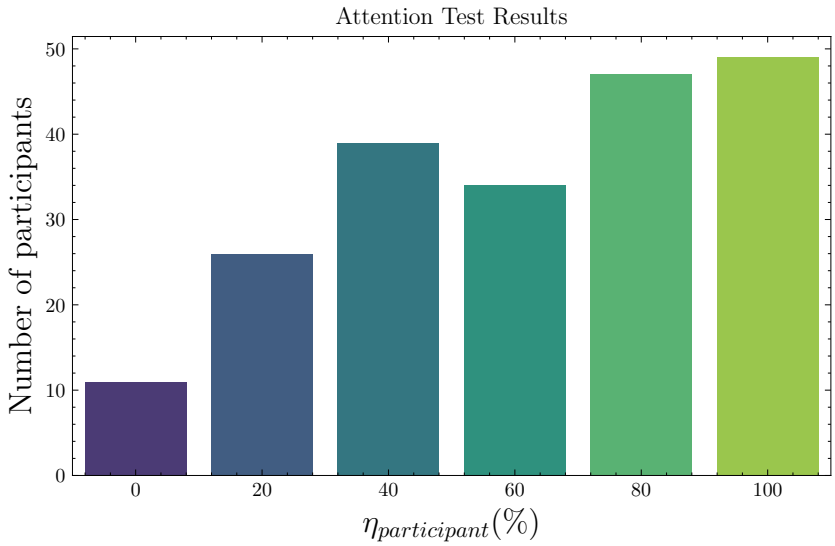


Fig. 4.1 Performance of participants on the "gold standard" attention test.

The final experiment was composed of $R = 8$ rounds of $c = 6$ tactons each. Five of those rounds were dedicated to the "global" experiment, while three were dedicated to the persona experiment. At each round, participants were asked to produce *at least* two groups according to how similar they perceived the VT stimuli. Although a group would technically comprise two or more tactons, we allowed participants to produce singletons, so long as the rule above was satisfied. In order to deter participants who enter random answers, we also required participants to have played each tacton at least three times before submitting a grouping.

Our pilot studies (see Section 4.1.3) did not indicate that the participants interpreted the word "similar" in the same manner, some would look for tacton *equality* rather than *closeness*. For this reason, prior to starting the experiment, participants were provided with the following explanation:

> We are NOT asking you to tell us whether the vibrations are the SAME; instead we are looking for your gut feeling as to whether they feel similar to you. A synonym to

> "similar" would be "close", "comparable", "near", "alike", or "resembling without necessarily being identical."

The complete instruction form that was presented to the participants on AMT is attached in Appendix A.

Two tactons were identical in each round of the global experiment. This constituted our "gold standard" attention test: participants passed the test if they grouped those two tactons in the same cluster. We used this score for compensation and for evaluating the performance of each annotator, η^r, $r \in R$ (see Section 3.5.1). Note that we can calculate a performance indicator on a per-round basis, η^r, or on an aggregated version per-participant, $\eta_{participant}$. This latter per-participant score corresponds to the accuracy of the attention tests across all five rounds of the global experiment:

$$\eta_{participant} = \frac{1}{R} \sum_{r=1}^{R} \eta^r \tag{4.1}$$

We used the η^r in compensation and for active sampling (see Section 3.5.1) and used $\eta_{participant}$ for assessing if the data from a participant was reliable enough to be included in our pool of valid data (see Section 4.2.1).

Upon completion of the experiment, participants were asked to complete a small demographic information questionnaire comprising a dozen questions (see Appendix B), along with a unique token to enter on AMT to validate the HIT. The completion tokens were stored in a database and HITs were approved only if the corresponding completion tokens were genuine. This procedure is in accordance with common anti-cheating strategies on AMT, because it avoids compensating workers who issue made-up confirmation codes, who resubmit previously generated codes, or who submit empty tasks and claim they did not get a code upon task completion [82].

There were 210 participants in total; we retained 129 because of attention test screening. In total, there were 1032 cluster ratings across 6192 binary tactons.

4.2 Global Similarity Assessment

This section presents the results of the global experiment, where we assess the perceptual similarity of actively sampled randomly generated tactons. First, we present general re-

sults on the attention tests. Then, we evaluate the active sampling procedure and show its relevance. Further, we analyze the data from the similarity experiments, and we extract ID characteristics from them. Finally, we perform community detection on the network of tacton ratings to obtain similarity clusters.

4.2.1 Attention Tests

In Figure 4.1, we present the global performance of all participants on the "gold standard", that is, if they correctly identified the two identical stimuli at each round. Participants who scored 80% or more on these questions were awarded a bonus, and we did not make use of the data in further experiments for participants who scored below 50%.

4.2.2 Active Sampling

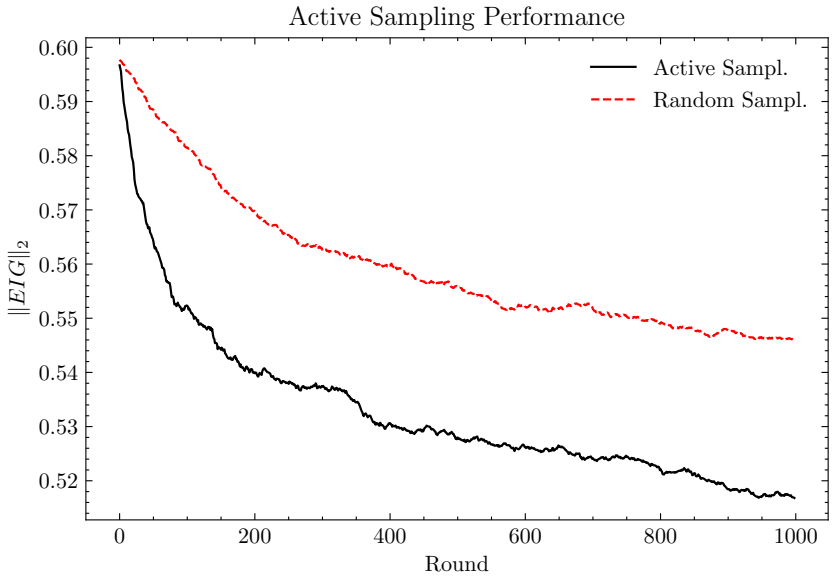


Fig. 4.2 Performance of active sampling vs. a baseline that randomly selects the same number of pairs at each round. A lower EIG signifies greater certainty in the groupings.

We evaluated the performance of the active sampling method presented in Section 3.5.3 by comparing it against a baseline of random sampling. The results are shown in Fig-

ure 4.2. Both curves were calculated on a constant amount of 30 tacton pairs over 1000 rounds. The curves depict the Euclidean norm of the EIG on all possible pairs of tactons: a lower norm indicates lower uncertainty with respect to the groupings, and is thus better. As such, we can see, especially in the first 100 rounds, high improvement in the groupings' information content, indicating that the active sampling strategy selects samples more efficiently than random. In fact, we can quantify the improvement that our active sampling strategy has in terms of time budget: it takes, on average, around 6.5 times more rounds to achieve a comparable EIG norm for the random sampling as compared to our option. Note that while in this plot the curves seem to saturate past the 500th round; in true experimental conditions, the number of tactons is not constant and the information gain tends to increase naturally, constantly yielding high EIG opportunities in areas of the tacton space.

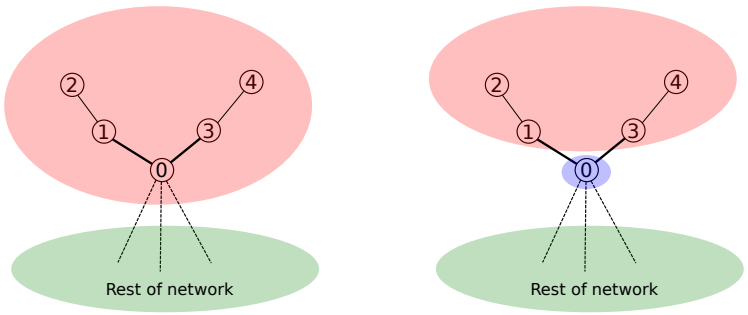


Fig. 4.3 Well-connectedness in the communities found by the Leiden algorithm (left). The right cluster becomes disconnected because node 0 was sent to another community, thereby dismantling the previously formed red community. Inspired by Traag *et al.* [1].

4.2.3 Perceptual Similarity Aggregation

4.2.3.1 Community Detection

Analogous to detecting "cliques" of people in social networks, we use the tacton network (see Figures 4.4 and 4.5) to detect communities of tactons that share common characteristics and that are perceived similarly on a global scale. We make use of the Leiden algorithm, an extension of the widely used Louvain algorithm for community detection

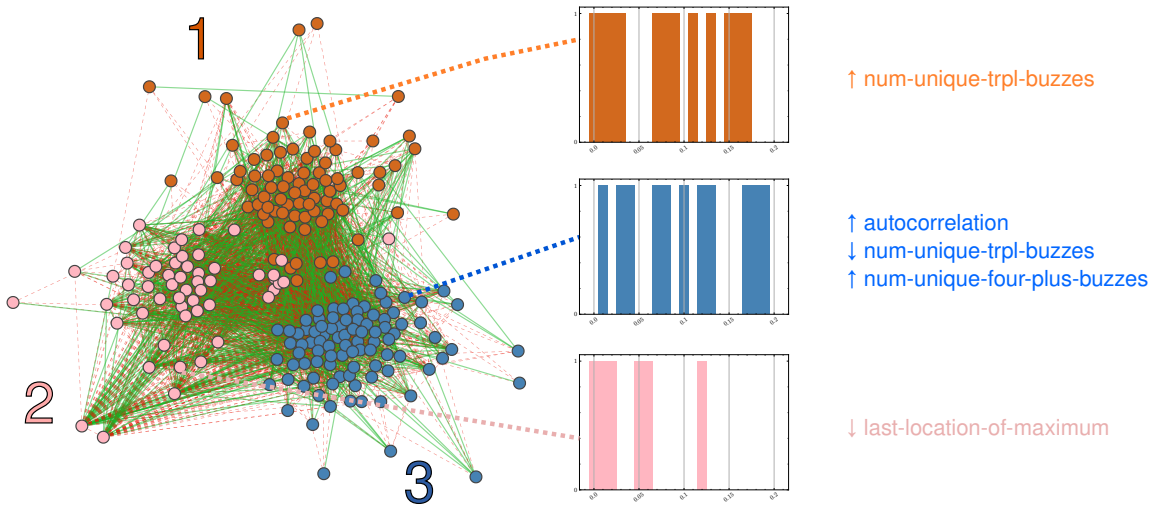


Fig. 4.4 Community detection using the Leiden algorithm in the similarity network. Edges representing similarity are depicted in green, and those representing dissimilarity edges are depicted in red.

in networks [1]. The Leiden algorithm is robust to different seeds, supports weighted graphs, and detects structure in a way that guarantees well-connected communities. Well-connected communities are communities that do not contain distinct disconnected subgraphs (see Figure 4.3). As opposed to the Louvain algorithm, the Leiden algorithm ensures that all tactons in one community have some connected path with the others. This is vital because the similarity links represent the data collected from the participants; ignoring well-connectedness would invalidate the results for the purposes of our study. Intuitively, while this procedure has a tendency to form a lower amount of bigger clusters, it ensures that there are no biases in the community information.

We show the result from the Leiden algorithm in Figure 4.4, where we represent each community by a distinct node color. The procedure not only detects similarity clusters, but also provides a hierarchy of similarity for each cluster, which we depict in the dendrogram in Figure 4.5. We notice that the orange (#1) and pink (#2) communities are closer to each other than to the blue community.

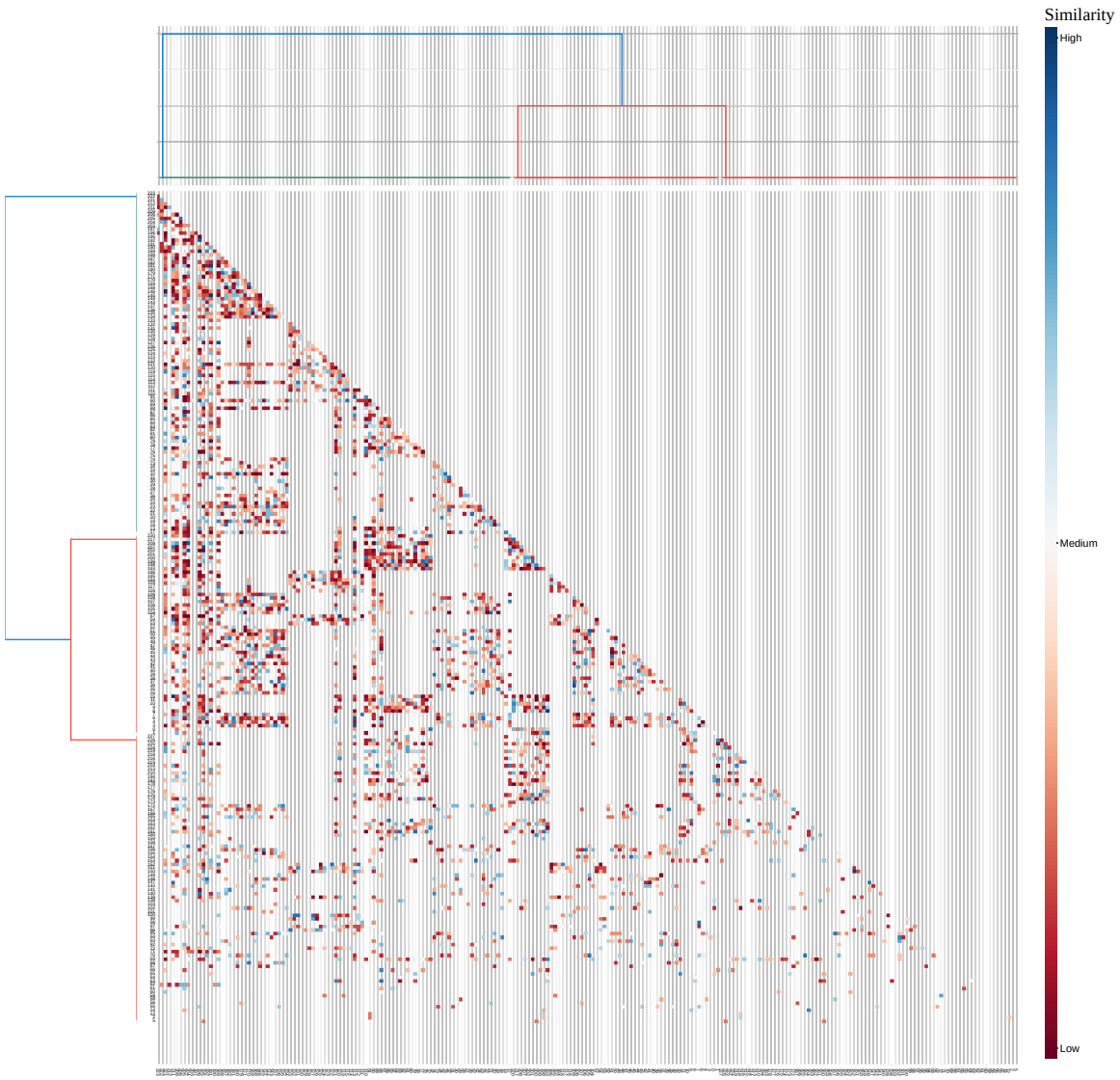


Fig. 4.5 Pairwise similarity ratings across tactons and participants.

4.2.3.2 Community Characterization

We characterize the communities by analyzing the feature distribution of tactons among each one, and show the results in Figure 4.6. A more detailed description of all the tacton

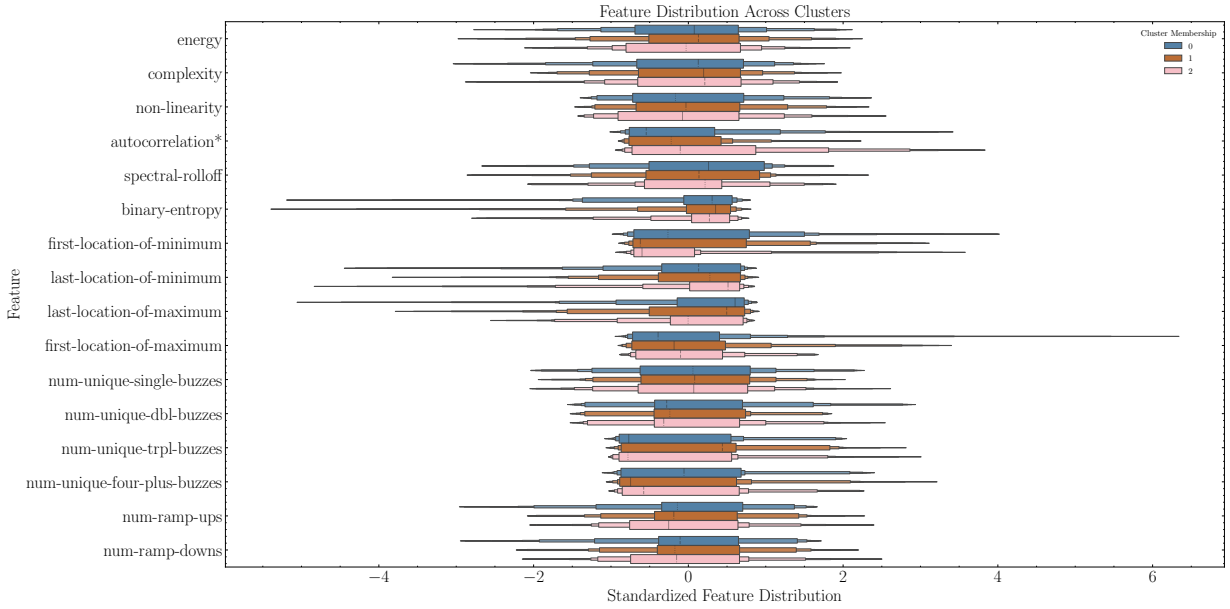


Fig. 4.6 Feature distribution across all clusters. The distributions were normalized so as to compare them on a similar scale. The asterix represents statistical significance across all groups for that feature according to a Kruskal-Wallis H-test with 95% confidence.

features used for analysis is available in Appendix D. The distributions were standardized so as to make a fair and clear comparison between the different scales of each feature. We notice that the blue and orange communities exhibit strong preference for distinct features, whereas the pink community does not seem to be characterized by any of the features used in this study. We performed a Kruskal-Wallis H-test to determine which features could significantly tell apart all three communities with 95% confidence. The only feature that is statistically significant is the autocorrelation of the signal.

4.2.3.3 General Representation

Figure 4.5 presents a heatmap of the logistic regression on the similarity ratings. A higher value on the heatmap depicts stronger similarity. Tactons are represented in community order, on the rows and columns. Each square in the matrix corresponds to one similarity logistic regression on the pairwise evaluations for those two tactons. Alongside the heatmap, we present a dendrogram which portrays the similarity groups that were

extracted using the Leiden algorithm [1] on the undirected graph represented by the adjacency matrix. For details on this procedure, see Section 4.2.3.1. The dendrogram also serves to depict the resemblance between the different clusters obtained.

Observing closer, one might notice the presence of big "holes" in the heatmap. These holes represent tacton pairs that were never queried for comparison. This is the result of the active sampling process, which learns the inter-dependencies between the tacton similarities as the experiment progresses, and avoids comparisons that are redundant or wasteful. The tactons that were never compared were thus "judged" by the probabilistic model to have a high probability of being inferred from other relevant ratings.

A second observation is that as the number of tactons grew, the number of ratings diminished greatly. This is an artifact of MST sampling (see Section 3.5.3). As the number of tactons grows, the MST only allows to sample each tacton once per batch. In the case of a high number of tactons, there is a greater chance that a tacton pair would require more than a single rating to infer "true" similarity values; this introduces irregularities in the global modeling where the more tactons there are, the less likely the active sampling algorithm will sample them consecutively.

4.2.3.4 Characterizing Individual Differences

While the probabilistic model gives the average rating of the perceptual similarity of pairs of tactons, our active learning framework can also give insight into the information gain from querying that pair of tactons. This can be used as a proxy for the *confidence* in the similarity ratings. Not only does this confidence score provide a global picture of the ratings, but it can also be decomposed to show what features account for the most variability in perception across individuals.

To do so, we first calculated the difference in feature values for each pair of tactons. We standardized those feature values differences and all EIGs across all tactons to be able to compare the EIGs. In Figure 4.7, we plotted the histogram of the feature differences with respect to the EIG. A lower EIG indicates a higher certainty that two tactons are similar. Darker colors indicate higher count of tacton ratings. One would expect that a high difference in feature values would lead to a lower uncertainty, i.e., a high degree of discrimination between the pair of tactons. For instance, if two tactons that have energy

Fig. 4.7 Characterizing individual differences by decomposing the ratings for each feature, and looking at the uncertainty (measured by the EIG) of the probabilistic model with respect to the difference in feature values in the tacton pairs.

values of 0.2 and 0.8 (out of a possible range of 0 to 1, thus a high feature difference) are found to have a low EIG, we can confidently surmise that energy differences do not explain individual differences in tacton perception.

Features that exhibit low EIG when their respective standardized features differ by more than two standard deviations in relative terms are more prone to characterizing individual differences. As such, we are looking for high density in the rightmost part of each graph. This would indicate that high differences in the feature's values for each tacton in the pair have led to high certainty in the ratings. Conversely, we want to avoid features whose density is high in the left side.

We see that features such as autocorrelation and first location of minimum, as well as most features that are discrete in nature (ramp-ups, tempo-related features), tend to be good for characterizing individual differences.

4.2.4 Discussion

4.2.4.1 Running studies on AMT

All in all, the results of our experiment strongly support the relevance of large-scale data collection platforms such as AMT in the evaluation of haptic perceptual tools. We however caution novices from repeating the same errors that we made (see Section 4.1.3): it took several iterations to get to a point where the data were reliable and usable. Running haptic studies is already complex; running them in a remote, unsupervised environment required multiple iterations, robust monitoring tools and data analysis. All things considered, the lessons learned are that scientists should (1) make use of the bottom-up approach described in Section 3.1.1 whenever possible, (2) avoid having a calibration phase, especially when gathering affective ratings (Section 4.1), (3) in evaluation studies, keep the number of items for comparison low (< 8) to reduce the cognitive load of the task and prevent workers from focusing on different tasks (Section 3.2), and (4) use the active sampling scenario described in Section 3.5.2 to reduce the person-time budget required for the task.

One crucial piece of advice for scientists is to avoid gathering too much information from a single participant; one should prefer gathering *less* data across *more* participants. Anecdotally, we noticed that the quality of ratings was reduced significantly beyond a dozen minutes of experiment length: we attribute this mainly to sensory fatigue and loss of concentration in the participants.

4.2.4.2 Participant performance

Our haptic-related HITs are done on smartphones, outside of the traditional AMT ecosystem. Moving away from this ecosystem tends to attract dishonest, malicious or simply distracted participants, because they perceive these studies as less robust to cheaters. Visualizing the performance of the participants on the attention tasks can better inform researchers when designing studies to run on AMT. To get a better idea of this distribution, we plotted the attention test accuracy in Figure 4.1.

Through our pilot studies, we estimated that even someone with experience with haptics may make mistakes about the similarity of the same two tactons about one fifth of the time. Given the objective difficulty of the task of sensing differences in tactons that may be very subtle for users who have no experience with haptics, we decided to discard the data from participants whose attention tests results were below 50%, thereby discarding around one third of the annotators. On one hand, the fact that the amount of annotators who obtained above 50% is around two-thirds demonstrates that running remote haptic studies on AMT is a promising direction. On the other hand, the fact that one-third of the data were discarded highlights the importance of implementing attention tests.

4.2.4.3 Probabilistic modeling

Contrary to previous literature that relied on MDS to offer an understanding of the similarity space, our probabilistic model attempts to model both the similarity between pairs of tactons as well as the *uncertainty* surrounding those ratings; we then leveraged this uncertainty to efficiently sample the tacton space. Figure 4.2 shows that this approach greatly reduced uncertainty as compared to a random sampling baseline for a constant number of tactons. The reduction in uncertainty is key to analyze, because haptic evaluation is typically very noisy. Unlike more traditional approaches such as MDS, it also allows us to naturally integrate an analysis of the individual differences in the modeling (see Section 4.2.3.4). In that sense, our model can be thought of as a hybrid between a Bayesian and a frequentist appoach to modeling similarity ratings: Bayesian because we have prior beliefs about the similarity pairs that we update sequentially and the purpose

of the data collection process is to model the distribution; frequentist because the logit model of regression does not rely on any prior information.

However, our perceptual similarity model does have its limitations. For instance, our similarity model relies greatly on the "triangle inequality" [97] for global similarity assessment. An example of the triangle inequality is that while red is similar to purple, and purple is similar to blue, in contradiction, red is not similar to blue. This goes to show that in reality, the internal mechanisms by which we distinguish two similar stimuli may not be based on the same grounds. A second limitation is that similarity may not always be symmetric. For instance, Tversky [98] reported that most people believe the similarity of North Korea to China to be greater than the similarity of China to North Korea. In summary, while the probabilistic model has advantages by modeling both the percepts and the decision-making processes as probabilistic, it fails to consider the *source* of the uncertainty generated by the ratings.

4.2.4.4 Global similarity assessment

All in all, the global experiment presents the first successful attempt at modeling perceptual similarity in a "bottom-up" fashion, providing a new perspective on tacton evaluation. While previous haptic evaluation experiments for similarity typically involved 100 or less participants [64, 65, 99, 67], the high number of participants with diverse backgrounds that took part in this study along with the outside-the-lab setting make the results more representative of the population in general. Mapping global perceptual similarity can be further used in rating propagation in applications like tacton generation, or in smartphone applications where one haptician might want users to users to discern coarse changes in the VT patterns.

The main finding is that the approximately 200 binary tactons that were evaluated could be grouped into three main perceptual similarity clusters, or "communities." From examples of tactons associated with each family, and their most salient features, we anecdotally labelled the pink, orange and blue communities, "short", "coarse" and "jittery" respectively. As seen in Fig. 4.4, the pink community tends to have short tactons, with a low value of last-location-of-maximum ("short"); the orange cluster has a high number of

buzzes of duration longer than a single period ("coarse"); the blue community has high autocorrelation and a high number of buzzes, indicating fast tempo ("jittery").

While this number may seem low, it is in-line with Seifi *et al.* [49] who found that participants typically preferred higher-level changes to vibrations when asked to design them. This may very well be the case here: the amount of perceptual similarity groups that can be distinguished in the whole tacton population at a high-level could be quite low.

Once the communities were determined, we analyzed the features of the distinct tactons that were in those communities in order to gain insight into what the population as a whole used to judge (dis)similarity. Naturally, due to the fact that we were limited to binary tactons, the tacton features are less descriptive than those of amplitude-modulated tactons. In this case, most, if not all, features depict some characteristic of the rhythm or the tempo of the VT pattern. The only feature that exhibited statistical significance across all groups is the autocorrelation. The low amount of statistically significant features was expected, as with only three communities and a low number of clusters the features are disparate inside each cluster. Given our data, our advice to designers is that if you want simple, binary tactons to be perceived differently, autocorrelation seems to be the main parameter to tune (see Figure 4.6). Based on our observations and although it was found to be marginally non-statistically significant, the already widely used approach of modifying the number of distinct buzzes (e.g., the triple-buzzes, as seen in Fig. 4.6) in a tacton also seems to be a promising factor to consider to maximize tacton differentiability.

4.2.4.5 Characterizing individual differences

We implemented a methodology to detect the features that were most useful in finding individual differences from a global consensus on the ratings. We saw that the concepts of information gain and certainty/uncertainty were closely linked, and that we could use the latter as a proxy for evaluating the former. Results in Figure 4.7 have shown that tacton features such as autocorrelation and tempo-related features (i.e., number of ramp ups or down, number of unique buzzes, etc.) were more discernible of the tactons than more complex features that had to do with frequency information or signal processing. In a deeper analysis that would analyze the distinctions between the perception across

participants, one would therefore want to focus on these features to discern groups of perceptual similarity.

These results corroborate the comments received gathered through the early pilot experiments where participants gave the feedback that they were mostly focused on the number of individual "buzzes" in the tactons more than anything else.

One shortcoming is that the number of features for binary tactons is limited. In the case of more complex tactons, we anticipate that the feature diversity would be greater. Gathering data on the perceptual similarity of amplitude-modulated tactons would likely lead to a smoother characterization of the tacton space.

4.3 Personas

As described in Section 4.1.4, we used the same six randomly generated tactons to evaluate the persona experiment. We queried the tactons three times per participant, thus obtaining three groupings on the same group of tactons. We combined these annotations, and used them to map inter-user haptic similarity perception across all users.

The aggregation procedure is as follows: we first flattened the pairwise ratings inside the clusters, such that we obtained, given 6 tactons in a round, $\frac{6\times(6-1)}{2}$ pairwise comparisons; we then reduced the dimensionality of each set using UMAP [100]; finally, we used HDBSCAN [101, 102] to cluster the lower-dimension embeddings into *personas*. An example of the flattening process for the 15 pairwise comparisons obtained from the groupings for a single round can be seen in Figure 4.8. The final aggregation is the sum of the flattened strings of the three rounds.

Contrarily to the more traditional principal component analysis (PCA), UMAP preserves the global structure in the data – this is useful to later perform meaningful clustering on the low-dimensional embeddings. HDBSCAN is a clustering algorithm that can find clusters without the need for *a priori* specifying the number of clusters. Not only does it leave items unclustered, it can elect *not* to cluster an item in any group and provide a confidence score for each item. The hyperparameters used in both of these algorithms are available in Appendix E.1, and the linkage tree for splitting the UMAP low-dimensional embeddings created by HDBSCAN can be seen in Figure 4.9.

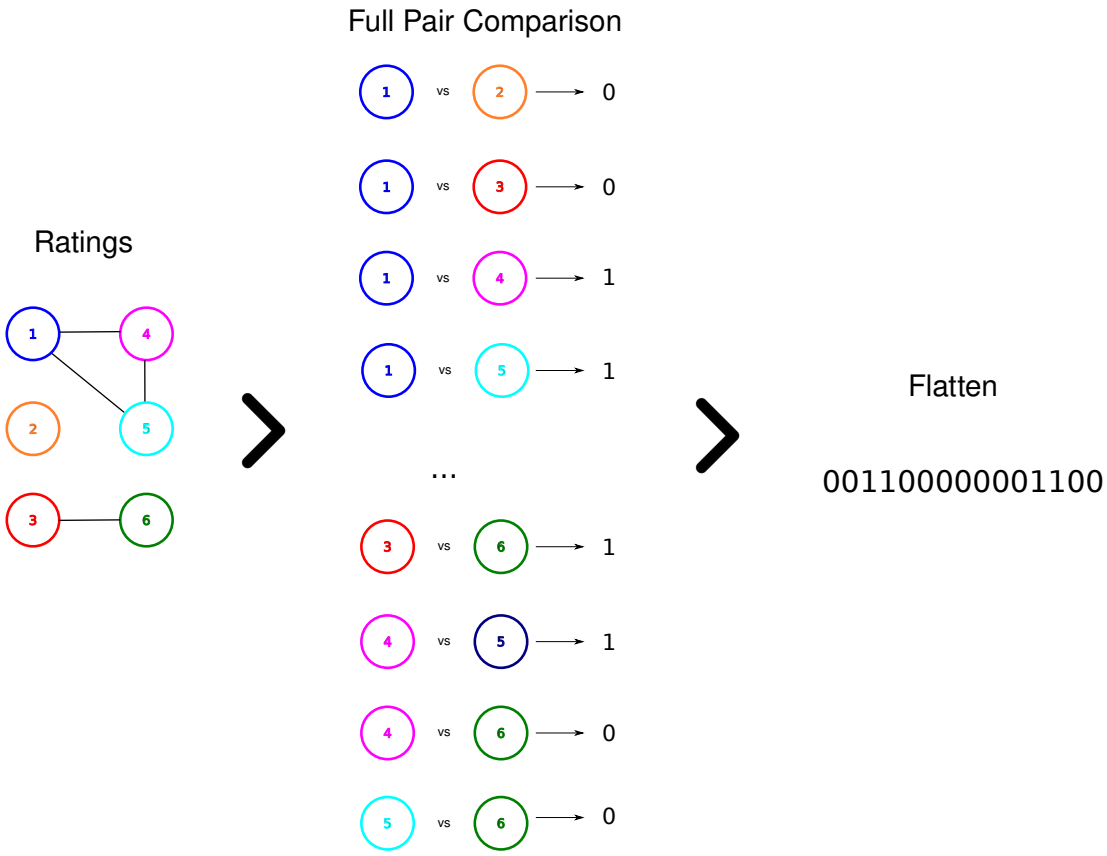


Fig. 4.8 Flattening the persona ratings.

Internally, HDBSCAN splits the various data points into clusters in a hierarchical fashion. The linkage tree is useful to visualize the process by which the individual participants were split. The cluster count is the highest level grouping that can be formed from branches of the linkage tree at a given distance level. The depth of the linkage tree plot depicts the distance of the radii of the cluster centroids in the space of participants. This points to a trade-off: a high distance will indicate a coarser view of the space, while a lower distance will produce a tighter one. This is specifically depicted on the three plots on the right, where we show that the same participants were respectively split into three groups (clusters) at distance 6.1, eight groups at distance 4.1, and seventeen groups at distance 2.05.

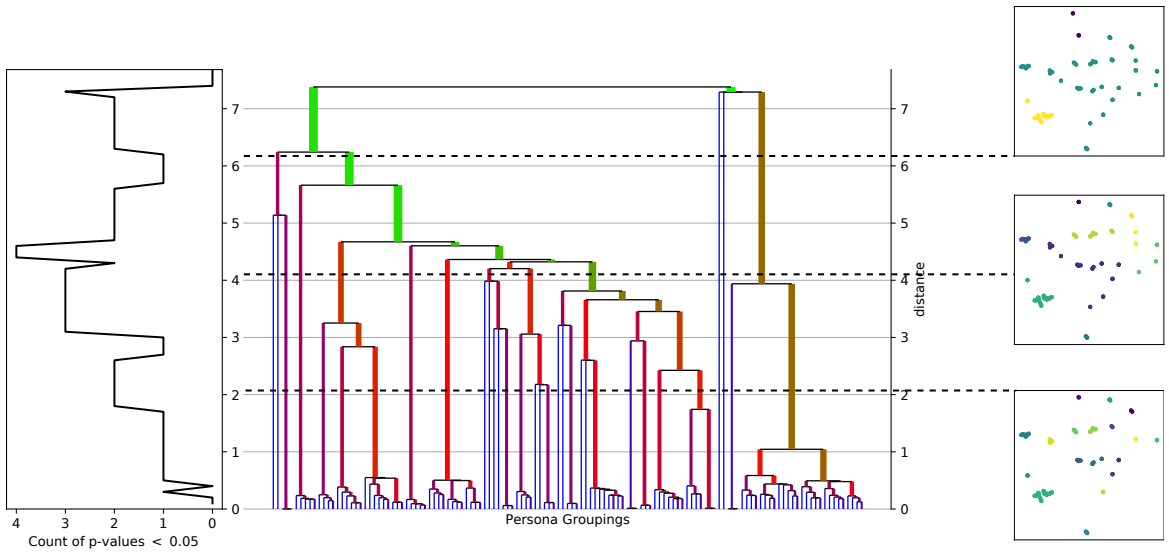


Fig. 4.9 HDBSCAN clustering linkage tree for grouping the personas. A smaller distance indicates a finer clustering. We settled on a splitting distance of 4.4 because it exhibits the highest agreement between the personas and the tacton features.

We found the most descriptive persona grouping with respect to the tacton perception as follows. At each round of the persona experiment, we counted the features of the tactons that were found similar. At each distance, we performed logistic regression on the cluster labels using this aggregated count. We summed the count of statistically significant features (confidence level 95%) for each regressive model (one per distance) and plotted it against the distances (see Figure 4.9 on the leftmost plot). The intuition is that a higher count indicates that more tacton features explained the personas, thus pointing to a higher relevance of the clusters (persona groupings) at that distance level.

Distance level 4.4 was found to be most explanatory of the features, and yielded six different persona groups. At that level, HDBSCAN excluded 4 participants from being in any group. The detailed flattened ratings for each persona group can be found in Appendix C.

4.3.1 Feature Saliency across Personas

For each persona group, we evaluated the feature distribution of the *global* tactons that that group found similar. We standardized each distribution by subtracting its mean and dividing by its standard deviation across all groups in order to get a comparable scale for each group and feature. We plotted the mean of each distribution in Figure 4.10.

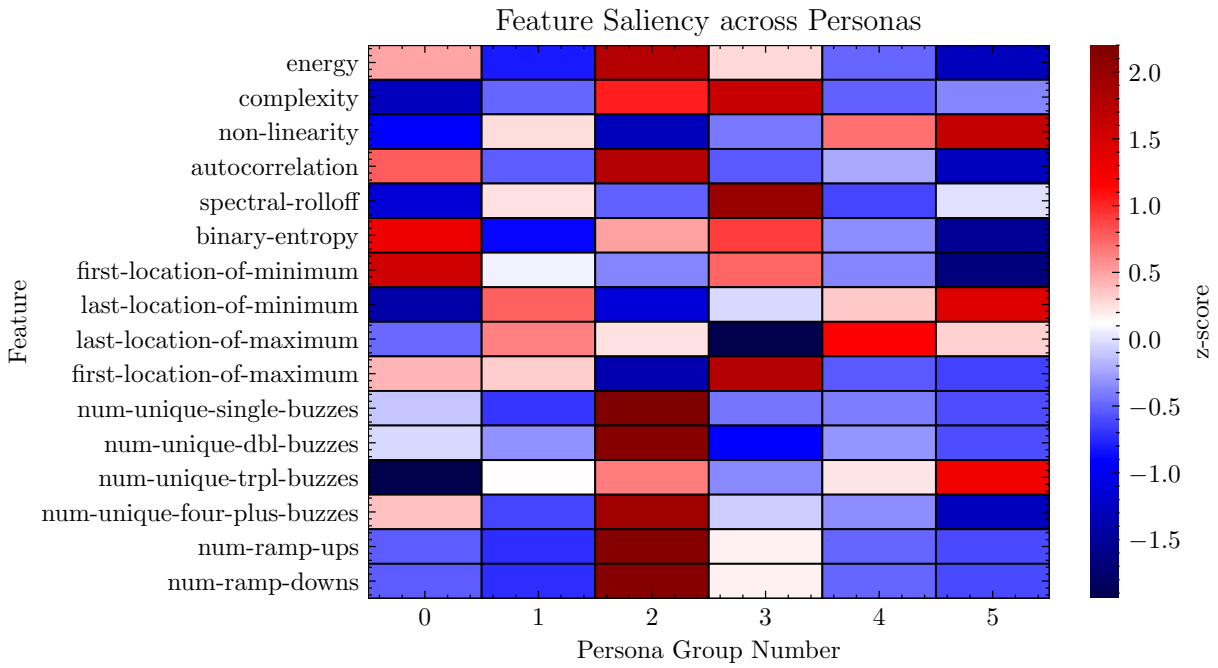


Fig. 4.10 Feature saliency for each persona group and tacton feature.

We notice that the participants in persona groups #2 and #5 exhibit inversely correlated feature distributions for multiple features, and groups #4 and #5 are sensitive to the same features as a whole. Groups #0, #1 and #3 do not appear to have correlations with the other groups.

4.3.2 Demographic Information and Personas

We collected demographic information about each participant after completion of the experiment (see Appendix B for details of the information collected). We one-hot encoded the demographic information and used Lasso regression to predict the persona groups.

We then used Student's t-test to find the demographic information that was statistically significant predictors of the persona groups. The results are presented in Figure 4.11.

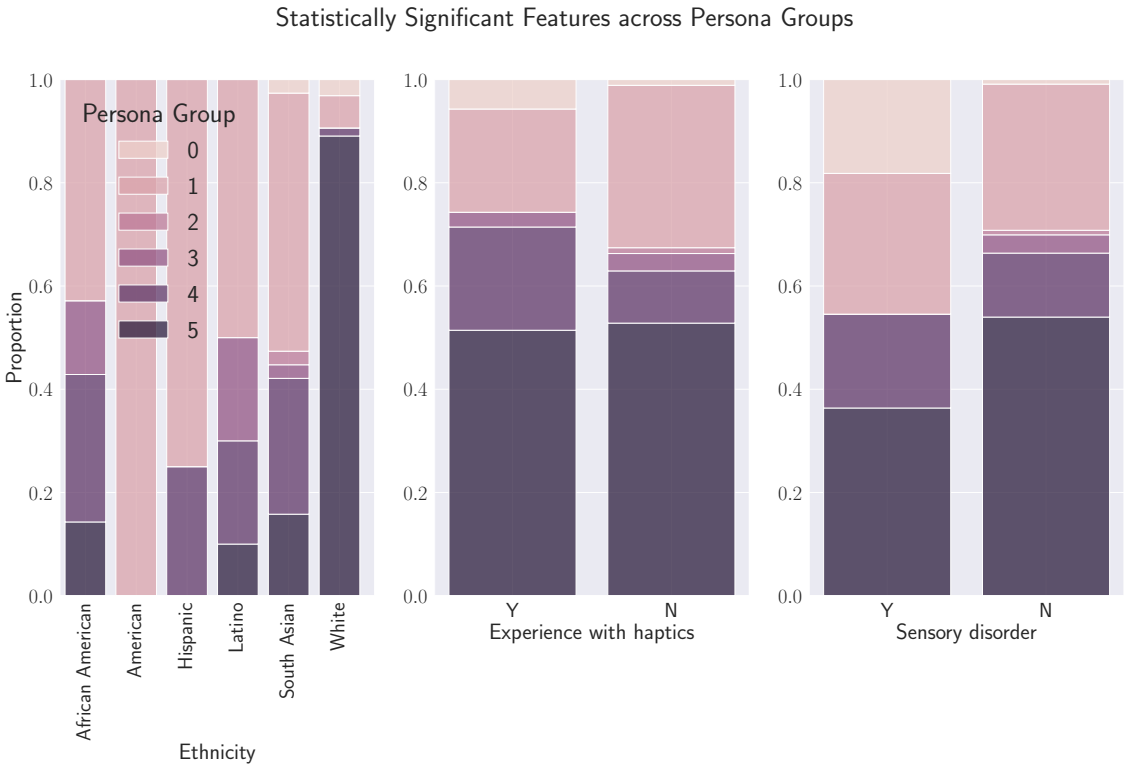


Fig. 4.11 Demographic information distribution for each persona group.

We found that most participants with white ethnic background were classified into the same persona group (#5), while all Native Americans were classified into group #1. Also, nearly all south Asians were classified in group #2. Hispanics were mostly in group #1, along with Latinos. The persona group distribution is similar among African Americans, Hispanic and Latinos as they seem to share many similarities with respect to how they perceive tactile stimuli. The plots for "experience with haptics" and "presence of sensory disorder" appear to be mostly due to the unanimous presence of group #2 in the negative responses; this does not appear to be conclusive.

4.3.3 Discussion

The results of the present section are, to our knowledge, the first evidence of haptic evaluation tools that explicitly consider and analyze the inter-user differences in the perception of haptic stimuli through "personas." Figure 4.9 shows that it is possible to "slice" the linkage tree of a clustering to extract coarser or finer representations of the population. Our haptically motivated method to find the optimal slice has shown promise in differentiating the personas on a perceptual level (see Figure 4.10), using the feature characteristics of the global tactons that were presented to each persona group's participants. We found that specific persona groups exhibited strong preference towards certain features of tactons for perceptual similarity. This evidence strongly supports the hypothesis that tactile perception is not uniform across the whole population.

Further supporting this claim are the results of Figure 4.11. Obvious elements are that a great majority of Asian participants were found to be in the same persona group; the same goes for White participants, and participants with Hispanic/Latino origins. We caution in interpreting these results as proof that haptics is perceived differently across ethnicities: this could simply be an artifact of the response process on AMT. Further work that dives deeper into cultural differences in haptics would be needed to confirm or refute this.

As a side node, we made sure to check the correlation between responses and the cellphone brands that participants used across the persona groups and could not find any meaningful link between the two.

Nevertheless, the results in this section reveal promising directions for future research with regards to looking at differences in haptic perception across the population in general. The next section will discuss how we can leverage these personas to gain a greater understanding of the VT haptic perceptual similarity landscape as a proxy for stimuli personalization.

4.4 Predicting Similarity Ratings

4.4.1 Graph Representation Learning

The idea behind graph representation learning is to learn a faithful and exploitable representation of the graph structure. We can achieve this by finding a higher-dimensional representation of each tacton (node) in the graph that combines information from (1) its temporal characteristics, (2) its one-hop neighbors and their associated similarity weight, and (3) its n-hop proximity that enforces context sharing (i.e., nodes that share common neighbors but are not directly connected). We refer to this higher-dimensional representation as the node *embedding*, $\psi(T)$.

Embeddings are not directly interpretable, but the vector similarity between the embeddings can give insight into how close two tactons are in the space, and we can perform supervised or unsupervised learning on these embeddings in a generalizable manner. To learn these embeddings, we minimize the triplet loss, which maximizes the distance between an "anchor" tacton A to a "negative" tacton N, while minimizing the distance between the anchor and a "positive" tacton P. The process is depicted in Figure 4.12, and can be described mathematically as:

$$\mathcal{L}(A,P,N) = \frac{1}{B}\sum_{i=1}^{B} max(\|\psi(A_i) - \psi(P_i)\|_2^2 - \|\psi(A_i) - \psi(N_i)\|_2^2 + \alpha, 0) \tag{4.2}$$

where α is the margin parameter, used to avoid convergence to trivial solutions, and B is the batch size. Notice that there is no gain when $\|\psi(A_i) - \psi(P_i)\|^2 < \|\psi(A_i) - \psi(N_i)\|^2 + \alpha$, so in practice we use hard-triplet mining: we only take into account hard or semi-hard triplets that yield a positive loss. Anchor swap [103] was used to improve the convergence of the algorithm. We repetitively sample triplets of tactons from the graph, run them through the GCN model, compute the triplet loss, and backpropagate the error to learn the embeddings $\psi(T)$. This is not the first instance of triplet loss being used in haptics: Priyadarshini *et al.* also used triplet loss to learn perceptual (dis)similarities in the context of haptic textures [104].

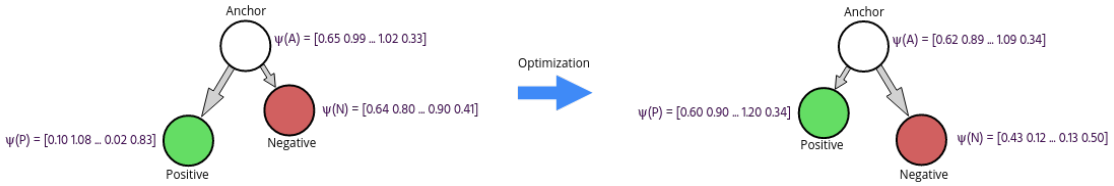


Fig. 4.12 The triplet loss maximizes the distance between the anchor embedding $\psi(A)$ and the negative embedding $\psi(N)$ whilst minimizing the distance between the anchor embedding and the positive embedding $\psi(P)$.

The following section will describe how we evaluated the embedding procedure learned by the GCN with triplet loss, and will show how the information captured by the embeddings has predictive power over tacton similarity.

4.4.2 Predicting Similarity on the Global Experiment Data

As depicted in Figure 4.12, we learned an embedding $\psi(T)$ for every tacton that summarized its relationship with other items in the graph as well as its physical characteristics. To validate our embedding in the tacton space, we use them as inputs for a weighted link prediction model, i.e., a model that can predict the degree of similarity between two tactons.

We implemented two simple regressors to evaluate the performance of our node representations: a linear regressor and a gradient boosting tree regressor, and trained both models. The parameters for the gradient boosting tree model are located in Appendix E.3.

All of the results presented in Table 4.2 are averaged over a five-fold cross-validation scheme and over five different seeds for the gradient boosting tree regressor.

4.4.3 Extending Similarity Prediction to Personas

Learning the global similarity with the graph approach is useful, but it would be even more useful if we could tailor the similarity groups to suit individual participants. Previously, in Section 4.3, we grouped participants into a number of personas that described their perceptual similarity. To evaluate these personas, we train the supervised learning edge weight prediction model for each one and average the results. The performance of

the weighted link prediction task then acts as a surrogate objective to evaluate the performance of our unsupervised embedding procedure (the node representation learning).

We aggregate the results of the inter-persona similarity prediction experiment in Table 4.2, and detail the R^2 for every persona group in Figure 4.13. Appendix E.2 details the triplet loss performance and the hyperparameters for tuning our graph learning models. All models were created using Pytorch Geometric [105] and optimized with Adam gradient-based optimizer [106].

Table 4.2: Average error on the testing set for the similarity prediction task. All results are averaged over five different seeds (for the gradient boosting regressor), and cross-validated in a five-fold scheme.

Experiment		R^2	RMSE	MAE
All Tactons	Lin. Reg.	0.081 ± 0.042	0.249 ± 0.010	0.246 ± 0.005
	Grad. Boost.	0.152 ± 0.012	0.129 ± 0.018	0.120 ± 0.020
Avg. by group	Lin. Reg.	0.132 ± 0.015	0.238 ± 0.044	0.221 ± 0.050
	Grad. Boost.	0.272 ± 0.109	0.091 ± 0.032	0.089 ± 0.038

The first observation is that, as is typically the case, gradient boosting performed better on the task than linear regression. The second is that, as indicated by the R^2 scores, the models were indeed better than predicting the mean of the dataset. The third is that the average by persona group improves upon the global experiment on average, indicating that the persona groups were meaningful in the prediction of perceptual VT similarity.

4.4.4 Discussion

4.4.4.1 Performance of learning from the similarity graph

Predicting human behavior has a tendency to be noisy and unreliable, because humans are complex machines. It is therefore quite encouraging to see that VT tacton perceptual similarity predictions were consistently better-than-random. While the range of R^2 obtained is quite low, we achieved approximately 10% root mean squared error on the test set. Although, this score would be deemed mediocre in some contexts, we believe

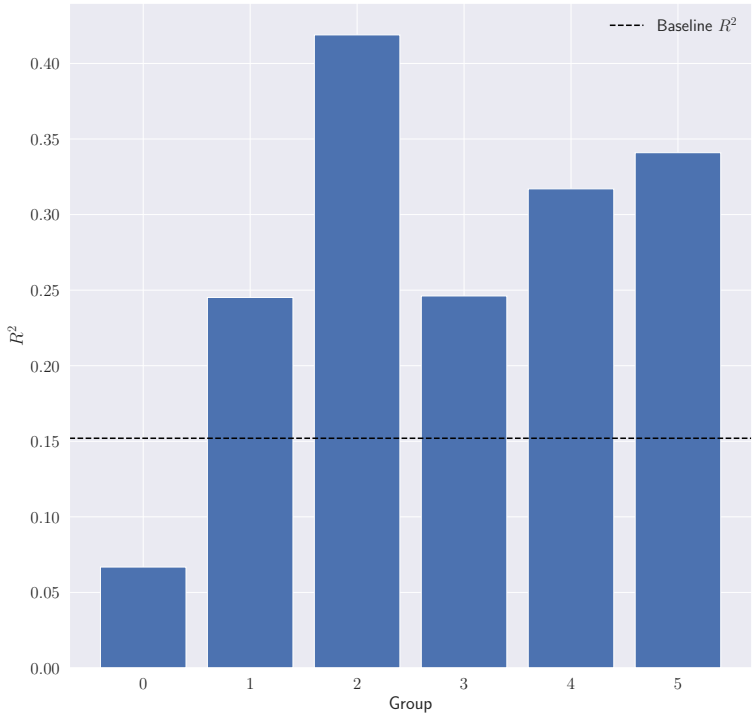


Fig. 4.13 Detailed R^2 results for all persona groups for the gradient boosting regressor.

that in haptics it could provide useful information to the designer, or even help further investigate tacton generation.

4.4.4.2 Performance on learning from the persona similarity graphs

As seen in Figure 4.13, on average, isolating the networks of personas to predict the similarity between pairs of tactons increased the prediction scores. Although we cannot directly know the cause of this increase in score (due to the lack of interpretability of the learned embeddings), a safe assumption is that the personas groups contain meaningful information about the perception of VT tactons. We hypothesize that the augmentation in predictive power from the personas models further corroborate the relevance of our methodology to finding perceptual similarity personas. The personas found were not simply an artifact of the collection process but contain information about the tacton

ratings that would not have been found had we looked only at aggregate data from all participants.

In addition, the results show the relevance of the graph approach for modeling tacton similarity: aggregating information from a tacton's neighbors helps predict similarity from that same tacton to others. In that regard, our approach to learning from the graph is conditional on the validity of the IIA axiom (see Section 3.2 and 4.2.4.3). In reality, violations of this axiom may account for generally low predictability of human behavior: biases in the response, shifts in selective attention and other factors may account for most of the error in the predictions; our machine learning model did not account for the uncertainty in the similarity ratings.

The authors also wish to further highlight the inductiveness of the graph similarity learning process: there is no need to re-train a model from scratch in order to predict the similarity between never-seen-before tactons and tactons in the graph. Improvements to the methodology include making use of graph-centrality features for the tactons, such as PageRank or the Adamic/Adar index.

Chapter 5

Conclusion

5.1 Summary

Past work in vibrotactile tacton evaluation has typically focused on designing tactons *a priori*, and then conducting user studies to evaluate them. This has allowed scientists to research meaningful parameters that influenced our perception of tactons, but has failed to provide a common ground to build robust tooling to support research in haptics, thereby forcing the community to "reinvent the wheel" on each new study (for a more detailed explanation, see Section 2.1.2).

In order to prevent this "reinvention of the wheel" on each study, we introduced in Section 3.1.1 a new paradigm to conduct haptic research, in which tactons are not designed upfront but rather randomly generated and iteratively presented to the user for comparison. We then extracted patterns and meaning from data mining techniques on the collected data. This had three main effects: first, all tactons are built from the same base; second, it mitigated the bias induced by the human design of tactons; third, it enabled efficient tacton research outside of typical laboratory conditions. This methodology required a larger number of samples, which we solved by outsourcing the task to workers on the Amazon Mechanical Turk platform (see Section 3.3). In order to further simplify the problem, we restricted the VT tactons to the binary case (Sections 3.1.2 and 3.1.3), and developed an active strategy to sample pairs of items that optimizes the exploration of

the space as compared to traditional within-subject studies (Section 3.5.3). This led to a drastic drop in the required person-time experimental budget by 6.5 times.

As described in Section 2.2, we chose to focus on the evaluation of tactile perceptual similarity, as it is a core element in achieving an understanding of the variables that motivate behavior and mediate affect [2], and as such is a stepping stone towards the validation of the bottom-up approach.

In order to test the methodology, we conducted a series of pilot studies to properly design AMT studies for haptics, the takeaways of which can be found in Section 4.2.4.1. In order to help scientists conduct remote haptic perceptual evaluation studies on AMT, we suggest (1) avoiding having a calibration phase, especially when gathering affective ratings, (2) keeping the number of items for comparison low (< 8) to reduce the cognitive load of the task and prevent workers from focusing on different tasks, and (3) using the active sampling scenario described in Section 3.5.2 to reduce the person-time budget required for the task. We anticipate that this data-intensive process will allow these researchers to gain greater insight into the underlying mechanisms by which we perceive and interpret haptic stimuli.

Through data mining, we mapped VT haptic perceptual similarity at scale in Section 4.2.3.1. We found that modeling haptic percepts as probability distributions (Section 3.5.2) helped characterize the uncertainty in ratings, and this led to greater insight into our perception of them (Section 4.2.3.4). Particularly, in our study, we found three distinct clusters, or communities, of binary tactons that were perceived similarly across all participants in our crowdsourced study. Each cluster is qualified by specific binary tacton parameters, such as the autocorrelation of the signal or the number of distinct buzzes (see Appendix D).

The uncertainty characterization led to a deeper analysis into "personas," or groups of users who share characteristics with regards to their perception of VT stimuli (Section 4.3). We provided evidence that the users in these groupings exhibited sensitivity to specific tacton characteristics (Section 4.3.1), and that some of them could be linked to demographic information such as cultural background ("ethnicity") or "prior experience with haptics."

We further found that graphs represented a natural way of modeling haptic similarity (Section 3.5.4). In a haptic graph, the nodes represent the tactons, and the edges the relationships between them. Graphs provide a scalable, compact way to represent perceptual interactions, and they enable us to easily perform machine learning on the haptic data. From our graph representation, we built a model that predicted the perceptual similarity between pairs of tactons from the large amount of data gathered through our experiments. The model first learned embeddings for all tactons in the graph using the triplet loss (Section 4.4.1), and then predicted similarity with either a linear regression or gradient boosted decision trees (Section 4.4.3). The results showed that the RMSE was consistently lower when using gradient boosting trees as opposed to linear regression (0.249 vs. 0.129), and the R^2 improved from 8% to 15%. While these are relatively low scores, this nevertheless showed promise in the relevance of the predictive model. Adding persona information to the inputs of the model increased the scores across the board (0.129 vs 0.091 for RMSE, 0.152 vs 0.272 for R^2), thereby highlighting the pertinence of the persona groupings.

5.1.1 Shortcomings

Given the constraints on the experiments, three main shortcomings can be identified. First, due to hardware limitations on smartphones at the time of conducting early experiments on this work, only binary tactons could be considered. This greatly limited the amount and the diversity of tacton features that could be analyzed. These days, the majority of smartphones contain haptic actuators that support amplitude modulation. Running the studies without the binary tacton constraint, thereby leading to an expansion of the tacton space, would allow a more complete experiment in the space of possible VT signals.

Second, results from AMT are often polluted with data from malicious workers. While screenings and attention tests can alleviate the problem, the screening process is not perfect, and some of the noise persists in the data.

Third, although our methodology could be applied to other domains of haptics, our experiments exclusively focused on vibrotactile stimuli, making the results only applicable to that particular subdomain.

5.1.2 Potential Impact and Future Work

The potential impact of this study in the haptic community is broad. Firstly, the flipped approach to bottom-up tacton evaluation along with the probabilistic model are useful to model perception in a more succinct, directed and generalizable way. The extraction of persona groups in our participant pool points to the potential for personalizing and tailoring haptics to suit the users' needs, which has been identified as essential to increase haptic adoption in consumer devices [15].

Secondly, an issue raised is that perceptual evaluation studies in haptics typically involved designing a set of tactons with little to no regard to prior work on how to *tie* the results to other work in the literature. Every study typically evaluates a subset of the tacton space, but the results are rarely aggregated into a coherent, organized set. Because this study employs a bottom-up approach by creating all tactons from the same *common ground*, extensions to it could therefore constitute the building blocks to "glue" multiple studies or libraries of tactons together.

Evaluating similarity in haptic tactons represents a base level from which to extract more complex perceptual characteristics or interpretations. For instance, knowing that two tactons are perceived similarly enables us to transfer our knowledge of one tacton to another (evaluating a single one of the two should tell us plenty about the second). This further reduces the need for "reinventing the wheel" at every perceptual study.

Additionally, our research suggests that there are specific groups of people in the population who exhibit perceptual similarity in haptics, and that these groups are more than an artifact of the experiment or of the data because they possess greater predictive power for perceptual similarity than the entire participant population. While we could not identify with certainty the *provenance* of these groups, we expect that future work should be directed at persona identification from demographic and cultural characteristics.

In the same line of thought, characterizing the personas would enable the classification of our perception of tactons on a finer level, i.e., digging deeper into what groups of people share haptic perceptual characteristics. One could predict how an individual will react to a stimulus without having ever collected data on them, but merely from how users with similar demographic characteristics have reacted to the stimulus.

Further lines of research should also be considered in the domain of affective computing. Our probabilistic model and graph theoretic framework, and by extension our active sampling scheme, could easily be extended to preference rather than similarity. The main difference is that similarity is symmetric while preference implies a direction in the judgement of the percept. In similarity, if tacton A is similar to tacton B, the opposite is assumed to be true as well. However, in preference evaluation, we attempt to infer a ranking from the candidates. For instance, the study could be re-run by asking participants to group tactons based on their perceived aggressiveness: the objective would then be to find the *most aggressive* tacton, which implies ranking the aggressiveness of the tactons. While a similarity graph is undirected, in this case, due to the fact that aggressiveness is not symmetric (tacton A is more aggressive than tacton B does not imply the opposite, in fact, it lowers the probability of the opposite), the "aggressiveness" graph becomes directed. This goes back to the Bradley-Terry model of preference for paired comparisons, as discussed in Section 3.5.2.

References

[1] V. A. Traag, L. Waltman, and N. J. van Eck, "From Louvain to Leiden: guaranteeing well-connected communities," *Scientific Reports*, vol. 9, p. 5233, Mar. 2019.

[2] F. Ashby and D. Ennis, "Similarity measures," *Scholarpedia*, vol. 2, no. 12, p. 4116, 2007.

[3] F. A. Geldard, "Some Neglected Possibilities of Communication," *Science*, vol. 131, pp. 1583–1588, May 1960.

[4] K. MacLean, "Designing with haptic feedback," in *Proceedings 2000 ICRA. Millennium Conference. IEEE International Conference on Robotics and Automation. Symposia Proceedings (Cat. No.00CH37065)*, vol. 1, pp. 783–788 vol.1, Apr. 2000. ISSN: 1050-4729.

[5] E. Kim and O. Schneider, "Defining Haptic Experience: Foundations for Understanding, Communicating, and Evaluating HX," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, (New York, NY, USA), pp. 1–13, Association for Computing Machinery, Apr. 2020.

[6] S. A. Brewster and L. M. Brown, "Non-visual Information Display Using Tactons," in *CHI '04 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '04, (New York, NY, USA), pp. 787–788, ACM, 2004. event-place: Vienna, Austria.

[7] A. Israr, S. Zhao, K. Schwalje, R. Klatzky, and J. Lehman, "Feel Effects: Enriching Storytelling with Haptic Feedback," *ACM Transactions on Applied Perception*, vol. 11, pp. 1–17, Sept. 2014.

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

[9] O. Schneider, S. Zhao, and A. Israr, "FeelCraft: User-Crafted Tactile Content," in *Haptic Interaction: Perception, Devices and Applications* (H. Kajimoto, H. Ando, and

K.-U. Kyung, eds.), Lecture Notes in Electrical Engineering, pp. 253–259, Tokyo: Springer Japan, 2015.

[10] D. Tam, K. E. MacLean, J. McGrenere, and K. J. Kuchenbecker, "The design and field observation of a haptic notification system for timing awareness during oral presentations," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '13*, (Paris, France), p. 1689, ACM Press, 2013.

[11] O. S. Schneider, H. Seifi, S. Kashani, M. Chun, and K. E. MacLean, "HapTurk: Crowdsourcing Affective Ratings of Vibrotactile Icons," in *Proceedings of the CHI Conference on Human Factors in Computing Systems*, (New York, NY, USA), pp. 3248–3260, ACM, 2016. event-place: San Jose, California, USA.

[12] P. Knees and M. Schedl, *Music Similarity and Retrieval: An Introduction to Audio- and Web-based Strategies*. The Information Retrieval Series, Berlin Heidelberg: Springer-Verlag, 2016.

[13] O. Schneider, K. MacLean, C. Swindells, and K. Booth, "Haptic experience design: What hapticians do and where they need help," *International Journal of Human-Computer Studies*, vol. 107, pp. 5–21, Nov. 2017.

[14] H. Seifi and K. E. MacLean, "Exploiting haptic facets: Users' sensemaking schemas as a path to design and personalization of experience," *International Journal of Human-Computer Studies*, vol. 107, pp. 38–61, Nov. 2017.

[15] H. Seifi, *Personalizing Haptics: From Individuals' Sense-Making Schemas to End-User Haptic Tools*. Springer, June 2019.

[16] J. B. F. v. Erp and M. M. A. Spapé, "Distilling the Underlying Dimensions of Tactile Melodies," in *Eurohaptics 2003 proceedings*, 2003.

[17] Brown, "A first investigation into the effectiveness of Tactons - IEEE Conference Publication," 2005.

[18] L. M. Brown and T. Kaaresoja, "Feel Who's Talking: Using Tactons for Mobile Phone Alerts," in *CHI '06 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '06, (New York, NY, USA), pp. 604–609, ACM, 2006. event-place: Montréal, Québec, Canada.

[19] J. Luk, J. Pasquero, S. Little, K. MacLean, V. Levesque, and V. Hayward, "A Role for Haptics in Mobile Interaction: Initial Design Using a Handheld Tactile Display Prototype," in *Proceedings of the SIGCHI Conference on Human Factors in Computing*

Systems, CHI '06, (New York, NY, USA), pp. 171–180, ACM, 2006. event-place: Montréal, Québec, Canada.

[20] 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*, MobileHCI '06, (New York, NY, USA), pp. 231–238, ACM, 2006. event-place: Helsinki, Finland.

[21] S. Brewster and A. Constantin, "Tactile Feedback for Ambient Awareness in Mobile Interactions," in *Proceedings of the 24th BCS Interaction Specialist Group Conference*, BCS '10, (Swinton, UK, UK), pp. 412–417, British Computer Society, 2010. event-place: Dundee, United Kingdom.

[22] M. Obrist, S. A. Seah, and S. Subramanian, "Talking about tactile experiences," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '13*, (Paris, France), p. 1659, ACM Press, 2013.

[23] J. Smith and K. MacLean, "Communicating emotion through a haptic link: Design space and methodology," *International Journal of Human-Computer Studies*, vol. 65, pp. 376–387, Apr. 2007.

[24] H. A. Osman, A. F. Pilon, and A. E. Saddik, "Mobile phone short message tacton notification based on mood and urgency," in *2014 IEEE International Symposium on Haptic, Audio and Visual Environments and Games (HAVE) Proceedings*, pp. 76–81, Oct. 2014.

[25] 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)*, pp. 235–240, June 2015.

[26] J. Ferguson, J. Williamson, and S. Brewster, "Evaluating Mapping Designs for Conveying Data Through Tactons," in *Proceedings of the 10th Nordic Conference on Human-Computer Interaction*, NordiCHI '18, (New York, NY, USA), pp. 215–223, ACM, 2018. event-place: Oslo, Norway.

[27] M.-W. Lin, Y.-M. Cheng, and W. Yu, "Using Tactons to Provide Navigation Cues in Pedestrian Situations," in *Proceedings of the 5th Nordic Conference on Human-computer Interaction: Building Bridges*, NordiCHI '08, (New York, NY, USA), pp. 507–510, ACM, 2008. event-place: Lund, Sweden.

[28] D. J. Barber, L. E. Reinerman-Jones, and G. Matthews, "Toward a Tactile Language for Human–Robot Interaction: Two Studies of Tacton Learning and Performance," *Human Factors*, vol. 57, pp. 471–490, May 2015.

[29] T. Pakkanen, R. Raisamo, and V. Surakka, "Audio-Haptic Car Navigation Interface with Rhythmic Tactons," in *Haptics: Neuroscience, Devices, Modeling, and Applications* (M. Auvray and C. Duriez, eds.), Lecture Notes in Computer Science, pp. 208–215, Springer Berlin Heidelberg, 2014.

[30] P. Brooks, B. Frost, J. Mason, and K. Chung, "Acquisition of a 250-word vocabulary through a tactile vocoder," *The Journal of the Acoustical Society of America*, vol. 77, pp. 1576–9, May 1985.

[31] Y. Jiao, F. M. Severgnini, J. S. Martinez, J. Jung, H. Z. Tan, C. M. Reed, E. C. Wilson, F. Lau, A. Israr, R. Turcott, K. Klumb, and F. Abnousi, "A Comparative Study of Phoneme- and Word-Based Learning of English Words Presented to the Skin," in *Haptics: Science, Technology, and Applications* (D. Prattichizzo, H. Shinoda, H. Z. Tan, E. Ruffaldi, and A. Frisoli, eds.), Lecture Notes in Computer Science, (Cham), pp. 623–635, Springer International Publishing, 2018.

[32] S. Zhao, A. Israr, F. Lau, and F. Abnousi, "Coding Tactile Symbols for Phonemic Communication," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, (New York, NY, USA), pp. 1–13, Association for Computing Machinery, Apr. 2018.

[33] M. F. d. Vargas, A. Weill-Duflos, and J. R. Cooperstock, "Haptic Speech Communication Using Stimuli Evocative of Phoneme Production," in *2019 IEEE World Haptics Conference (WHC)*, pp. 610–615, July 2019.

[34] H. Tan, C. Reed, Y. Jiao, Z. Perez, E. Wilson, J. Jung, J. Martinez, and F. Severgnini, "Acquisition of 500 English Words through a TActile Phonemic Sleeve (TAPS)," *IEEE Transactions on Haptics*, vol. PP, pp. 1–1, Feb. 2020.

[35] L. A. Jones and A. Singhal, "Perceptual dimensions of vibrotactile actuators," in *2018 IEEE Haptics Symposium (HAPTICS)*, pp. 307–312, Mar. 2018.

[36] E. E. Hoggan and S. A. Brewster, "Crossmodal Icons for Information Display," in *CHI '06 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '06, (New York, NY, USA), pp. 857–862, ACM, 2006. event-place: Montréal, Québec, Canada.

[37] E. Hoggan and S. Brewster, "New parameters for tacton design," in *CHI '07 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '07, (New York, NY, USA), pp. 2417–2422, Association for Computing Machinery, Apr. 2007.

[38] D. Ternes and K. E. MacLean, "Designing Large Sets of Haptic Icons with Rhythm," in *Haptics: Perception, Devices and Scenarios* (M. Ferre, ed.), Lecture Notes in Computer Science, pp. 199–208, Springer Berlin Heidelberg, 2008.

[39] E. Hoggan, R. Raisamo, and S. A. Brewster, "Mapping Information to Audio and Tactile Icons," in *Proceedings of the 2009 International Conference on Multimodal Interfaces*, ICMI-MLMI '09, (New York, NY, USA), pp. 327–334, ACM, 2009. event-place: Cambridge, Massachusetts, USA.

[40] M. Azadi and L. Jones, "Identification of vibrotactile patterns: building blocks for tactons," in *2013 World Haptics Conference (WHC)*, pp. 347–352, Apr. 2013.

[41] H. Qian, R. Kuber, and A. Sears, "Tactile Notifications for Ambulatory Users," in *CHI '13 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '13, (New York, NY, USA), pp. 1569–1574, ACM, 2013. event-place: Paris, France.

[42] M. Ernst and A. Girouard, "Exploring Haptics for Learning Bend Gestures for the Blind," in *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, CHI EA '16, (New York, NY, USA), pp. 2097–2104, ACM, 2016. event-place: San Jose, California, USA.

[43] 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)*, pp. 52–58, Apr. 2016.

[44] T. Stein, M. Seeger, B.-B. Borys, and L. Schmidt, "Design Recommendations for Tactons in Touch Screen Interaction," *IADIS International Journal*, p. 16, 2017.

[45] D. C. Egloff, M. M. Wanderley, and I. Frissen, "Haptic display of melodic intervals for musical applications," in *2018 IEEE Haptics Symposium (HAPTICS)*, pp. 284–289, Mar. 2018.

[46] H. Seifi, M. Chun, and K. E. Maclean, "Toward Affective Handles for Tuning Vibrations," *ACM Trans. Appl. Percept.*, vol. 15, pp. 22:1–22:23, July 2018.

[47] S. Guest, J. M. Dessirier, A. Mehrabyan, F. McGlone, G. Essick, G. Gescheider, A. Fontana, R. Xiong, R. Ackerley, and K. Blot, "The development and validation of sensory and emotional scales of touch perception," *Attention, Perception, & Psychophysics*, vol. 73, pp. 531–550, Feb. 2011.

[48] O. S. Schneider and K. E. MacLean, "Improvising design with a Haptic Instrument," in *2014 IEEE Haptics Symposium (HAPTICS)*, (Houston, TX, USA), pp. 327–332, IEEE, Feb. 2014.

[49] H. Seifi, C. Anthonypillai, and K. E. MacLean, "End-user customization of affective tactile messages: A qualitative examination of tool parameters," in *IEEE Haptics Symposium (HAPTICS)*, pp. 251–256, Feb. 2014.

[50] Clark, "Predictable and distinguishable morphing of vibrotactile rhythm - IEEE Conference Publication," 2017.

[51] D. N. Osherson, E. E. Smith, O. Wilkie, A. López, and E. Shafir, "Category-based induction," *Psychological Review*, vol. 97, no. 2, pp. 185–200, 1990.

[52] R. L. Goldstone, "The role of similarity in categorization: providing a groundwork," *Cognition*, vol. 52, pp. 125–157, Aug. 1994.

[53] J. Hampton, "Similarity-Based Categorization and Fuzziness of Natural Categories," *Cognition*, vol. 65, pp. 137–65, Feb. 1998.

[54] F. W. Young and R. M. Hamer, *Multidimensional scaling: history, theory, and applications*. Hillsdale, N.J.: L. Erlbaum Associates, 1987. OCLC: 13792004.

[55] D. M. Ennis and N. L. Johnson, "Thurstone-Shepard Similarity Models as Special Cases of Moment Generating Functions," *Journal of Mathematical Psychology*, vol. 37, pp. 104–110, Mar. 1993.

[56] J. D. Carroll and J.-J. Chang, "Analysis of individual differences in multidimensional scaling via an n-way generalization of "Eckart-Young" decomposition," *Psychometrika*, vol. 35, pp. 283–319, Sept. 1970.

[57] G.-D. Guo, A. Jain, W.-Y. Ma, and H.-J. Zhang, "Learning similarity measure for natural image retrieval with relevance feedback," *IEEE Transactions on Neural Networks*, vol. 13, pp. 811–820, July 2002.

[58] E. Xing, M. Jordan, S. J. Russell, and A. Ng, "Distance Metric Learning with Application to Clustering with Side-Information," in *Advances in Neural Information Processing Systems* (S. Becker, S. Thrun, and K. Obermayer, eds.), vol. 15, pp. 521–528, MIT Press, 2003.

[59] L. L. Thurstone, "A law of comparative judgment," *Psychological Review*, vol. 34, no. 4, pp. 273–286, 1927.

[60] T. Jehan, "EVENT-SYNCHRONOUS MUSIC ANALYSIS / SYNTHESIS," in *Proceedings of the COST G-6 Conference on Digital Audio Effects (DAFx-04)*, pp. 361–366, 2004.

[61] D. F. Silva, C.-C. M. Yeh, G. E. A. P. A. Batista, and E. J. Keogh, "SiMPle: Assessing Music Similarity Using Subsequences Joins," in *ISMIR*, 2016.

[62] M. Casey, "General sound classification and similarity in MPEG-7," *Organised Sound*, vol. 6, pp. 153–164, Aug. 2001.

[63] M. Cooper and J. Foote, "Automatic Music Summarization via Similarity Analysis," *Proceedings of the 3rd International Conference on Music Information Retrieval, ISMIR*, Aug. 2002.

[64] J. Pasquero, J. Luk, S. Little, and K. MacLean, "Perceptual Analysis of Haptic Icons: an Investigation into the Validity of Cluster Sorted MDS," in *14th Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems*, (Alexandria, VA, USA), pp. 437–444, IEEE, 2006.

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

[66] G. Park and S. Choi, "Perceptual space of amplitude-modulated vibrotactile stimuli," in *2011 IEEE World Haptics Conference*, pp. 59–64, June 2011.

[67] I. Hwang, J. Seo, and S. Choi, "Perceptual Space of Superimposed Dual-Frequency Vibrations in the Hands," *PLOS ONE*, vol. 12, p. e0169570, Jan. 2017.

[68] T. Kaaresoja and J. Linjama, "Perception of Short Tactile Pulses Generated by a Vibration Motor in a Mobile Phone," in *Proceedings of the First Joint Eurohaptics Conference and Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems*, WHC '05, (USA), pp. 471–472, IEEE Computer Society, Mar. 2005.

[69] A. Israr, S. Zhao, and O. Schneider, "Exploring Embedded Haptics for Social Networking and Interactions," in *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*, CHI EA '15, (Seoul, Republic of Korea), pp. 1899–1904, Association for Computing Machinery, Apr. 2015.

[70] E. Hoggan, C. Stewart, L. Haverinen, G. Jacucci, and V. Lantz, "Pressages: augmenting phone calls with non-verbal messages," in *Proceedings of the 25th annual ACM symposium on User interface software and technology - UIST '12*, (Cambridge, Massachusetts, USA), p. 555, ACM Press, 2012.

[71] F. A. Geldard, "Adventures in tactile literacy," *American Psychologist*, vol. 12, no. 3, pp. 115–124, 1957.

[72] L. A. Jones and N. B. Sarter, "Tactile Displays: Guidance for Their Design and Application:," *Human Factors*, Feb. 2008.

[73] T. Nukarinen, *Assisting Navigation and Object Selection with Vibrotactile Cues*. Tampereen yliopisto, 2019.

[74] M. Pastor, B. Day, E. Macaluso, K. Friston, and R. Frackowiak, "The Functional Neuroanatomy of Temporal Discrimination," *The Journal of neuroscience : the official journal of the Society for Neuroscience*, vol. 24, pp. 2585–2591, Apr. 2004.

[75] N. Ailon, "Reconciling Real Scores with Binary Comparisons: A New Logistic Based Model for Ranking," in *Advances in Neural Information Processing Systems* (D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, eds.), vol. 21, pp. 25–32, Curran Associates, Inc., 2009.

[76] J. Farrington, "Seven plus or minus two," *Performance Improvement Quarterly*, vol. 23, no. 4, pp. 113–116, 2011.

[77] E. Law, B. Settles, and T. Mitchell, "Learning to Tag from Open Vocabulary Labels," in *Machine Learning and Knowledge Discovery in Databases* (D. Hutchison, T. Kanade, J. Kittler, J. M. Kleinberg, F. Mattern, J. C. Mitchell, M. Naor, O. Nierstrasz, C. Pandu Rangan, B. Steffen, M. Sudan, D. Terzopoulos, D. Tygar, M. Y. Vardi, G. Weikum, J. L. Balcázar, F. Bonchi, A. Gionis, and M. Sebag, eds.), vol. 6322, pp. 211–226, Berlin, Heidelberg: Springer Berlin Heidelberg, 2010.

[78] E. Law, K. West, and M. Mandel, "EVALUATION OF ALGORITHMS USING GAMES: THE CASE OF MUSIC TAGGING," *Oral Session*, p. 6, 2009.

[79] C. Eickhoff and A. P. de Vries, "Increasing cheat robustness of crowdsourcing tasks," *Information Retrieval*, vol. 16, pp. 121–137, Apr. 2013.

[80] E. Law, M. Yin, J. Goh, K. Chen, M. A. Terry, and K. Z. Gajos, "Curiosity Killed the Cat, but Makes Crowdwork Better," in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems - CHI '16*, (Santa Clara, California, USA), pp. 4098–4110, ACM Press, 2016.

[81] F. Daniel, P. Kucherbaev, C. Cappiello, B. Benatallah, and M. Allahbakhsh, "Quality Control in Crowdsourcing: A Survey of Quality Attributes, Assessment Techniques, and Assurance Actions," *ACM Comput. Surv.*, vol. 51, pp. 7:1–7:40, Jan. 2018.

[82] M. Allahbakhsh, B. Benatallah, A. Ignjatovic, H. R. Motahari-Nezhad, E. Bertino, and S. Dustdar, "Quality Control in Crowdsourcing Systems: Issues and Directions," *IEEE Internet Computing*, vol. 17, pp. 76–81, Mar. 2013.

[83] E. Law, K. Z. Gajos, A. Wiggins, M. L. Gray, and A. Williams, "Crowdsourcing As a Tool for Research: Implications of Uncertainty," in *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '17, (New York, NY, USA), pp. 1544–1561, ACM, 2017. event-place: Portland, Oregon, USA.

[84] A. Hultman, *Simple Affective Hapticons using Web Technologies*. Student thesis, KTH, School of Electrical Engineering and Computer Science (EECS)., 2019.

[85] R. A. Bradley and M. E. Terry, *The rank analysis of incomplete block designs 1: the method of paired comparisons*. Blacksburg, VA: Virginia Agricultural Experiment Station, 1952. OCLC: 48031947.

[86] G. King and L. Zeng, "Logistic Regression in Rare Events Data," *Political Analysis*, vol. 9, pp. 137–163, 2001.

[87] J. P. Dotson, J. R. Howell, J. D. Brazell, T. Otter, P. J. Lenk, S. MacEachern, and G. M. Allenby, "A Probit Model with Structured Covariance for Similarity Effects and Source of Volume Calculations," *Journal of Marketing Research*, vol. 55, pp. 35–47, Feb. 2018.

[88] X. Chen, P. N. Bennett, K. Collins-Thompson, and E. Horvitz, "Pairwise ranking aggregation in a crowdsourced setting," in *Proceedings of the sixth ACM international conference on Web search and data mining - WSDM '13*, (Rome, Italy), p. 193, ACM Press, 2013.

[89] L. J. Savage and J. Wiley & Sons, Inc., "The foundations of statistics.," *Naval Research Logistics Quarterly*, vol. 1, no. 3, pp. 236–236, 1954.

[90] J. C. Handley, "Comparative Analysis of Bradley-Terry and Thurstone-Mosteller Paired Comparison Models for Image Quality Assessment," *Proc. IS&T's Image Processing, Image Quality, Image Capture, Systems Conference*, p. 5, 2001.

[91] J. Li, R. K. Mantiuk, J. Wang, S. Ling, and P. L. Callet, "Hybrid-MST: A Hybrid Active Sampling Strategy for Pairwise Preference Aggregation," *arXiv:1810.08851 [cs, stat]*, Oct. 2018. arXiv: 1810.08851.

[92] N. Perraudin, N. Holighaus, P. Majdak, and P. Balazs, "Inpainting of Long Audio Segments With Similarity Graphs," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, pp. 1083–1094, June 2018.

[93] W. Rafique, M. Khan, N. Sarwar, M. Sohail, A. Irshad, I. S. Bajwa, F. Kamareddine, and A. Costa, *A Graph Theory Based Method to Extract Social Structure in the Society*. Communications in Computer and Information Science, Singapore: Springer, 2019.

[94] C. Theoharatos, V. K. Pothos, N. A. Laskaris, G. Economou, and S. Fotopoulos, "Multivariate image similarity in the compressed domain using statistical graph matching," *Pattern Recognition*, vol. 39, pp. 1892–1904, Oct. 2006.

[95] T. N. Kipf and M. Welling, "Semi-Supervised Classification with Graph Convolutional Networks," *arXiv:1609.02907 [cs, stat]*, Feb. 2017. arXiv: 1609.02907.

[96] W. L. Hamilton, R. Ying, and J. Leskovec, "Inductive Representation Learning on Large Graphs," *arXiv:1706.02216 [cs, stat]*, Sept. 2018. arXiv: 1706.02216.

[97] W. James, *The Principles of Psychology*. Cosimo, Inc., 1890. Google-Books-ID: TMrJfcaC8bYC.

[98] A. Tversky, "Features of similarity," *Psychological Review*, vol. 84, no. 4, pp. 327–352, 1977.

[99] N. Gaißert, H. H. Bülthoff, and C. Wallraven, "Similarity and categorization: From vision to touch," *Acta Psychologica*, vol. 138, pp. 219–230, Sept. 2011.

[100] L. McInnes, J. Healy, and J. Melville, "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction," *arXiv:1802.03426 [cs, stat]*, Dec. 2018. arXiv: 1802.03426.

[101] R. J. G. B. Campello, D. Moulavi, and J. Sander, "Density-Based Clustering Based on Hierarchical Density Estimates," in *Advances in Knowledge Discovery and Data Mining* (J. Pei, V. S. Tseng, L. Cao, H. Motoda, and G. Xu, eds.), Lecture Notes in Computer Science, (Berlin, Heidelberg), pp. 160–172, Springer, 2013.

[102] L. McInnes and J. Healy, "Accelerated Hierarchical Density Clustering," *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, pp. 33–42, Nov. 2017. arXiv: 1705.07321.

[103] V. Balntas, E. Riba, D. Ponsa, and K. Mikolajczyk, "Learning local feature descriptors with triplets and shallow convolutional neural networks," in *Procedings of the*

British Machine Vision Conference 2016, (York, UK), pp. 119.1–119.11, British Machine Vision Association, 2016.

[104] P. K, S. Chaudhuri, and S. Chaudhuri, "PerceptNet: Learning Perceptual Similarity of Haptic Textures in Presence of Unorderable Triplets," *arXiv:1905.03302 [cs, stat]*, May 2019. arXiv: 1905.03302.

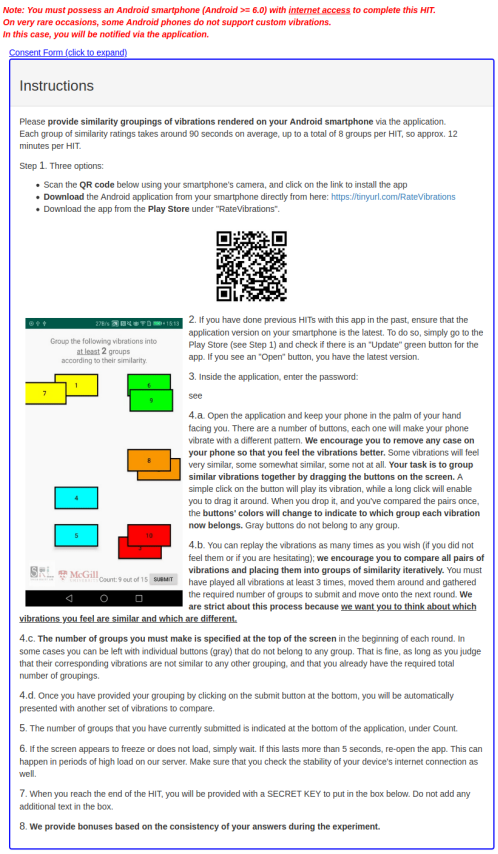
[105] M. Fey and J. E. Lenssen, "Fast Graph Representation Learning with PyTorch Geometric," *arXiv:1903.02428 [cs, stat]*, Apr. 2019. arXiv: 1903.02428.

[106] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *arXiv:1412.6980 [cs]*, Jan. 2017. arXiv: 1412.6980.

Appendices

Appendix A

Instruction Form



Note: You must possess an Android smartphone (Android >= 6.0) with internet access to complete this HIT.
On very rare occasions, some Android phones do not support custom vibrations.
In this case, you will be notified via the application.

Consent Form (click to expand)

Instructions

Please **provide similarity groupings of vibrations rendered on your Android smartphone** via the application. Each group of similarity ratings takes around 90 seconds on average, up to a total of 8 groups per HIT, so approx. 12 minutes per HIT.

Step 1. Three options:

- Scan the **QR code** below using your smartphone's camera, and click on the link to install the app
- **Download** the Android application from your smartphone directly from here: https://tinyurl.com/RateVibrations
- Download the app from the **Play Store** under "RateVibrations".

Group the following vibrations into at least 2 groups according to their similarity.

Count: 9 out of 15 SUBMIT

2. If you have done previous HITs with this app in the past, ensure that the application version on your smartphone is the latest. To do so, simply go to the Play Store (see Step 1) and check if there is an "Update" green button for the app. If you see an "Open" button, you have the latest version.

3. Inside the application, enter the password:

see

4.a. Open the application and keep your phone in the palm of your hand facing you. There are a number of buttons, each one will make your phone vibrate with a different pattern. **We encourage you to remove any case on your phone so that you feel the vibrations better.** Some vibrations will feel very similar, some somewhat similar, some not at all. **Your task is to group similar vibrations together by dragging the buttons on the screen.** A simple click on the button will play its vibration, while a long click will enable you to drag it around. When you drop it, and you've compared the pairs once, the **buttons' colors will change to indicate to which group each vibration now belongs.** Gray buttons do not belong to any group.

4.b. You can replay the vibrations as many times as you wish (if you did not feel them or if you are hesitating); **we encourage you to compare all pairs of vibrations and placing them into groups of similarity iteratively.** You must have played all vibrations at least 3 times, moved them around and gathered the required number of groups to submit and move onto the next round. **We are strict about this process because we want you to think about which vibrations you feel are similar and which are different.**

4.c. **The number of groups you must make is specified at the top of the screen** in the beginning of each round. In some cases you can be left with individual buttons (gray) that do not belong to any group. That is fine, as long as you judge that their corresponding vibrations are not similar to any other grouping, and that you already have the required total number of groupings.

4.d. Once you have provided your grouping by clicking on the submit button at the bottom, you will be automatically presented with another set of vibrations to compare.

5. The number of groups that you have currently submitted is indicated at the bottom of the application, under Count.

6. If the screen appears to freeze or does not load, simply wait. If this lasts more than 5 seconds, re-open the app. This can happen in periods of high load on our server. Make sure that you check the stability of your device's internet connection as well.

7. When you reach the end of the HIT, you will be provided with a SECRET KEY to put in the box below. Do not add any additional text in the box.

8. **We provide bonuses based on the consistency of your answers during the experiment.**

Fig. A.1 Instruction form presented to participants before completing the experiment.

Appendix B

Post-experiment Questionnaire

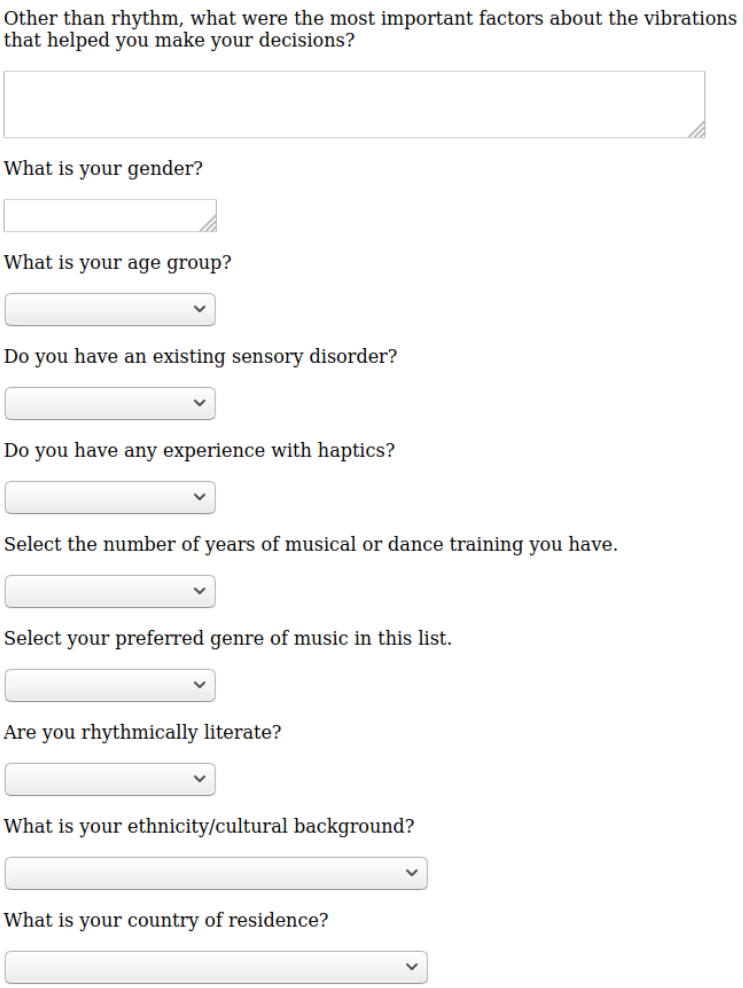


Fig. B.1 Post-experiment questionnaire.

Appendix C

Persona Groups

Table C.1: Persona groups composition.

	FPC ratings
0	000020131200011, 011010022300001, 011011130310100
1	000000000120100, 000002212211210, 000010030001010, 000031210002212, 000100021000000, 000202002101100, 000210022010101, 000302103212200, 002000012001001, 002000020010000, 002000020102000, 002000021001000, 002000030203000, 002012020100001, 003000012000000, 003000020003000, 003000021002000, 003000021010000, 003000030002000, 003000030003000, 003000030102000, 003000031002001, 003000130003000, 003010030010010, 003012020001011, 003030030000030, 003101023201100, 010000020000010, 010000030000000, 012001031002011, 013010030011010, 013202011120200, 020000030000200, 020010101000201, 021000030001020, 022010031100021, 023000003200000, 023011030010010, 023030030202030, 023300002220300, 030000003000300, 030000010000200, 030012110001322, 033000003300001, 033020030302020, 033030030303030, 100110010010010, 110020000002110, 112010003100000, 113110030103110, 120032221110300, 121010112001210, 130001000000300, ...
2	000010302000020, 000100211000021, 010000301000010, 010302300010000, 102020202111031
3	000210010300000, 032000020200000, 110120110200000, 211230030301010, 222011110200000
4	000020200020000, 001001100011101, 001002000000000, 001010012110100, 002000000011000, 002000000111000, 002000001010001, 002000001111000, 002002003002101, 002002012011002, 002100001000002, 002133020020020, 003001001011001, 003001010200000, 003003032032002, 003021031211021, 003032030121030, 003303003003300, 010032000000000, 011000101010000, 012000002121001, 012000011011003, 012000100101000, 021020200203001, 021200202210101, 031100100110001, 100020002220111, 101001300100002, 112010001120001, 202000000211003
5	000100120201000, 003000000001000, 003000000003000, 003000000012000, 003000001012000, 003000003030000, 003000012020000, 003000100002001, 003001022020000, 003003030030000, 003010000000010, 012000012020010, 013000001102000, 013000013120000, 101100030100000, 102000010012100, 102100010010200

Appendix D

Description of the Features

Feature	Formula	Description
Energy	$\frac{T \cdot T}{N}$	Signal energy; in the case of binary tactons, it is equivalent to the sum of the signal.
Complexity	$\sum_{i=2}^{N} \sqrt{(x_i - x_{i-1}) \cdot (x_i - x_{i-1})}$	Complexity measures the energy of the signal derivative.
Non-linearity	$\sum_{i=1}^{N} S \cdot roll(T,4) \cdot roll(T,2)$	Non-linearity measures the average difference between two delayed versions of a signal, and serves to evaluate how many switches in frequency are in the signal.
Autocorrelation	$argmax(T \cdot conv(T))$	Traditional cross-correlation of the signal with a 1-lagged version of itself.
Spectral Rolloff	(outside scope)	Frequency at which signal harmonics are filtered out. Calculating Spectral Rolloff is outside the scope of this table, and should be done using specialized libraries.
Binary entropy	$-log_2(\frac{T}{1-T})$	Measure of uncertainty in the signal.
First location of minimum	-	Index of first value of 0.
Last location of minimum	-	Index of last value of 0.
First location of minimum	-	Index of first value of 1.
Last location of minimum	-	Index of last value of 1.
Num. unique dbl buzzes	-	Number of unique sequences of pairs 1's in the signal.
Num. unique trpl buzzes	-	Number of unique sequences of triplets 1's in the signal.
Num. unique four plus buzzes	-	Number of unique sequences of four or more 1's in the signal.
Num. ramp ups	-	Number of passages from 0 to 1 in the signal.
Num. ramp downs	-	Number of passages from 1 to 0 in the signal.

Appendix E

Learning Experiments Details

E.1 Persona Clustering

E.1.1 Hyperparameters

Table E.1: Hyperparameters for Persona Clustering using UMAP and HDBSCAN.

Parameter	Value
UMAP number of neighbors	2
UMAP minimum distance	0
UMAP number of components	2
HDBSCAN minimum cluster size	2

E.2 Graph Representation Learning

E.2.1 Hyperparameters

Table E.2: Hyperparameters for graph representation learning.

Parameter	Value
UMAP number of neighbors	5
UMAP minimum distance	0
UMAP number of components	2
HDBSCAN minimum cluster size	6
GCN number of layers	2
GCN hidden size dimension	16, 32
GCN normalization at each layer	true
Batch size (1 batch per epoch)	32
Number of epochs	1250
Learning rate	5e-4
margin α	1e-4

E.2.2 Train/Test Loss Curves

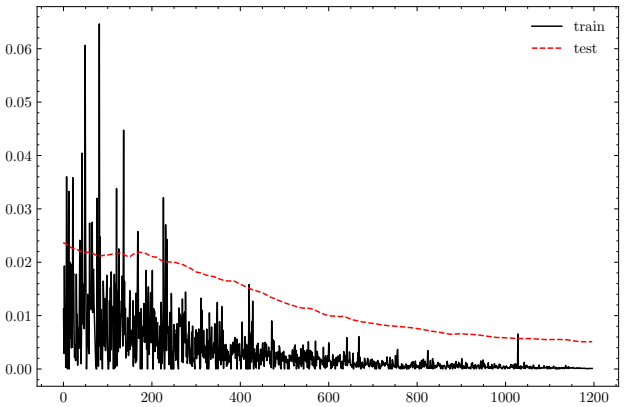


Fig. E.1 Triplet loss training and testing loss curves, plotted for 1 seed and a single fold for ease of view.

E.3 Regressor Hyperparameters

Table E.3: Hyperparameters for gradient boosting.

Parameter	Value
Number of estimators	2000
Learning rate	0.06
Number of leaves	750
Regularizer α	2.40
Regularizer λ	0.0
Fraction of features at every split	0.95
Bagging fraction	0.96
Minimum child weight	4e-7
Objective	L2 regression