Aalto University
School of Science

Master's programme in Computer, Communication and Information Sciences

Deciphering Multimodal Correspondence using Exploratory Data Analysis

Asutosh Hota

Master's Thesis
2023

Aalto University
School of Science

Author Asutosh Hota

Title Deciphering Multimodal Correspondence using Exploratory Data Analysis

Degree programme Computer, Communication and Information Sciences

Major Human-Computer Interaction, SCI3097

Supervisor Dr. Tapio "Tassu" Takala, Professor Emeritus (retired), Aalto University

Advisor MFA. Jaana Okulov, Doctoral Researcher, Aalto University

Date 28.05.2023 **Number of pages** 78+29 **Language** English

Abstract

Artistic and creative processes rely on integrating information from multiple sensory modalities. However, understanding the complex interplay between these modalities and how they correlate remains a challenge. The methods followed in conventional behavioural and psychological experiments have been consistently qualitative and the correlations/correspondence have been traditionally found on the basis of the choices that the human participant thinks (pair-matching). These have proven to be the existential foundation of multimodal correlation studies however, a lack of a quantitative approach limits this experimental methodology to test only a few numbers of participants. Conventional pair/pattern matching experiments may not fully capture the underlying correlations in sensory multimodal data and Exploratory Data Analysis (EDA) based approaches can reveal hidden trends and insights. This thesis proposes Primary Evaluator for Multimodal Correlation (PEMC), a novel framework which provides a data-driven approach for exploring correlations between two or more sensory modalities. The framework emphasizes the importance of EDA techniques in identifying hidden patterns in sensory multimodal data, which may not be captured through conventional pair/pattern matching experiments. Utilizing various EDA techniques, such as dimensionality reduction, unsupervised clustering, and correlation analysis, we propose the Correlation Analyser (CA), an integral part of PEMC. CA is used to identify correlations between two modalities. PEMC framework tries to conduct a preliminary evaluation of the existence of underlying correlations in sensory data using CA in 3 unique test settings. The results suggest that there exist multimodal correlations and recommend whether more controlled experiments are needed to establish the presence of universal multimodal correlations.

In this thesis, we conduct an in-depth analysis of sensory multimodal data extracted from audio responses, pen movement responses, and colour transition data as stimulus data using the PEMC. Our findings reveal moderate to strong correlations in the features of audio and pen movement data in response to colour transition data, providing valuable insights into how different modalities interact and influence each other. Potential limitations of the framework, best practices and many applications of the correlation analysis are also discussed giving directions to future studies.

Keywords Multimodal Correlation, Machine Learning, Exploratory Data Analysis

Preface

I want to thank my supervisor Prof. Tapio Takala and my advisor MFA Jaana Okulov for their support and guidance throughout the thesis process. This journey has been one of immense personal and intellectual growth. During my study at Aalto University, it has been a privilege to collaborate with esteemed mentors, researchers, and peers, who have provided guidance, support, and stimulating discussions. Special gratitude to Prof. Antti Oulasvirta and Prof. Luis A. Leiva for giving me the opportunity to work in the User Interfaces group and contribute to applied AI and design research. This has paved the way for my future doctoral studies and work in the Finnish data and AI industry.

This work is dedicated to my support system: my parents Mr Bibhuti Bhusan Hota and Mrs Sagarika Hota and my dear wife Priyanka Mishra for constantly supporting me. Special regards to my loving family members, Lt. Mr Chitta Ranjan Mohapatra and Lt. Mrs Arnapurna Hota whom I lost during the Covid-19 pandemic years.

Otaniemi, 25.05.2023

Asutosh Hota

Contents

1 Abbreviations

2 Symbols

ΔC	Change in the Color space
ΔH	Change in the Hue space
ΔS	Change in the Saturation space
ΔL	Change in the Lightness/Luminance space

3 Introduction

A promising paradigm that blends human creativity and the computational powers of Artificial Intelligence (AI) is the Human-AI co-creation of art, which involves artists working alongside AI systems. How to effectively allow communication and collaboration between artists and AI technologies is a fundamental difficulty in this field. Multimodal communications, which is the use of several forms of communication, including text, audio, and visual, can be extremely important in resolving this issue. DALLE 2 [1] and Midjourney [2] are AI tools which use language models for artistic output with AI. By mixing concepts, traits, and styles, DALLE 2 and Midjourney can generate unique, realistic visuals and art from a text description. They can also overcome specification in three ways: changing the style, setting, and time; drawing the same thing in various contexts; and producing an image of an object with particular text put on it. However, these artistic creations have been mostly limited to text or images as input. In order to find ways in which humans and AI can communicate and create art from different modalities, understanding the relationship between the different types of data in different modalities becomes an important starting point. This work investigates some of the challenging avenues of research for finding multimodal correlations from heterogeneous data using exploratory data analysis. Finding methods to create quality multimodal datasets and conducting experiments to improve our understanding of multimodal correlations can further help researchers in advanced AI domain, cognitive science and psychology fields to understand how humans experience different modalities from their sensory perception.

Let's start with an example of why sensory multimodal correlation is important. Imagine you are driving a car on a fairly empty highway and you have to take a turn in about 1km. Just before making the turn, before a few hundred meters, suddenly you see another car speeding towards you on the same lane and you anticipate it is about to take the turn as well. However, the car doesn't indicate it explicitly and you are left with 2 options to decide, either slow down and let the car take the turn first or speed up to stay ahead of the other car. It is to be noted here that there are multiple information sources that you need to process internally to reach one of the possible outcomes: the visual cues from the rear-view mirror, the speed of the car is increasing rapidly or not, the indicators from the car, any honking/signal sounds etc. Some of this information might be highly useful for you to make the decision, in the above case as there were no indications from the car or any honking sounds, the information of a speeding car towards you might be the most important one to help you decide your actions. Hence, while deciding on a certain action, more emphasis is given to one of the many modalities present during the task. These individual subjective experiences that a person is currently in could be considered as 'qualia', subjective experiences of the real world. Qualia contributes to the very way we interact with the environment. From a Human Computer Interaction (HCI) perspective, humans while experiencing their surroundings focus primarily, but not exclusively, on four modalities: natural language or speech (audition), which is either oral or handwritten; touch (tactition), for understanding textural sensation (thermoception), for example, heat or cold; visual signals (vision), visual perception of the world; and vocal signals (audition) which

encode sounds and para-verbal information such as prosody and vocal expressions.

> "Modality refers to the way in which something happens or is experienced and a research problem is characterized as multimodal when it includes multiple such modalities" [3].

The study of building intelligent systems to process information from multiple modalities referred to as Multimodal Machine Learning (MML), is of increasing importance as a multidisciplinary field. Multimodal correlation is crucial for AI because it enables computers to interpret and comprehend input from several sources, including images, audio, text, and other types of data. AI systems can develop a more thorough and accurate understanding of the world by merging these several modalities, which can aid them in developing stronger prediction skills. For example, a self-driving automobile needs to cross a busy junction. The car can accurately recognize and react to traffic lights, pedestrians, and other vehicles in real-time utilizing a combination of visual data from cameras, auditory data from microphones, and sensor data from the car's Global Positioning System (GPS) system. Sensory multimodal tendencies, for example, the match between brightness-pitch, elevation-pitch, shape-odour, sound-odour, flavour-shape, and weight-hue are found universally and these cross-modal pairings seem to exist between all possible sensory modalities [4]. [5] acknowledges that the research so far has focused on studying each modality in isolation, however, most of the responses that humans give while interacting with the surroundings are generated from the effects of multiple modalities simultaneously. For example, while watching a horror movie in the theatre, the audio-visual stimulus instils the sensation of fear in the viewer's mind from the effective use of sound and darkness. Hence, it becomes necessary to study multimodality from a continuous effect of multiple modalities and not only study individual modalities in isolation. Cross-modal studies are seeking correlations in modalities often through static stimuli even though the literature clearly implies that the phenomenon is relative and can be studied through dynamic comparisons. These are the connections most often appearing in the literature: darker colours are related to quieter sounds [4], the high pitch is associated with lightness, brightness to height in the vertical direction, and the smaller size of an object (e.g., [6, 7]). High pitch is also associated with an angular shape, whereas lower tone is associated more often with a U-shape, Pseudo words "maluma", "baluma" and "bouba" are matched to a globular rounded shape and "takete" or "kiki" to straight-edged angular shape [8], and "mil" is more often linked to small objects when "mal" to large ones [4]. Despite the advancements in analysis tools and improvements in deep learning, there exist a few challenges which make the study of these multimodal correlations quite difficult:

1. Different modalities have different alignments, for example, continuous and sparse signals

2. There exists a difference in noise levels due to the heterogeneous nature of multimodal data.

3. Multimodal data is complex to represent and translate from one modality to another, extracting different features from multiple data types might be another challenging aspect.

4. It can be challenging to transfer knowledge from one modality to another.

For systematically studying these different modalities and modelling computational models that can represent, translate or map sensory multimodal data, it becomes essential to extract quality features from the data and find suitable correlations in them to find out which information is most relevant. The way these features are extracted and selected can affect the very way these computational models could be designed. For example, [9] shows that selective attention in case of a cluttered display can actually help to identification of a visual shape, or the location information determines the features information of the visual shape proficiently. A similar study also uses a location-based attention mechanism while investigating the emotional effects of audiovisual, auditory and visual in emotion recognition tasks. The study supports the view that affective information from face and prosody converges at higher association cortices in the human brain [10]. Attention mechanisms are a fairly new concept in deep learning literature, also called attention models, machine attention, neural attention or artificial attention. The attention mechanism was built to help memorize long source sentences in Neural Machine Translation (NMT) tasks but soon it was realised that they worked well with image data in image classification tasks as well. So far attention mechanisms have mostly been studied from the visual perspective but in order to understand truly the phenomenon in general, other modalities need to be studied as well. For instance, orientation to a peripheral, salient sound affects visual processing. It enhances visual perception by not only boosting visual-cortical responses but the visual cortex activity is modulated even before the visual object actually appears [11]. Machine attention has been inspired by Human attention in many ways. Human attention has been studied by psychologists for over a century now and they have found the importance of it in enhancing the performance of humans in many tasks, i.e., due to the selective visual attention to a location in space or an object. The advent of neural attention in the deep learning paradigm has brought up a sense of 'interpretability' of how the neural network works in visual space, for example, showing salient areas where the machine focuses attention in case of an object detection task [12]. But in order to design multimodally capable systems in the future, the relationship between human attention and neural attention needs to be understood, not only in the visual space but also in other multimodal spaces so that future deep learning systems could be trained to understand the multimodal even better.

There are many types of sensory data that can be collected for a multitude of tasks in the multimodal domain but there is not a general approach or framework that exists to extract features from these highly heterogeneous data types and find suitable correlations amongst them. In this thesis, I examine multimodal correlation from two perspectives: 1) Machine Learning (ML) and statistical techniques and 2) artistic research (explained in detail in Background and Motivation). The key idea here is

to study the advancements in these multidisciplinary fields of research and leverage the techniques, ideas and concepts of each discipline to propose 'PEMC', a simple yet efficient framework to analyse sensory multimodal correlation. An experiment is designed and conducted taking ideas from artistic research and the collected data would be used to show how PEMC can be leveraged to find preliminary sensory multimodal correlations. It is to be noted here that future experiments might be necessary to test the presence of the preliminary correlations that emerge during the study conducted for this thesis.

3.1 Motivation

This thesis builds upon prior research of Jaana Okulov, a Doctoral candidate at Aalto University whose work investigates the concepts of Qualia and the realisation of these concepts in computational systems. She has worked extensively on these topics from an artistic point of view while working at Olento Collective, a small team of artists and creative individuals dedicated to developing "Olento", an artificial intelligence capable of learning via a modicum of multimodal information streams. This thesis collaborates with the artistic research standpoint from a quantitative approach and takes an exploratory data analysis route to realise the existence of more cross-modal correspondences in sensory data collected from humans. A data-driven approach makes it easier to collect more data from different modalities and analyse them systematically. The work of the Olento collective shows how multidisciplinary collaboration can lead to the finding of new ideas from the amalgamation of existing theories and concepts. For example, the collective noticed how certain changes in one modality (stimuli) were more attended to than others and expressed effectively. For example, they have discovered some unique findings relating to the behaviour of users' affective relationships while attending to colour, drawing lines, singing and dancing movements. One of the experiments they conducted involved a singer and a dancer trying to develop a mutual affective language through improvisation. Many similarities evolved which have been previously found in other seminal works in psychology research, for example, the well-studied multimodal correlation of pitch and brightness manifested itself in the research settings. Interestingly, new unique correlations emerged as well. For instance, a saturated lime-yellow colour was notably the "fastest" colour. Participants verbally described lime-color as the fastest colour and their bodily expression in relation to that colour had a "fast" quality associated with it. [13]. This very fact suggests that there could be more universal correlations that have not yet been realised. Hence, there is a need for a systematic approach to study the different kinds of heterogeneous data and find correlations that can help to solve other challenging problems in the multimodal domain like mapping, translation etc.

3.2 Research Questions

The Research Question (RQ) I have explored in this study are:

1. Can we find a framework that could help us test the sensory cross-modal

correlations in 2 or more modalities?

2. What kind of universal/individual correlations emerge in modalities (Stimulus: Color data in Hue, Saturation,Lightness/Luminance (HSL) space, Responses: Audio/Verbal responses and Pen Gestures drawing) when they are multimodally compared?

3. What are the advantages and use cases of finding sensory multimodal correlations?

The first and second question is explored by proposing a data-driven experiment framework, PEMC, to evaluate the existence of multimodal correlations. A systematic and controlled study with human users is conducted by leveraging the framework to find correlations in 3 heterogeneous data types (modalities). It was important to conduct an experiment collected from real-world samples which can make it easier for practitioners from other multidisciplinary fields to replicate in their respective use cases to evaluate multimodal correlation in the future. The third question is approached from a more theoretical standpoint, where I would like to discuss how can we use multimodal correlations in other relevant fields and their applications and use cases.

3.3 Contributions

The key contribution of my thesis to the field of sensory multimodal correspondence studies are:

1. Propose a Preliminary Evaluator for Multimodal Correlation (PEMC), a framework to study multimodal data correlations and find whether underlying universal correlations emerge. A user study is part of the experiment to see the efficiency of the approach. (RQ 1 and RQ 2)

2. Leverage statistical and machine learning techniques to propose a CA, an exploratory data analysis method to check for correlations in 2 modalities of data. (RQ 1)

3. Discussion on the future of use cases and applications of multimodal correlation. (RQ 3)

3.4 Methodological Frameworks

The approach followed in the thesis is both theoretical and experimental. 1) Understanding the concepts of correlations in a multimodal setting from the point of view of artistic studies and their suitable applications and use cases (theoretical) and 2) Using the advancements in the field of statistics and machine learning to analyse, visualize and study cross-modal heterogeneous data types (experimental). This type of study can well be classified as experimental psychology, studying different psychological phenomena through controlled studies but using a data-driven approach. This section briefly describes the multi-disciplinary and diverse nature of the different concepts which has helped to study the aforementioned research questions.

3.4.1 Artistic Studies

The work by J. Okulov [14] establishes that in art, "it is quite common to build an artwork by letting the environment influence the aesthetic expression, but the process is often implicit, intuitive, and un-quantified". These practices are a way to bring the sensation a stimulus is causing into expression. The original stimulus (in modality 1) and created gesture (in modality 2) should correlate in their qualia, and therefore tell something about the underlying phenomena. Recorded expressions can be used again as stimuli and triangulation can verify the nonverbal data. This approach of finding correlations of different modalities from human collected data to study and perform artistic phenomenons that is very intuitive for humans but difficult for a machine enforces the system to have an "intuition"(good understanding of the representation of multimodal data space) about the surroundings. Intuition is a human trait that involves instinctively knowing something without the need for conscious reasoning. In contrast to intuition, AI systems are designed to learn from data and make decisions based on that data. However, AI systems can be trained to spot patterns in data that humans may not see, leading to discoveries that appear intuitive. However, these insights are based on data and algorithms rather than gut or intuition, as humans do. This creates a unique opportunity to explore this space of artistic research which has the capability to realise other advanced research concepts like the hard problem of consciousness or qualia in behavioural psychology.

3.4.2 Machine Learning and Statistical techniques

Machine learning and statistical techniques can be deployed to draw information from a continuous stream of data. Aesthetics and non-verbal information such as audio, hand gestures, drawing with a stylus on a screen etc. are different multimodal data types which could be complex and high-dimensional in nature. There have been recent advancements in the field of many multimodal applications with the use of attention models in areas like audio-video localisation, image captioning, video captioning etc. which have paved the way for the exploration of machine attention in a much more explicit way. Apart from these, the studies have helped in a better representation of multimodal data which can be further processed and analysed systematically. But these studies have been limited to the visual and textual data space. Future experiments should be planned to study underlying principles of human psychology and computational models could be built on top of these data to experiment quantitatively on many aspects of human behaviour and how machines could be trained to perform those behaviours better.

3.5 Structure

The following sections will explain the background that led to the development of the PEMC Framework followed by a detailed outlook on the possible use cases and applications of the framework. The Background section discusses seminal works in the field of multimodality, psychology, machine and deep learning etc. to understand the

current state of multimodal research and the advancements that have made it possible to study the field further. Preliminary Evaluator for Multimodal Correlations (PEMC) section introduces the proposed framework for finding multimodal correlations and explains each component of the framework, their theoretical foundations and the techniques that are used in it. Experiments: Multimodal Correlation Using PEMC framework section shows how the framework was used to evaluate the multimodal correlation between multimodal data types of 3 modalities collected from human users. The key findings are discussed in the Discussions and Analysis section followed by Use Cases and Applications section to explore the potential applications of multimodal correlations. Future Work and Conclusion section discusses briefly the contribution and usefulness of the framework and concludes the thesis.

4 Background

In the past few decades researchers from different disciplines such as psychology, philosophy, neuroscience, cognitive science, HCI and deep learning, have approached the study of multimodality, qualia and consciousness. This section draws inspiration from those works and formulates the background of this thesis. In this section, I review the different theories associated with the key ideas of multimodal correspondences. The goal is to understand the commonalities between the different theories and find ways to design, develop and test a general framework that could be further used for detecting sensory multimodal correlations or correspondences in heterogeneous multimodal data.

4.1 Cross Multimodal Correspondences and Conventional Correlation Studies

In this section, I review the existing methodology and current challenges that exist in the cross-multimodal correlations domain. Humans while interacting with the surroundings around them sense many unimodal signals through their sensory organs at any given time. The human mind is capable enough to distinguish between each of these unimodal entities and combine them to form inferences fairly easily and quickly. For example, watching a horror movie on your phone compared to watching the same movie in a movie theatre. The visual cues as well as audio combines together to inform you about the surrounding probably more in the movie theatre giving you a better immersive experience. Those primarily arise due to the conditions, dark-lit rooms, better quality audio and even 3-Dimensional (3D) motion pictures. What happens is the brain perceives this joint representation of two or more modalities (Multimodal representation task) to inform you about the situation which will help you experience the movie better compared to watching it on a phone. Hence, to understand the world better and build systems that can model the world similarly to humans, it becomes necessary to find the correlations existing in the different unimodal stimuli and then integrate those features in the application areas.

Correlation research in psychology and cognitive studies is an area of research which is often non-explorative in nature and the conventional methods measure two variables and assess the statistical significance between them. No effort is given to control extraneous variables affecting the measured variables, hence sometimes leading to misinformed correlations. Conventional methods have been adopted where subjects experience a stimulus and chose from a set of responses. These approaches have been predominantly qualitative in nature and many experiments of this kind have been conducted in the past. [15] investigates spatial localization of audio-visual stimuli and finds out that the vision dominates in a scenario where the visual localisation of stimuli is good. For visuals which are blurred, sound localisation dominates. "Precision of bimodal localization is usually better than either the visual or the auditory unimodal presentation [15]." This kind of study is qualitative in nature and users are shown different instances of the task and they were required to localize in space light "blobs"

or sound "clicks". Early works of [16] distinguish between the different ways we can interpret the interaction between sources of sensory information. They divide these interactions into "sensory combination" (interaction between non-redundant signals) and "sensory integration"(interaction between redundant signals). The idea of multi-sensory perception is supported by the use of the sensory combination, when a single modality is not enough to come up with an estimate of a given surrounding/event, information from multiple modalities can be taken into consideration to remove "disambiguation" and bring clarity to the situation. But when there is more than one estimate present during the task, then the perceived estimate most likely to be chosen by the human is an integration of more than one estimate. For example, suppose you are baking a cake and you want to know whether it has been cooked well or not, you can take visual cues like the colour and size of the cake to check whether it's ready (sensory combination). But very often, what you do would be to check the cake by touching it with a knife to see if the core is cooked properly and then integrate the prior information to come to an estimate of the "cooked" state of the cake. [16] also proposes a "Perception-action loop" (figure 1) which shows the reconstruction of the environment from the sensory data. One of the key challenges of this study was to find a method to solve the "correspondence problem" for sensory integration, i.e., how to systematically study and find suitable correspondence emerging from different sensory integration tasks or simple, continuous multimodal data in the temporal domain.

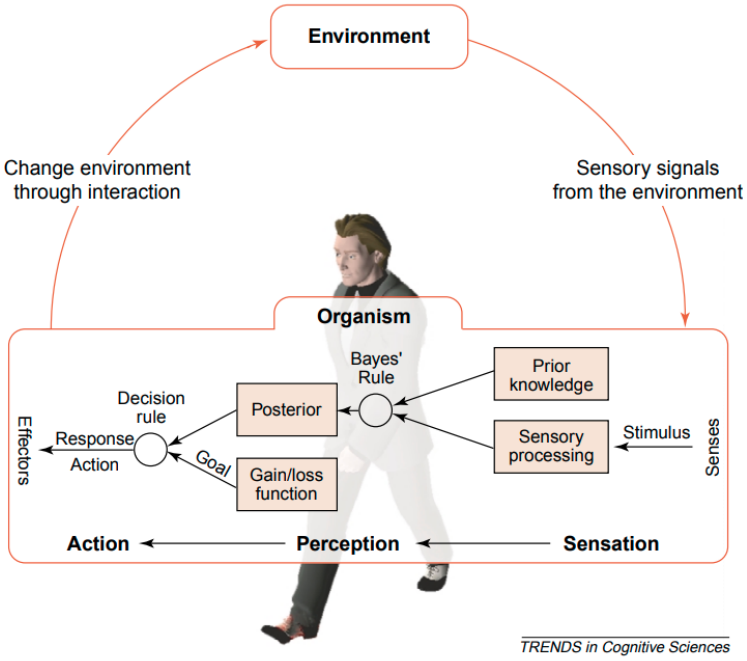


Figure 1: Perception-Action loop model, the dependence of actions on prior knowledge and multimodal sensory processing [16]

Another popular correspondence study was conducted by [8] where Spanish-speaking participants were shown forms depicted in figure 2 and participants chose which shape corresponds to one of the two sounds "takete" and "baluba". There was

a strong correspondence between the jagged shape with the sound "takete" and the globular shape with "baluba". The experiments were redesigned and conducted using similar sounds "bouba" and "kiki" in 2001 [17] using American college undergrad students and Indian-Tamil speakers. Interestingly, over 95 per cent of participants chose the round shape as "bouba" and the pointed shape as "kiki". This qualitative study establishes the fact that human brains are wired in a similar way which attaches abstract meanings to shapes and sounds consistently across different populations of different regions, languages and cultures across the world. This finding also establishes the importance of finding out structured, systematic methods of analysis of multimodal data which can lead to a better understanding of human perception in general. A lot of similar experiments have been conducted by researchers where the users are generally asked to match one modality to another, for example, a user could be asked what kind of sound can correspond to a particular colour. Sound and colour correlations have been studied extensively using this method of correspondence matching. [18] conducted an experiment with school children who were asked to choose a colour that comes to their mind immediately after listening to 6 pure tone frequencies. It was found that blue and violet were selected for stimuli being a lower frequency sound, red and orange corresponding to middle-pitched tones and yellow and green corresponding to higher frequencies of sound stimuli. Similarly, [19] conducted an experiment finding out that brightness and loudness were cross-modally correlated. Even more challenging associations were studied, for example, music with pictures was studied by [20] where the users had to listen to a musical selection on a phonograph and chose coloured reproductions of paintings.

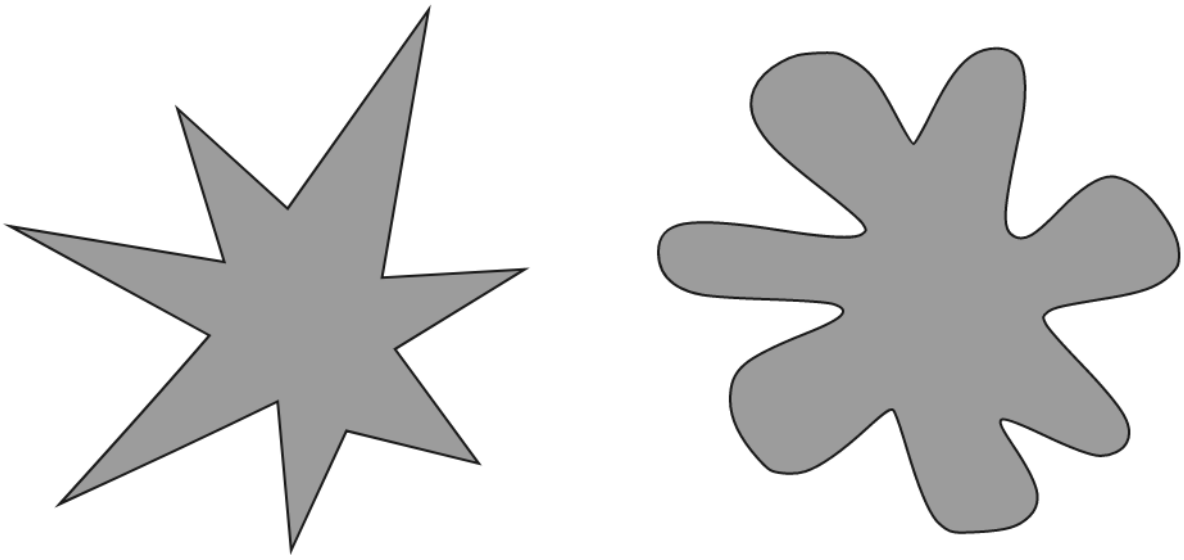


Figure 2: Baluba-Takete effect, the globular shape corresponds to Baluba and the pointed shape corresponds to Takete[8]

Depending on the study objective and the kinds of data being examined, many methodologies can be utilized in sensory multimodal correlation analysis. Here are a few typical approaches:

1. Psycho-physical investigations: In psycho-physical studies, stimuli that change along several sensory dimensions are presented to participants, and they are asked to rate or compare the sensations. Participants might be asked, for instance, to compare the brightness and hue of various colours or to score the sweetness and sourness of various food items. Quantitative measurements of the interaction between various sensory modalities can be obtained through psycho-physical investigations. For instance, scientists have employed psychophysical studies to examine how the interaction of visual and auditory cues in speech perception. [21]

    (a) Threshold detection: It is an experiment where participants are asked to report when they can just barely detect a stimulus (such as a sound or a light) when it is shown to them at different intensities. A subject might be asked to indicate when they can just barely hear a sound, for instance, after hearing it at various volumes. [22]

    (b) Discrimination exercises: Ask participants to identify which of two stimuli is distinct by providing them with two similar stimuli that differ in some way, such as two tones with slightly different frequencies. A person might be asked to choose which of two noises, which have slightly differing frequencies, is higher.[22]

    (c) Scaling experiments: Using a subjective scale, such as a numerical rating or a visual analogue scale, participants are asked to rate the intensity of a stimulus using a scaling approach. Participants might be asked to rate a sound's volume on a scale from 1 to 10, for instance. [22]

    (d) Adaptation experiments: In adaptation experiments, subjects are exposed to a stimulus over an extended period of time, and their sensitivity to the same or a different stimulus is subsequently measured. For instance, after listening to a sound at a certain frequency for a while, participants can be asked to say if a different sound is higher or lower in pitch. [22]

    (e) Cross-modal investigations: In cross-modal experiments, participants are exposed to stimuli from many sensory modalities (such as sound and touch), and the experiments look at how the sensory data is combined. Participants might be asked to indicate whether a sound and a vibration are emanating from the same source after being exposed to both. [22]

2. Brain imaging methods: It is possible to quantify cerebral activity related to several sensory modalities using brain imaging methods like functional Magnetic Resonance Imaging (fMRI) and Electroencephalogram (ECG). For instance, fMRI might be used to examine the brain activity connected to the simultaneous perception of visual and aural inputs. Researchers can pinpoint the parts of the

brain that are engaged in sensory multimodal processing by looking at patterns of neural activity. For instance, research has examined how the brain integrates data from many sensory modalities using brain imaging techniques. [23]

3. Deep neural networks are one example of a machine learning technology that can be used to learn complex correlations between several sensory modalities. To anticipate how a participant would interpret a novel stimulus, for instance, researchers may train a neural network using information from many sensory modalities. For instance, environmental sensor networks can utilize machine learning algorithms to analyze sensor data.[23] These techniques could however be complex and might work well only in a well-defined task.

It is to be noted that several other similar experiments have found cross-modal similarities and correspondences in different modalities other than sound and colour, such as vision and touch, colour and odour, pitch and smell etc [4]. Although, in the above experiments, multimodal correlations have emerged, the methods followed in almost all the experiments have been consistently qualitative and the correspondence has been always found on the basis of the choices that the human participant thinks. Certainly, these have proven to be the existential foundation of multimodal correlation studies however, a lack of quantitative approach limits this experimental methodology to test only a few number of participants. There are many challenges that can possibly lead to a miscalculated correlation and possibly an effect can emerge which can hardly be generalised. For instance, in [20], the number of musically trained participants and the number of non-musical participants were 18 and 15 respectively. Music is diverse and people with different musical backgrounds could possibly react to different stimuli differently. Perception in general, is dependent on the prior knowledge of the individual, see figure 1, hence, finding correspondences only based on the user's thought process could lead to a false correspondence. Furthermore, it is difficult to find participants who are able to express themselves using different modalities like singing, sketching etc. which makes the process of designing and developing a framework for systematic verification of the correlations subjective and qualitative. I believe that so far the experiments performed to find multimodal correlations have been bereft of the technical advancements in the field of other areas of research like computer vision, computational creativity, HCI, machine and deep learning etc. There is a certain need to shift the analysis methodology from a qualitative perspective of experimental studies to a quantitative-based objective study leveraging the advancements in other disciplines of science and technology. The qualitative aspects need not be completely ignored but rather used as a validation tool for verifying the findings from the quantitative study. Therefore, I would like to propose a data-driven framework for finding out multimodal correlations quantitatively and then, the traditional methods of experimental psychology research can be used to validate the findings, making the results more reliable for further studies to come. These findings would further support the idea of multi-sensory integration in terms of Bayesian integration theory. The key concept is that humans may integrate inputs in a statistically optimal way by integrating prior knowledge and sensory information and ranking each by its relative reliability. It seems only natural,

then, to investigate how the concept of cross-modal correspondences may be portrayed as a sort of prior knowledge within such a framework [4]. Further work by Marc Ernst and H.B. Helbig [24] shows that our sensory system's prior knowledge of what is cross-modally correlated can determine the strength of cross-modal coupling. Thus, if a computational model can suitably understand the cross-modal correlations of the given multimodal data, it can leverage this knowledge to take better estimates in a certain situation using the stronger cross-modal coupling in the case of a multi-sensory integration task.

4.2 Multimodality and Machine Learning

With the innovations and advancements in the field of sensor technologies, it is now easier to collect a different variety of data at an unprecedented scale. This variety of data types creates a problem of complexity in understanding the correlation between the different data entities due to the heterogeneous nature of the data. The correspondences between different modalities are essential to understand any natural phenomenon and leveraging this knowledge to be used in other applications. Baltrusaitis et al. have come up with broad categories of challenges we need to tackle in order to experience the full potential of the multimodal setting [3]. The categories are:

1. Representation: The process of representation and summarizing of multimodal data to exploit the complementary and redundancy of multiple modalities.

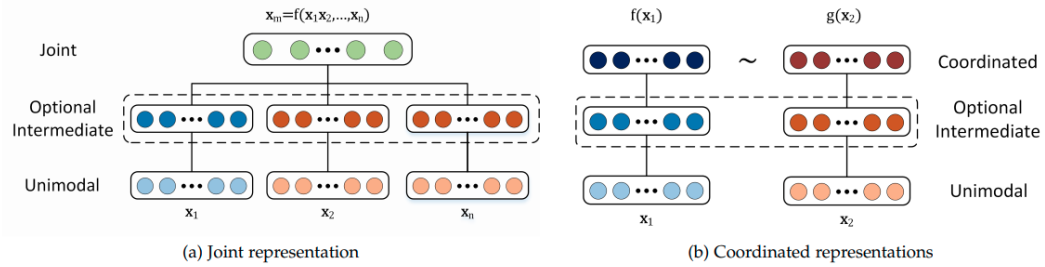
2. Alignment: Processing different types of continuous or discrete data types to measure similarity between them and deal with probable long-range dependencies and ambiguities.

3. Translation: The process of mapping heterogeneous data from one modality to another where the relationship between modalities is often open-ended, subjective and determined by qualitative studies. For example, the correlation between audio and image data.

4. Fusion: Join information from two or more modalities to perform a prediction task. Here the key challenge is that information comes from different modalities and possibly has varying predictive power and noise topology, with missing data in one or more modalities.

In the context of this thesis, the two categories of prime importance are multimodal representation and translation for the following reasons, 1) for finding cross-modal correspondences it is essential to represent the heterogeneous data of different modalities in a comparable space 2) for performing any kind of prediction task, the information from multiple modalities might be integrated to reach to a decision. Advancements in deep learning have shown many advancements in a lot of similar translation tasks and the methods used can be studied further to bridge the gap between machine and human perception. The following sections explain the state-of-the-art techniques and methods that have shown significant results in the representation and translation of multimodal data.

4.2.1 Methods of Multimodal Representation



(a) Joint representation (b) Coordinated representations

Figure 3: Comparison of Joint Representation v/s Coordinated representation of multimodal data[3]

Baltrusaitis et al. broadly categorise representations of multimodal data into 3 kinds: joint, sequential and coordinated representations [3]. Joint representations project unimodal representations together into a multimodal Space. It is preferred to be used when there is a presence of multimodal data types both during training the deep learning models as well as finding inferences based on the predictions. Neural networks are often considered a suitable choice for unimodal data representation [25]. They are used to represent visual, acoustic, and textual data, and are increasingly used in the multimodal domain [26, 27, 28]. This is suitable because of the very way the neural network architectures are constructed, in a layered form. Figure 3 shows the joint representation of multimodal data in a joint space, where each modality has its own network architecture and the outputs of the different modalities concatenate to form a joint representation of multimodal data. Each NN model creates its own abstraction of the unimodal data type it is trained on, concatenating to form an induced joint representation. On the other hand, if the nature of the data types is continuous such as audio, videos etc then sequential data representation is preferred. Different variants of Recurrent Neural Network (RNN), such as Long-Short Term Memory (LSTM) networks [29], have shown tremendous results in sequence modelling of various tasks [30]. With the increase in complexity of each individual modality, it becomes quintessential to learn separate representations of each modality and coordinate them through a constraint to represent in a single space, that method is known as coordinated representation. They enforce similarity between representations, moving onto coordinated representations that enforce more structure on the resulting space [3]. For example, A method to learn such common binary space between sentence descriptions and corresponding images using end-to-end trainable deep learning techniques was proposed by Li and Yang [31]. One of the advanced methods of representation of coordinated spaces is Canonical Correlation Analysis (CCA) [32]. A linear projection which maximizes the correlation between two random modalities can be computed and orthogonality is enforced in the new space. CCA models have shown good results for cross-modal retrieval [33, 34, 35] and audiovisual signal analysis [36, 37]. Hardoon et al. [33] use CCA to learn a semantic relationship between website screenshots and text associated to them. [35] investigates the benefits

of explicitly modelling correlations between two modalities and modelling is effective in features with a higher level of abstraction. There are also several variants of CCA which can be effective in specific scenarios of multimodal data.

4.2.2 Methods of Multimodal Translation

A big part of multimodal machine learning is concerned with translating (mapping) from one modality to another. Given an entity in one modality, the task is to generate the the same entity in a different modality. The multimodal translation is a long-studied problem, with early work in speech synthesis [38], visual speech generation [39] video description [40], and cross-modal retrieval [35]. A particularly popular problem is visual scene description, also known as image [41] and video captioning [41], which acts as a great test bed for a number of computer vision and NLP problems. Researchers have devised numerous methods to leverage the use of deep learning to map one modality to another, for example, speech synthesis [38], visual speech generation [39], video description [40], cross-modal retrieval [35] and image/video captioning [41] which has created a new field of machine learning application with concepts from computer vision and Natural Language Processing (NLP). Baltrusaitis et al. [3] categorise these methods and applications into two types, example-based and generative-based. The former uses a dictionary for translation of multimodal entities whereas the latter has a model which has learnt to generate the translation.

Example-based models can be simply understood using similarity-based data retrieval, where the task is to find the closest sample from the dictionary and use that as a result. Ordonez et al. [42] used unimodal retrieval to generate image descriptions by using extracted global image features to retrieve captions for candidates. Hodosh et al. [43] use a multimodal K-Means Canonical Correlation Analysis (KCCA) space for image-sentence retrieval. Instead of aligning images and sentences globally in a common space, Karpathy et al. [44] propose a multimodal similarity metric that internally aligns the image fragments (visual objects) together with sentence fragments (dependency tree relations). However, similarity in unimodal space does not always imply a good translation and these similarity-based and distance metrics-based retrieval strategies in the unimodal spaces fail in complex situations. One approach to solve this problem is to find the semantic spaces of each unimodal data type which are more meaningful to retrieve from and similarly, it can be extended for each of the modalities. Furthermore, they allow for bi-directional translation, which is not straightforward with unimodal methods. However, they require manual construction or learning of such a semantic space, which often relies on the existence of large training dictionaries (datasets of paired samples). Another possible disadvantage of the dictionary-based approach is that the models become large and inference time becomes more, hence making the predictions of new samples slower. [3]

Generative approaches to multimodal translation construct models that can perform multimodal translation given a unimodal source instance. It is a challenging problem as it requires the ability to both understand the source modality and generate the target sequence or signal. As discussed in the following section, this also makes such methods much more difficult to evaluate, due to the large space of possible

correct answers. The key idea behind the generative approach of multimodal machine translation is to generate a translation from one or more unimodal source modalities to a target modality. The prime focus of this field has been focused on three modalities: language, vision and sound. There have been historical approaches to this problem [38] and modern approaches like [45, 46] as well. Encoder-decoder models have gained popularity and have been used to generate text, images [47, 48], and continuous generation of speech and sound [46]. Popular models to encode acoustic signals include RNNs [49] and Deep Belief Network (DBN) [50]. An encoder model learns the latent representation of each unimodal entity and uses this information to translate to another unimodal space. Decoding is most often performed by an RNN or an LSTM using the encoded representation as the initial hidden state [51, 47]. A number of extensions have been proposed to traditional LSTM models to aid in the task of translation. Rohrbach et al. [52] explore the use of various LSTM architectures (single layer, multi-layer, factored) and a number of training and regularization techniques for the task of video description.

It is important to note here that significant advancements in machine learning and deep learning research have been achieved in several applications to understand different modalities better and even perform prediction fairly accurately. For example, the application of image and video captioning systems [41, 41] can explain a visual scene quite well. This has been possible with the huge amount of data available for images and corresponding efforts in the computer vision fields that use data-driven techniques to collect, process and analyse the large quantity of data. However, there is a lack of quality datasets for other modalities where multimodal data of human expression is encapsulated. This is one of the primary reasons why there has been success in modalities like language(text), audio and video. Another challenge is to design and develop use cases for other modalities which can be tested with human users and have practical applications. I believe the sole reason why human perception is stronger than machines while interacting with the external world is the impeccable understanding of the different modalities (that we interact and respond to) through our multi-sensory channels. Humans outperform robots at perceptual tasks due to their capacity to generalize and detect patterns as a result of experience and learning [53]. Humans, on the other hand, may use their intuition and ingenuity to solve issues that robots cannot [53]. Machines, on the other hand, outperform humans in jobs requiring speed, accuracy, and consistency. They can also process enormous volumes of data quickly and reliably, making them valuable for data analysis and picture recognition [54]. The deep learning models can learn from data only when there exist patterns or relationships in training data. This currently limits the understanding of machines and the only way it could be further improved is by finding better relationships between the various modalities that machines can interact with.

4.3 Theories relating to Qualia, Consciousness and Multimodality

Modern theories on emotions, especially those used in affective computing, quite often evade the most essential question regarding emotional experience: the question of qualia. What is it like to be having a certain state? [55]. Understanding someone's emotions requires an ability to attend to their bodily state. Although there have been great advancements in machine attention, the presence of consciousness and qualia in machines is often ignored or almost absent and most debated. I believe the concepts of qualia and consciousness would play a vital role to study effectively the similarity between machine attention and human attention and could play a greater role in building more interpretable deep learning models in the future but a series of data-driven experiments needs to be planned and studies further to find interesting results in this field.

Daniel Dennet's Multiple Draft Theory provides an interesting approach to seeing how colours are perceived differently by humans and what makes 'qualia' an important topic of discussion in colour theory. He explains the importance of colour coding with a very simple example, "Consider the insects. Their colour vision co-evolved with the colours of the plants they pollinated, a good trick of design that benefited both. Without the colour-coding of the flowers, the colour vision of the insects would not have evolved, and vice versa. So the principle of colour-coding is the basis of colour vision in insects, not just a recent invention of one clever species of mammal."According to philosopher Jonathan Bennett (1965), the substance phenol-thio-urea, tastes bitter to one-quarter of the human population and is utterly tasteless to the rest. Which way it tastes to you is genetically determined. This is another example cited by Dennet in this argument that there are numerous modalities which can have different impacts on the person who perceives it. [14] argues about the importance of qualia having Dennet's theory as a basis. Dennet organizes his ideas into five intuition pumps that reflect the fundamental statements about qualia made by previous thinkers. Dennett's purpose is "to destroy our faith in the pretheoretical or "intuitive" concept" (1988), but he also manages to build a basis for qualia research in the process [14]. [14] describes the several criteria for qualia based on Dennet's arguments, 1) it should be possible to isolate qualia from any other things in the surroundings; 2) qualia must refer to the characteristic properties and features from the physical world; 3) the problem of studying and understanding qualia shall be done systematically; and 4) the presence of possible inter-subjective comparisons between different qualia. This work also introduces the idea of the relationship between human/machine perception, attention and qualia. Hence, the presence of qualia and the impacts on human beings might be well established but how can we extend this idea to deep learning systems is an interesting avenue of research.

Giulio Tononi in 2004 proposed a theory to explain the nature and source of consciousness. It claims that consciousness shall be considered as any other kind of information, and that can be measured mathematically.Integrated Information Theory (IIT) takes a neuro-scientific approach and suggests taking neuro-scientific descriptions of the brain as a starting point for understanding what must be true of a

physical system in order for it to be conscious. This theory states that consciousness requires a grouping of elements within any computational or neural system that have physical cause-effect power upon one another and holds the opinion that only those systems will a feedback loop (similar to recurrent neural networks) can be considered conscious. This theory advocates the usability of any such system with a feedback model to achieve machine consciousness and thus, generalizes its claims beyond human consciousness to animal and artificial consciousness.

Researchers have also advocated the importance of qualia in the realisation of machine consciousness. According to Haikonen [56], qualia are the primary ways in which the human sensory system responds to sensed stimuli. They are not mere properties of the physical world but rather a primary way in which sensory information is manifested inside a human mind. “To be conscious in the way that we experience it is to have qualia. True conscious machines must have qualia, but the qualities of machine qualia need not be similar to the qualities of human qualia”. Haikonen believes Qualia are private and subjective to a given individual and “there is no known objective method of detecting and transmitting the exact qualia to another person”. In the case of a successfully anaesthetic body, there is an absence of the qualia/experience of pain, hence we can state that there is no consciousness without qualia. However, Haikonen raises the question of ‘Can there be conscious machines without qualia?’ if not, then how can we detect qualia in machines or how can we design machines which experience (or have qualia)?

If qualia were effects created by the nervous system, then artificial excitation of sensory nerves should produce qualia. This opens up the prospect of inciting qualia in artificial systems as well. Loizou seems to show that this is indeed the case. In an experiment where a Cochlear implant was planted in a person’s ears with a hearing disability, the presence of an artificial hearing aid resulted in some kind of auditory enhancements in the person. [57] This demonstrates that both natural and artificial stimulation of sensory nerves leads to the sense of qualia, and so qualia arise within the brain. The quality of the evoked qualia depended upon the identity of the nerves chosen; thus giving rise to the generation of auditory qualia. An interesting thing to note here is that the nerve fibres responsible for the transmission of data to the brain are not labelled, so how does the brain perceive which qualia to evoke in a particular case? This suggests that signal orientation does not matter whereas the quality of the evoked qualia is somehow determined by the target area in the brain where the signals are received. These ideas and studies have compelled researchers to explore the idea of conscious machines. One such work is [58] which argues that "... Daniel Dennet and many others have argued that in fact there is no Hard Problem and that what we perceive as consciousness is just an illusion like many others". It takes "experiencing illusion" as a tool to verify whether machines can experience the sense of illusion just like humans do and if they do, can they be considered partially or completely conscious? This is similar to solving Completely Automated Public Turing test to tell Computers and Humans Apart (CAPTCHA) like puzzles on several login pages on websites where CAPTCHA is used to verify whether the user is an actual human user or a bot crawler. The idea is simple, show such examples of text which can be predicted by humans with ease but that computer vision models fail to recognize.

Figure 4 shows a set of 3 illusion-based tests presented in the study to determine the consciousness of an artificial agent/computation model. This work advocates that "experiencing something allows one to obtain knowledge about that experience, which is not available to someone not experiencing the same qualia. " Building upon this idea if two different agents (artificial and natural) experience the same thing using their multi-sensory receptors (sensory organs in case of humans and electronic sensors in case of machines), we can say that even the artificial agent has been trained to experience a certain state of being (qualia) just like the human counterpart.

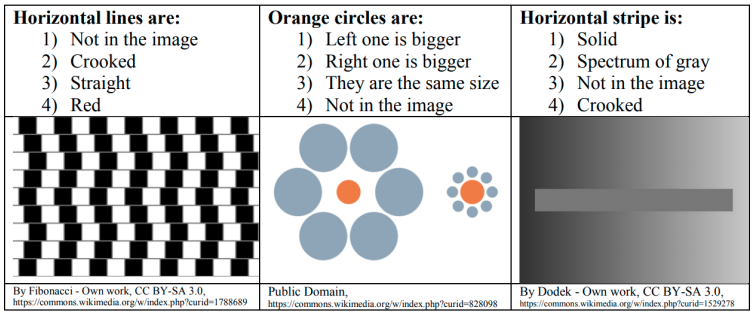


Figure 4: Tests to test consciousness of artificial agents [58]

The arguments of Daniel Dennets's theory although undermines the importance of qualia due to its subjective nature, the arguments establish one of the key foundations of how qualia needs to be studied. The recent work of Giulio Toroni [59] advocates that any computational system might be considered conscious IIT the system has a feedback loop. This opens up the proposition that deep learning architectures like Recurrent Neural Networks (RNN) can be considered conscious. Furthermore, Haikonen [56] believes that there is no consciousness without the presence of qualia. In a conscious system, there exists a presence of qualia so how can it realise in these systems?, how to find the correlation in multimodal datatypes? and how can we achieve human-like attention in computational systems? - I believe these questions will be purposeful in the realisation of better systems in the future of multimodal machine learning, consciousness and qualia research.

Neither providing a detailed explanation nor experimentation of the aforementioned theories/concepts are in the scope of this thesis. It is also not entirely sure to state that finding multimodal correlations can help in finding answers to the hard questions of consciousness and qualia. However, it is important to see that artistic creation is subjective in nature and it is equally important to explore these concepts of consciousness while designing future systems aimed at the co-creation of art by AI and humans. Training and implementing AI systems for specific use cases on multimodal data collected from human users and knowing the relationships between the different multimodal data types might be useful in building better performant and interpretable systems in the future.

4.4 Exploring Attention in Machines

The use of attention models in deep learning architectures has proven to be an integral part of machine translation, computer vision and Natural Language Processing(NLP). The concept of machine attention, focusing on certain input vectors from the sequential data while training a recurrent neural network, has worked well significantly in the aforementioned application domains. It was first proposed in [30] in a machine translation task. Encoder-Decoder networks have been predominantly used for this kind of task, where the encoder first encodes the source sentence into a fixed-length vector and then the decoder performs the translation from that vector. However, [30] shows that the performance of translation could improve significantly if another model is used to search for relevant sections of the sequential data to find the closest translation. As a test case, they considered an English-to-French task and the results with the attention mechanism model outperformed the existing encoder-decoder architecture significantly. Figure 5 shows one such translation from an English sentence to French, the weights shown in the figure (white colour means higher attention) are a visualization of how the attention mechanism works. The main difference between this approach and the basic encoder-decoder is that it does not attempt to compress a whole input sentence into a single fixed-length vector. Instead, it encodes the input sentence into a series of vectors and adaptively selects a subset of these vectors while decoding the translation [30].

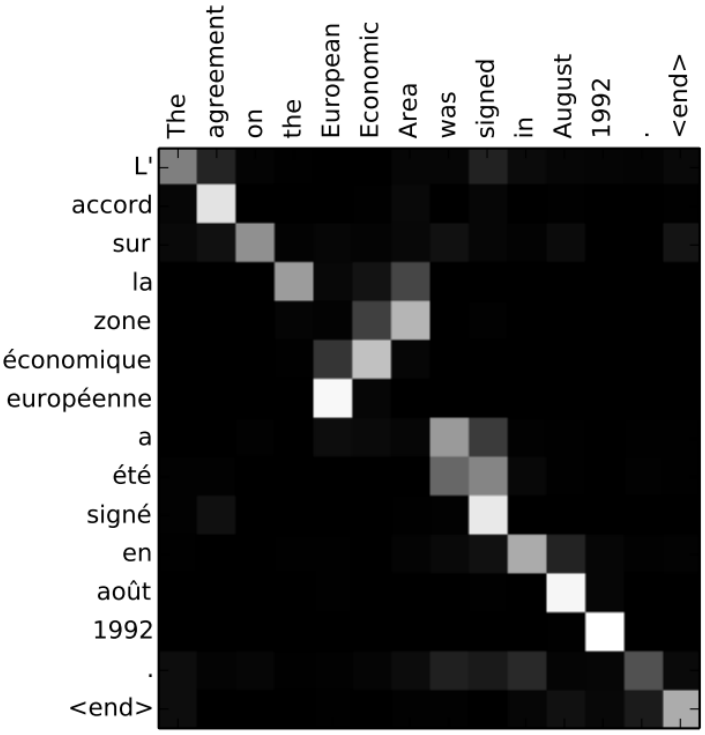


Figure 5: Visualization of attention weights in an English-to-French language translation task [12]

The methodology followed here is similar to what happens in human perception of surroundings. When humans interact with a surrounding, it tends to focus on things which are relevant or important in the context. The concept of attention mechanisms also helps to find interpretability in deep learning architectures. The nature of deep learning models doesn't clearly state why the model predicts something in a certain way and this methodology of looking at specific locations of the input data might help in inferring where the model actually looks! Similar work on the localization of features in input data has been performed by Zhou et.al. [12] where they localize important sections in an image in a classification task. Figure 8 shows the areas which have been highlighted using a class activation function, determining the most salient section of a given image. How this is implemented is that the global average pooling layer (to minimize over-fitting by reducing the total number of parameters in the model) is modified with the Class Activation Mapping (CAM) technique[12], thus allowing the Convolutional Neural Network (CNN) classifier model to both classify the image and localize class-specific image regions in a single forward-pass e.g., the toothbrush for brushing teeth and the chainsaw for cutting trees.

Figure 6: Localization of specific regions in an image classification task [12]

Finding multimodal correlations in data might assist AI attention models in better interpreting and integrating information from different modalities, resulting in enhanced performance, resilience, and accuracy in various AI applications. Attention models in AI are intended to selectively focus on relevant features or information, hence improving the performance of various AI applications such as image identification, audio recognition, and natural language processing. Attention models can better capture the relationship between different modalities and enhance their accuracy by discovering multimodal correlations in the data. They can also combine data from several modalities (such as text, photos, and audio) to complete jobs more correctly. Multimodal correlation analysis can aid in the identification of common features across different data variations and in the development of attention models generalizing to new data.

5 Preliminary Evaluator for Multimodal Correlations (PEMC)

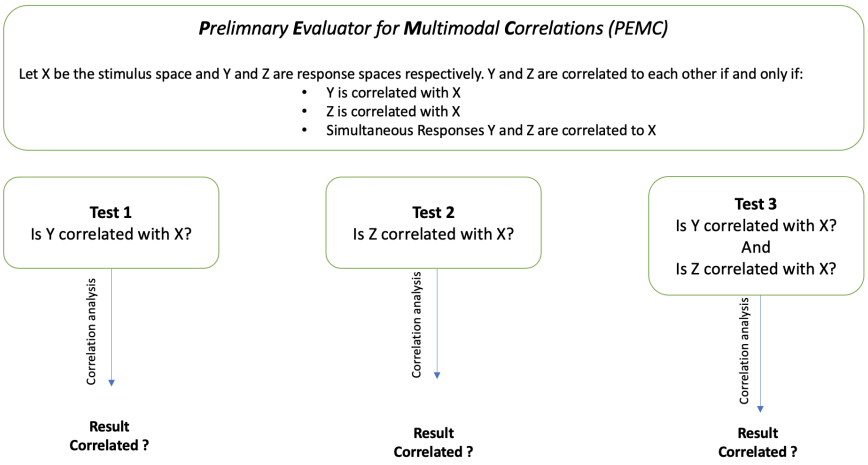


Figure 7: PEMC: Pictorial representation of the framework for data-driven multimodal correspondence with humans in the loop evaluation

This section explains the PEMC framework (figure 7) which is proposed in this thesis to find multimodal correlations. [60] discusses the transitive nature of correlations in general and if it is possible to establish correlation amongst 2 or more variables. Let X, Y and Z be three different modalities(feature variables for respective modalities are also valid, for example, Pitch, Power, and Mel-frequency Cepstral Coefficients (MFCC) of the audio signal) that we would like to measure the correlation between. According to [60], If X is correlated to Y, and X is correlated to Z then Y tends to be correlated to Z IFF the correlation coefficients of XY and XZ are close to 1. Based on this mathematical relation of transitivity, it makes it difficult to establish a correlation between the 3 or more data types as it is mathematically highly unlikely to get the correlation coefficients nearly equal to 1. To solve this problem, I propose PEMC which consists of 3 fundamental tests and the modalities are correlated only if all the 3 tests are satisfied.

> Let X be the stimulus space and Y and Z are response spaces respectively. Y and Z are correlated to each other IFF the following three tests are satisfied:
>
> 1. Test 1: Y is correlated with X, This means that as X changes, Y also changes in a systematic way.
> 2. Test 2: Z is correlated with X, this suggests that as X changes, Z also changes in a systematic way.

3. Test 3: Simultaneous Responses Y and Z are correlated to X, This means that the joint response of Y and Z together provides correspondence information about X.

Hypothesis More correlations will emerge from the multimodal data in test setting 1 and 2 and only a few correlations will emerge in test setting 3 which will serve as a guide for planning and designing more controlled experiments to test universal correlations in the data.

If all three tests are satisfied, it is proposed that Y and Z are correlated with each other. In other words, there is a relationship between the responses in Y and Z, and they provide consistent information about stimulus X in stimulus space. It's important to note that this theory is based on the assumption of correlation, which implies a statistical relationship between variables. Correlation does not necessarily imply causation and other factors or underlying mechanisms may be at play. It is to be noted that PEMC gives preliminary suggestions that correlations exist in the Y and Z features, which can be studied with more controlled experiments. The benefit of this approach is that more modalities can be tested easily in an uncontrolled experiment and future controlled experiments could be planned depending on the presence of emerging multimodal correlations. Further research and empirical evidence would be needed to validate or refute this theory in specific contexts or domains which go beyond the scope of this study. In this thesis, we focus on performing the preliminary experiments to see if correlations emerge in 2 or more modalities according to the framework.

The 3 tests for PEMC that are conducted to test the correlations present in the modalities are carried out with the CA process, a bimodal correlation analysis method. Figure 8 shows the workflow of the CA. Its purpose is to find the correlation between features extracted from the two modality types. The technical basis for CA is a combination of different techniques from statistics and machine learning; feature extraction, dimensionality reduction, unsupervised machine clustering and correlation analysis of the extracted features as shown in figure 8. The different techniques are explained in detail in the following sub-sections.

5.1 Correlation Analyser

Correlation Analyser is an integral part of PEMC. The tests that are fundamental to investigate the existence of any multimodal correlation according to PEMC are carried out using Correlation Analyser. Theoretically, it is a combination of several steps which can work with any sensory data type. The steps begin with a preliminary preprocessing step where feature extraction of raw data is done. This step is followed by a dimensionality reduction step since most of the multimodal data are highly dimensional in nature. After reducing the dimensionality of data, unsupervised clustering of data is created to understand any patterns or trends in the latent space of the features of the response data. Finally, a Correlation Analysis is conducted using Spearman's or Pearson's methods and concludes whether certain features of the

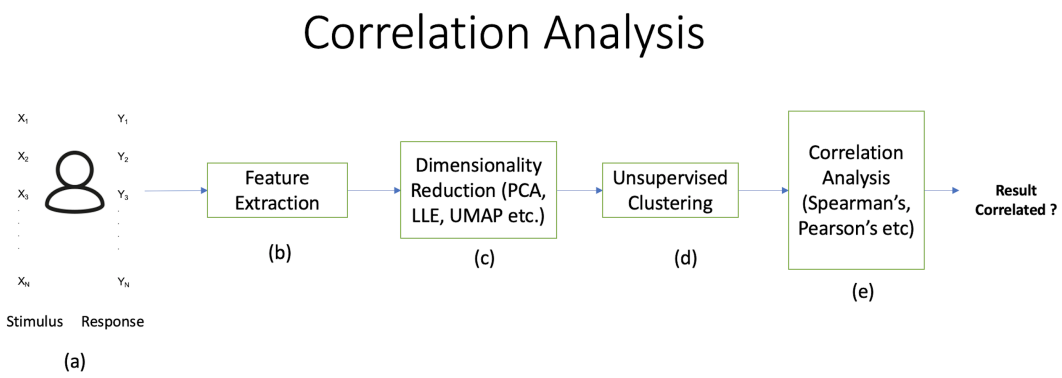


Figure 8: Correlation Analyser: Integral component of PEMC

response are correlated with the stimulus or not. The steps are modular in nature and more advanced dimensionality reduction, clustering or correlation techniques could also be used to analyse the data. The following subsections explain each of the steps in detail.

5.1.1 Handling Multimodal Datatypes

Multimodal data types could be in many forms, for example, audio, video, images, gestures and body movements, drawing etc. Fortunately, with the advancements in technology, today we have quality equipment and devices to record and collect the data and convert it into formats which are accessible. Considering most of the data types are dynamic in nature, There are different open-source libraries and tools which can process them into time-series data. The only thing that needs to be taken care of while preparing the data is that the multimodal responses to a given stimuli must be synchronised properly, hence utmost care needs to be taken while deciding the data collection techniques. Recording techniques which involve human intervention multiple times and manual recording of responses of different types simultaneously need to be avoided.

5.1.2 Feature Extraction

The CA method begins with the feature extraction of important features from the input data. The input to the CA is the feature vector consisting of the response vector Y to a given stimulus X, where $X = \{x_1, x_2, ..., x_n\}$, $Y = \{y_1, y_2, ..., y_n\}$ and n is the total number of samples in the stimulus and response vector set. Now depending on the type of data, different features could be extracted. The feature types are decided by the researcher and the number of feature types doesn't affect the use of the framework.

CA is highly modular and hence, different features of data Y and Z could be calculated and analysed together. It is recommended to extract all possible kinds of features for a data type (Audio, video, gestural data etc.) because we don't want to bias what kind of correlations we already expect from the data. We rather chose

more of an exploratory data analysis technique. In multimodality research, most of the experiments are designed in a way that only deals with certain features of the data and the researchers have a predetermined notion of correlations amongst them, hence limiting the possible outcome of the experiment. Collecting and creating multimodal datasets is a challenging task as that requires the involvement of experienced participants for certain tasks and hence, making the most out of the collected raw data would help us explore more universal correlation in an efficient way.

For example, if one of the responses would be audio data, then there exist different techniques to extract the audio features from the audio file. For instance, Librosa is a Python package for music and audio analysis. It provides the building blocks necessary to create music information retrieval systems. Different types of features could be extracted such as spectral features, rhythm features etc.

5.1.3 Dimensionality Reduction Techniques

Dimensionality reduction is the transformation of data from a high-dimensional space into a low-dimensional space so that the low-dimensional representation retains some meaningful properties of the original data, ideally close to its intrinsic dimension. This is an essential component in CA because correlation analysis requires the data to be in the format of two vectors $X = (X1, ..., Xn)$ and $Y = (Y1, ..., Ym)$, where X and Y consist of integral/decimal values. And most of the features that are extracted from multimodal data types are highly dimensional in nature. Hence, different dimensionality reduction techniques are used to find the best possible low-dimensional representation of the data. Given a D-dimensional dataset $X \in R^{NxD}$ with N number of samples, the reduction creates a projection of $Z \in R^{Nxd}$ where d«D such that a significant amount of data X is preserved in the data Z. The latent space of modality features could be seen using a variety of 2-Dimensional (2D) projection using scatter plots. Each of these algorithms generates a unique projection, which results in a unique clustering. Furthermore, given that contemporary projection algorithms are stochastic, it is possible for outcomes to vary between runs using the same set of hyper-parameters.

Principal Component Analysis (PCA) is mostly used in exploratory data analysis and for making predictive models [61]. It is often used to visualize genetic distance and relatedness between populations.PCA uses an orthogonal linear transformation to convert a set of possibly correlated observations into a set of linearly uncorrelated observations called principal components. The number of components is decided by the dimension d that you need to reduce the data to. For example, if d = 2, then the new data project Z ensures that the greatest variance lies on the first axis called PC1 and the second largest variance on the second axis. Basically, PCA produces a point cloud that represents the best linear approximation of the original dataset X. However, most of the datasets of multimodal nature are non-linear in nature and we need other non-linear techniques to reduce dimensionality.

Local Linear Embedding (LLE) is a method of Non-Linear Dimensionality reduction proposed by Sam T. Roweis and Lawrence K. Saul [62]. The LLE algorithm is an unsupervised method for dimensionality reduction. It tries to reduce these

n-Dimensions while trying to preserve the geometric features of the original non-linear feature structure. LLE tries to characterise the local geometry of the data by linear coefficients that reconstruct each data point x ∈ X from its neighbours. Reconstruction errors are then measured by the cost function which adds up the squared distances between all the data points and their reconstructions. Because LLE replaces each feature vector with a linear combination of their nearest vectors, the cloud tends to be dense and the distance ´between each cluster is higher.

Uniform Manifold Approximation and Projection (UMAP) is another non-linear dimensionality reduction technique that uses local manifold approximations to construct a topological representation of the high-dimensional data P and iterates by creating several low-dimensional topological representations Q and selecting the one that minimizes the cross-entropy between both representations [63]. One of the hyper-parameters of UMAP is the desired separation between close points in the embedding space which leads to spreading dissimilar objects out but also creates potentially densely packed regions for similar objects.

Classic Multi-Dimensional Scaling (MDS) is another dimensionality reduction algorithm that is commonly used in statistics, psychology, and computer science [64]. MDS works by computing a pairwise distance matrix between all of the data points in the high-dimensional space. This distance matrix is then used to create a new, lower-dimensional representation of the data using an optimization algorithm such as gradient descent. The optimization algorithm tries to minimize the difference between the pairwise distances in the original high-dimensional space and the pairwise distances in the new lower-dimensional space.

If the dimensionality reduction methods PCA, MDS, LLE, and UMAP perform differently on the same dataset, it indicates that the data contains complex and non-linear relationships between its variables.

PCA and MDS are linear dimensionality reduction methods that work well when the variables in the data have linear relationships. If PCA and MDS produce distinct findings, it could be due to non-linear relationships between variables that linear techniques cannot capture.

LLE and UMAP are non-linear dimensionality reduction methods that can detect non-linear relationships in data. If LLE and UMAP produce different results, it is possible that this is due to the fact that they use separate algorithms to capture these non-linear relationships. In general, the technique used to reduce dimensionality should be determined by the particular characteristics of the data and the objectives of the analysis. To identify the most appropriate approach for the specific problem at hand, it may be necessary to attempt multiple techniques and compare their results.

Relationships within the latent space of data might be inferred from the type of clusters formed from different algorithms. For example, if a dataset works well with the LLE dimensionality reduction technique, this indicates that the data may have non-linear relationships between its variables that LLE can capture. LLE is a non-linear dimensionality reduction method that seeks to preserve the data's local structure or the relationships between nearby data points in a high-dimensional space. It works by first identifying each data point's neighbours and then locating a low-dimensional representation of the data that preserves the same neighbour connections.

Similarly, PCA is a linear dimensionality reduction method that works by identifying the directions of maximum variance in the data and projecting the data along these directions onto a lower-dimensional space. The resulting principal components are linear combinations of the initial variables that capture the data's most significant patterns. If PCA effectively reduces the dimensionality of the data while keeping the majority of the variance, it implies that the data has linear relationships between its variables. This could be the case, for example, if the data represents a system in which various variables are directly proportional to each other or where the variables have a linear relationship.

5.1.4 Clustering

In CA, clustering is performed both on the stimulus feature vector and the response feature vectors. These feature vectors could be high dimensional as well hence, suitable dimensionality reduction techniques are also required if necessary prior to this step (explained in the above subsection). This is especially done to make the computation of the clustering algorithm computationally effective by suppressing the redundant noise in the feature vectors and speeding up our computation.

Clustering algorithms are used to find groups/clusters inside the feature vector dataset. This step of finding clusters is extremely necessary to perform the correlation analysis because the presence of even a small number of dissimilar samples can deteriorate the correlation coefficients and hence, even though there exists any correlation multimodally, the correlation coefficient's value would be closer to 0 (neither positive correlation nor negative). It is to be noted that running the correlation analysis on the whole dataset without clustering yields correlation coefficients nearly equal to 0 in all cases. Thus, it is required to divide the dataset $D = \{d_1, d_2....d_n\}$ of n number of feature vectors into a set $G = \{g_1, g_2, ..., g_k\}$ where 1<k«n.

In CA, the K-Means algorithm has been used to find the clustering of the groups. It is a very computationally efficient and robust algorithm which converges optimally quickly to the local minima. One of the parameters that need to be passed into the algorithm is the value of K. It basically defines the number of groups to divide the whole dataset into. To find the value of K, the elbow curve criterion is used in all the experiments presented in this thesis. The elbow criterion is a popular method for selecting the optimal number of clusters (K) in K-means clustering. It involves plotting the Within-Cluster Sum of Squares (WCSS) against the number of clusters and identifying the "elbow" point in the plot where the rate of decrease in WCSS slows down. This elbow point is often considered the optimal number of clusters and is decided manually by the user for each experiment. One can simply look at the elbow plot and locate the point where the WCSS begins to level off and resemble an "elbow". At this stage, the rate of WCSS decline starts to noticeably slow down. As it shows a fair compromise between WCSS minimization and avoiding overfitting with too many clusters, this point represents the ideal number of clusters. Figure 9 shows a sample elbow curve where the sum of squared distances between the clusters is plotted against the k value. The optimal K value is chosen manually after which there is not much variance in the data, in this particular case the optimal value is K=3.

The corresponding plot of the right-hand side of the figure shows that there are 3 almost clear clusters formed from synthetic data for demonstration purposes (clusters demarcated by the 3 distinct colours).

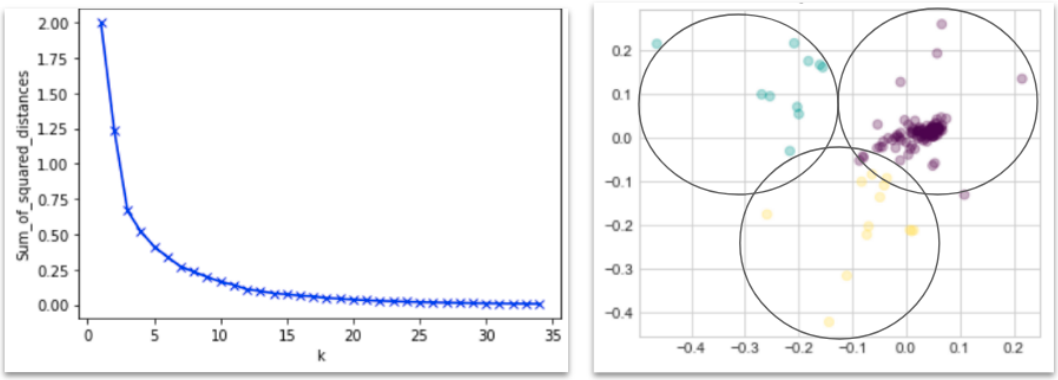


Figure 9: Clustering using the K-means algorithm on a dataset where K=3 yields the best possible results)

5.1.5 Correlation Techniques

In statistics, the coefficient of multiple correlations is a measure of how well a given variable can be predicted using a linear function of a set of other variables. It is the correlation between the variable's values and the best predictions that can be computed linearly from the predictive variables. The coefficient of multiple correlations takes values between 0 and 1; a higher value indicates a high predictability of the dependent variable from the independent variables, with a value of 1 indicating that the predictions are exactly correct and a value of 0 indicating that no linear combination of the independent variables is a better predictor than is the fixed mean of the dependent variable.

There are two popular techniques which are used to perform the correlation analysis, namely, Spearman's test and Pearson's test. If the data are correlated using Spearman's or Pearson's correlation, it means that there is a relationship between the two variables being compared. However, the type of correlation coefficient used can provide information on the nature of the relationship.

Pearson's correlation coefficient measures the linear relationship between two continuous variables. If the Pearson correlation coefficient is close to +1 or -1, it indicates a strong positive or negative linear relationship, respectively. A correlation coefficient close to 0 suggests little to no linear relationship. The Pearson correlation coefficient measures the linear relationship between two datasets. Strictly speaking, Pearson's correlation requires that each dataset be normally distributed. Like other correlation coefficients, this one varies between -1 and +1 with 0 implying no correlation. Correlations of -1 or +1 imply an exact linear relationship. Positive correlations imply that as x increases, so does y. Negative correlations imply that as x increases, y decreases.

On the other hand, Spearman's correlation coefficient measures the monotonic relationship between two variables, which means that it measures how well the relationship can be described by a monotonic function (e.g., a straight line or a curve). Spearman's correlation coefficient is also used to measure the strength and direction of the relationship between two ordinal variables. A Spearman correlation coefficient of +1 or -1 indicates a perfect monotonic relationship, while a coefficient close to 0 suggests little to no monotonic relationship. The Spearman rank-order correlation coefficient is a non-parametric measure of the monotonicity of the relationship between two datasets. Unlike the Pearson correlation, the Spearman correlation does not assume that both datasets are normally distributed. Like other correlation coefficients, this one varies between -1 and +1 with 0 implying no correlation. Correlations of -1 or +1 imply an exact monotonic relationship. Positive correlations imply that as x increases, so does y. Negative correlations imply that as x increases, y decreases.

6 Experiments: Multimodal Correlation Using PEMC framework

The PEMC framework intends to find a multimodal correlation between 2 or more multimodal data vectors. Figure 7 shows an ideal scenario where there are 2 responses (Y, Z) to a stimulus X , where $X = \{x_1, x_2, ..., x_n\}$, $Y = \{y_1, y_2, ..., y_n\}$ and $Z = \{z_1, z_2, ..., z_n\}$ are vector representations of the multimodal data features. n is the total number of samples in the stimulus and response vector set. For example, let's consider the modalities that have been experimented with in this thesis. Here, we have the HSL colour model as the stimuli space and verbal/audio and pen gestures in the response space. After feature extraction, one of the possible representations of the stimulus and response could thus be:

1. X: The change in Hue, Saturation and Lightness/Luminance in the colour space, ΔC
2. Y: Features extracted from the audio response, for example, pitch, power, MFCC coefficients etc.
3. Z: Features extracted from the pen gestures response, for example, average velocity, total distance etc.

6.0.1 Mode of the experiment and choice of the modalities

The number of modalities in this study is limited to 3, to simplify the complexity of the problem and understand the intrinsic correlation of features within each of the modalities. Colour is one of the easily studied modalities which induce behavioural responses in humans, for instance, people react differently to different colours and there are quite a lot of studies showing these trends in the psychology domain **??**. Hence, the stimulus in the experiments conducted with human users is changing colours in the HSL (Hue Saturation Lightness) space. The human users responded to this stimulus in two ways, 1) by expressing themselves vocally and 2) by drawing on a Wacom professional tablet. The user is not restricted to the type of sounds which is expressed or to the shapes that are drawn. The study is kept to be as open as possible to understand whether there exist any universal responses to the given stimulus.

6.0.2 Experimental Setting

Figure 10 shows the settings that were followed while collecting the multimodal data for the experiment. This setting was designed in collaboration with Jaana Okulov's work. A web platform was built to ensure that the different modalities of the collected data are synchronous to lead to better results. Synchronisation of different modalities while collecting the data is essential in this kind of experiment where the responses to the stimulus are sequential time series data.

The idea here was to collect data in three modalities, i.e., the stimuli space: the change in colour in the HSL colour space and response space: audio/verbal response

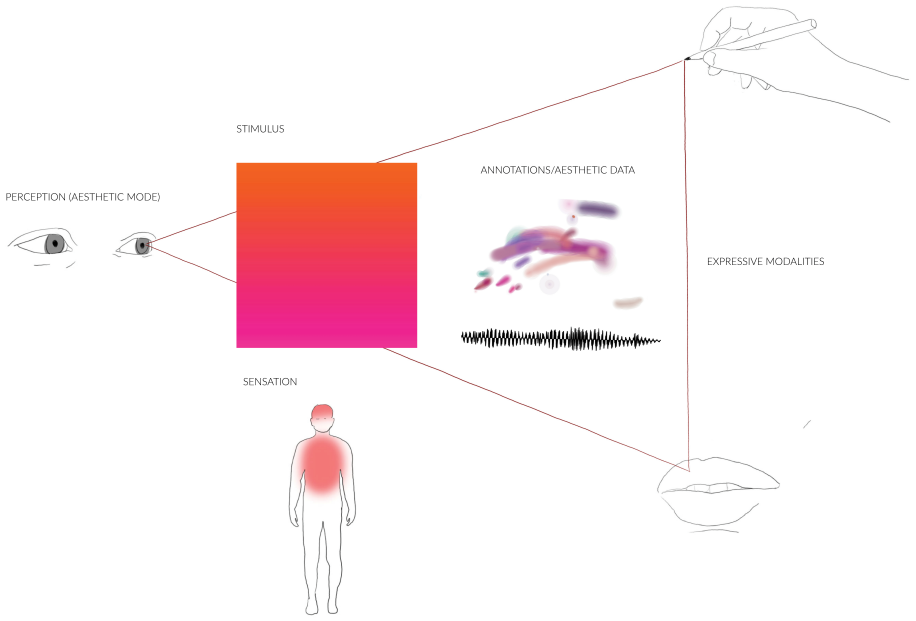


Figure 10: Experimental Settings for collection of data for the experiments, Figure Credits: Jaana Okulov

and pen gestures. The table 1 and table 2 shows the stimuli space and response space respectively.

No.	Stimuli	Description
1	ΔH	The change in Hue space
2	ΔS	The change in Saturation space
3	ΔL	The change in Lightness space

Table 1: Stimuli Space

No.	Response	Description
1	Audio/verbal	Verbal expressions made by human users
2	Pen gestures	The gestures drawn on Wacom tablet

Table 2: Response Space

The choice of participants was primarily people who are comfortable with expressing themselves vocally (some experience in singing) and people from artistic studies who can express themselves by drawing. Also, the users were not instructed on

what kind of sounds (whether songs, tunes, rhythms or just vocals) they shall make or any particular type of drawing they are supposed to do. The instructions were kept as such to find whether there exists any universal form of expression to a given modality.

The sequential data of the three modalities was discretised (into smaller chunks of continuous data) according to the time stamp of each stimulus. Different features from the two modalities (audio and pen gestures) were calculated and compared to the variation of the colours in each of the axis of the three-dimensional HSL space(H, S and L vectors were calculated respectively to account for the change in magnitude and direction along these axes).

6.0.3 Data recorder and dataset

In this study, the three modalities are the variation in HSL colour space (stimulus) and 2 responses, i.e., vocal expressions and gestures drawing using a Wacom tablet. HSL colour space is chosen due to the close resemblance with the colour space interpreted by human perception. It is an alternative to the RGB colour space and aligns with the way human vision perceives the colour-making attributes [65]. Figure 11 shows the values of the HSL colour space, the value of Hue space ranges from 0 to 360 (0 is red, 120 is green, 240 is blue), with other values of Lightness(0 for black and 100 for white) and Saturation (0 means a shade of grey and 100 is the full colour) ranging from 0-100. A web application was created to collect the data systematically and automatically to avoid human errors in synchronisation in the recording of both audio and pen data. The participants were briefed about the data collection methodology and the different settings of the experiments. There were 3 experimental settings:

1. Setting 1: Participant responds with only verbal expressions (audio-only responses).
2. Setting 2: Participant responds with only pen expressions (pen-only responses).
3. Setting 3: Participant responds with both verbal and pen expressions simultaneously (both audio and pen responses).

Figure 11: HSL colour space

The participants were also comfortable expressing with vocal expressions and pen gestures. The quality of the data obtained can be considered of high quality as measures were taken to ensure the synchronous nature of the time series data and experienced users from artistic backgrounds participated in the data collection process.

6.0.4 Feature Extraction

Sl.No	Feature Name	Description
1	Pitch Variation (PV)	Variation of Pitch in a time window
2	PVFT	Pitch Variation Fourier Transform (PVFT)
3	Chroma Energy	Normalised chroma energy of sound wave
4	MFCC	Short-term power spectrum of a sound
5	RMS	Root Mean Square (RMS) value for each frame
6	Spectral centroid	Centroid of spectrum of frequencies
7	Spectral bandwidth	Difference between upper and lower frequencies
8	Spectral contrast	Measure of clarity of the signal

Table 3: Feature set description for Audio Response

Sl.No	Feature Name	Description
1	Total Distance	Distance from the initial point to final point
2	Velocity	Average velocity of the pen movement on the Wacom
3	Peaks	Total number of peaks formed during one colour transition
4	pressure	Average pressure computed during one color transition

Table 4: Features description for Pen gestures response

To analyse and correlate the multiple modalities, all three types of data were processed to find out features for further comparison using the CA. Firstly, the H, S and L vectors were computed for each iteration, which informs about the magnitude and the direction of the change in the H, S and L axes respectively. For the audio data, an open-source audio processing library (Librosa) was used to find the different features explained in table 3. These features collectively explain the verbal expression which was in response to a given colour transition. Similarly, features were also calculated for the pen responses, explained in table 4. In order to find out patterns or trends in each participant's responses in the two response modalities, responses were visualised from their latent representations. This representation of modalities was computed in a data-driven method, automatically for each participant as the patterns in data might be different. Firstly, the dimensionality of the data points was reduced using techniques (UMAP, LLE, MDS, PCA) and then clustered using Kmeans, an unsupervised machine learning technique. K-means clustering is a method of vector quantization, originally from signal processing, that aims to partition n observations into k clusters in which

each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. The number of clusters was calculated automatically for each participant, choosing the value of k from the elbow curves of the modalities (refer to 5.1.4).

6.1 Pre-Study, Experiment 1 (N=2)

This experiment can be considered a pre-study to understand the challenges of the experiment in a tangible way. To test whether the choice of modalities and the PEMC framework works in practice, the sample size (N) of the participants was limited to N = 2, and the participants recruited had at least 3 years of experience in artistic studies. Both participants were from the School of Arts and Design, Aalto University and the data collection process was conducted by Jaana using the web application. 20 samples of responses to the stimuli were recorded for each of the 3 experimental tests(pen-only responses, audio/verbal responses and simultaneous pen and audio responses).

6.1.1 Pre-study Results and Discussions

The following unimodal correlations emerged while using PEMC for finding out the respective correlated feature set :

PEMC Test 1 (Stimulus: Transition in HSL Space (X), Response: Verbal/Audio (Y))

1. Saturation seems to be positively correlated with the pitch and clarity of the audio signals with a strong Spearman's correlation coefficient {0.58, p=0.003} and {0.52, p=0.04} (statistically significant) respectively.

2. Hue seems to have a negative correlation with power (MFCC coefficients) with a Pearson's coefficient of -0.54 (p=0.056).

3. Lightness seems to be positively correlated with power and pitch with Spearman's coefficient of 0.30 (p=0.03) and 0.38 (p=0.05)

PEMC Test 2 (Stimulus: Transition in HSL Space (X), Response: Pen gestures (Z))

1. Saturation seems to be positively correlated with pressure (Pearson's coeff = 0.34, p= 0.02), total distance (Spearman's coeff = 0.52, p = 0.04) and velocity(Pearson's coeff = 0.4, p= 0.03).

2. Lightness seems to be positively correlated with peaks (Spearman's coefficient = 0.35, p = 0.03) and pressure (Spearman's coefficient = 0.42, p = 0.04).

PEMC Test 3 (Stimulus: Transition in HSL Space (X), Response: Audio (Z) and Pen(Y) simultaneously)

1. Saturation seems to be positively correlated with pressure (Pearson's coefficient = 0.28, p= 0.04), total distance (Spearman's coefficient = 0.34, p = 0.03), pitch(Spearman's coefficient = 0.42, p = 0.007)

2. Lightness seems to be positively correlated with peaks (Spearman's coefficient = 0.28, p = 0.008) and pitch (Spearman's coefficient = 0.26 p=0.03)

The aforementioned features in the Response space (Y and Z) were found weakly correlated with the HSL space (X) in Test settings 1 and 2. In test setting 3, Features like Pressure and Total Distance (Z response) & Pitch (Y response) were positively correlated with Saturation. Hence, according to the PEMC framework, pressure and Total Distance could be positively correlated to Pitch. Similarly, Peaks and Pitch seem to be weakly correlated with each other. As the goal of PEMC is to provide a preliminary evaluation of the presence of possible correlations, more controlled studies might be conducted to verify the presence of correlations where the users can respond to Y space with Z as the stimulus or vice-versa. Furthermore, there were many features in the Y and Z response spaces which showed the presence of correlations and were statistically significant (p value < 0.05), hence future experiments were planned with more participants (N=8) and 140 samples of responses were collected for each of the test setting.

6.2 Experiment 2 (N=8)

6.2.1 User 1

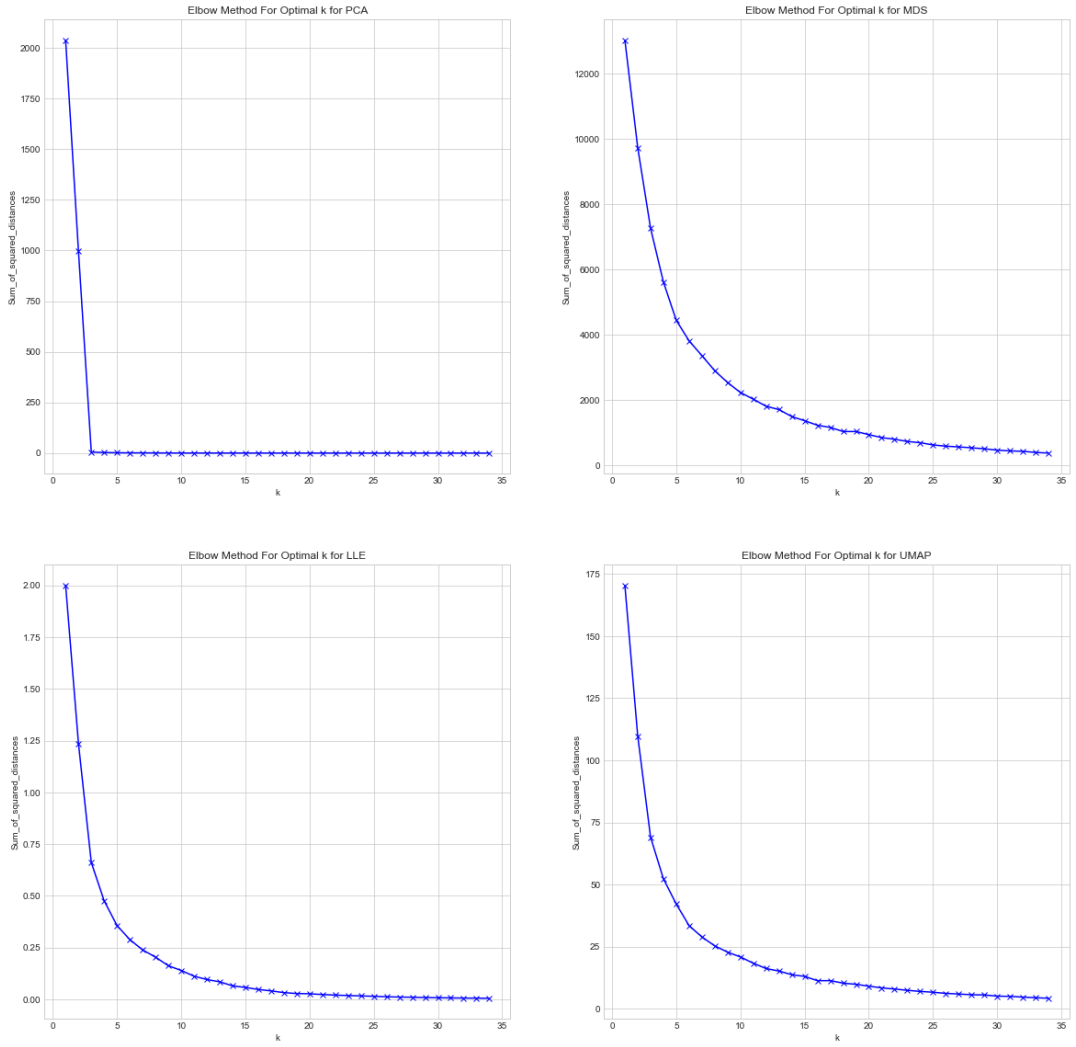


Figure 12: Elbow plots for finding optimal K value for clustering, audio/verbal only response, U1

The 2-D projections between WCSS against the number of clusters are shown in figure 12 for all dimensionality-reduction algorithms for audio/verbal only response by U1. Identifying the "elbow" point in the plot where the rate of decrease in WCSS slows down, it can be seen that K = 3 is a good K value to create the clusters considering the total number of samples is 120. The same principle will be used to find the K-Value and similar graphs for K-value predictions will be shown for each of the test settings. The most clearly separated clusters shall result in the best possible results for correlation analysis.

Figure 13: Latent Space Cluster Visualization of latent feature space, audio/verbal only response, U1

Figure 5 shows an example of visualization of the 2-D projections of the feature data and the clusters created by the K-Means algorithm. However, in all the experiments done in the thesis, the visualisation will be slightly different. Instead of the colours denoting different classes of clusters, shapes will be used to show the classes of clusters and the colours in the scatter plot will represent the stimulus colours shown to the user. These colours are simply the average of H, S and L values when the colour changes from one point in HSL space to another. The figure 13 can be taken as an example and all the results will be shown with scatter plots in this format.

For the sound/verbal responses to the change in HSL space by user 1, the projections produced for different dimensionality reduction techniques are shown in figure 13. The representations produced depend on the nature of the data. Since audio feature vectors are high-dimensional in nature, the PCA algorithm fails to produce clear clusters and most of the data projections lie in the same region with future outliers in others. Non-linear projections like MDS, LLE and UMAP tend to do a better job and LLE

creates clear groups with highly separated clusters.

No.	DR	C.Comp	ΔC	Feature	Correlation Type	Corr.coeff	$\rho-$value
1	LLE	H	35.72	MFCC	Pearsons	-0.45	0.032
2	LLE	H	38.48	MFCC	Pearsons	-0.36	0.036
3	LLE	H	18.66	PV	Pearsons	0.26	0.043
4	LLE	H	41.36	MFCC	Pearsons	-0.28	0.044
5	LLE	S	38.22	PV	Pearsons	0.43	0.032
6	LLE	S	36.40	PV	Pearsons	0.39	0.02
7	LLE	S	26.78	PV	Pearsons	0.24	0.038
8	LLE	L	42.66	PV	Spearmans	0.35	0.037
9	UMAP	L	45.33	Power	Spearmans	0.42	0.042
10	LLE	L	36.42	Power	Pearsons	0.36	0.03

Table 5: Table to Sound only correlation analysis, U1

In the test setting when the user is responding to change in HSL space with audio responses, many weaker correlations have emerged. MFCC features (capturing important spectral contents) are negatively correlated with ΔH. The pitch seems to be correlated to the ΔS. Furthermore, the Power of the audio signal is positively correlated with the ΔL.

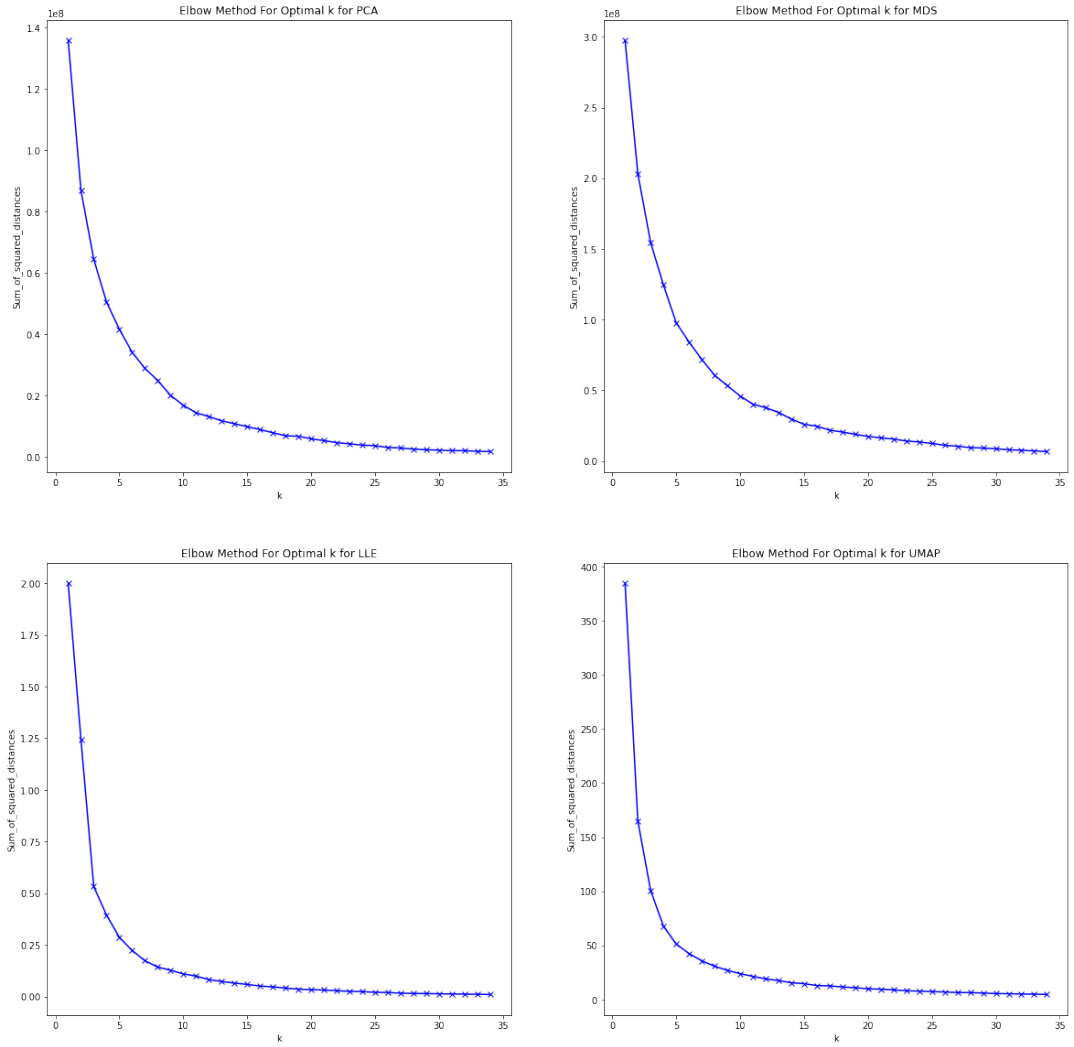


Figure 14: Elbow plots for finding optimal K value for clustering, pen gesture only response, U1

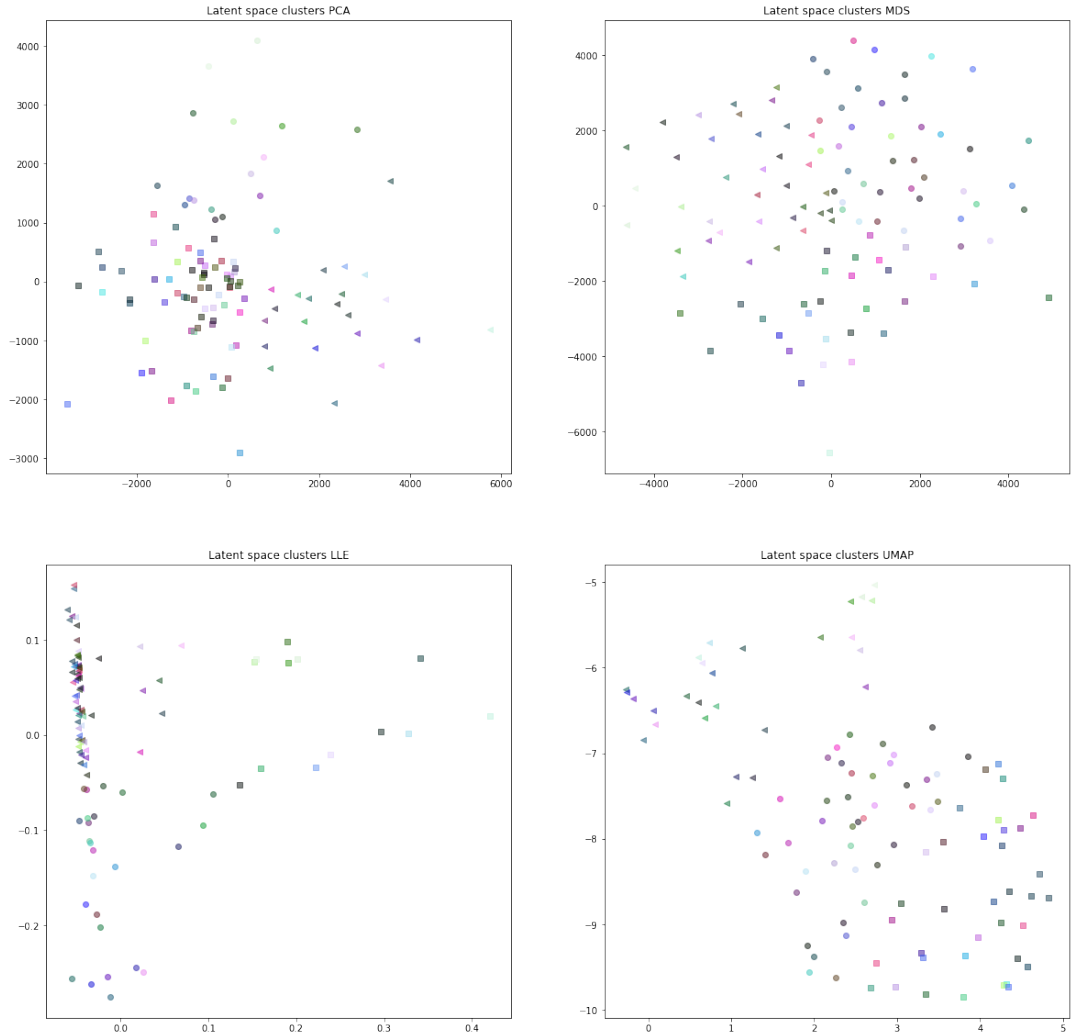


Figure 15: Latent Space Cluster Visualization of latent feature space, pen gesture-only response, U1

No.	DR	C.Comp	ΔC	Feature	Corr. Type	Corr.coeff	$\rho-$value
1	PCA	H	24.65	Peaks	Spearmans	0.245	0.042
2	MDS	H	24.33	Total Distance	Spearmans	0.319	0.098
3	MDS	H	24.33	Peaks	Spearmans	0.29	0.048
4	MDS	H	24.33	Peaks	Pearsons	0.4	0.005
5	LLE	H	24.39	Total Distance	Spearmans	0.379	0.068
6	LLE	S	32.11	Peaks	Pearsons	0.42	0.04
7	LLE	S	35.16	Pressure	Spearmans	0.37	0.03
8	PCA	L	45.22	Peaks	Spearmans	0.38	0.034
9	PCA	L	36.84	Total Distance	Pearsons	0.46	0.031

Table 6: Table to Pen only correlation analysis, U1

For test 2 of the PEMC framework, The 2-D projection graphs shown in figure 15 shows the clusters created with K = 3. The pen gesture data seems to be less complicated than the audio/verbal response and the clusters created with the PCA algorithm show the presence of some linearity in the data. However, the clusters created by LLE are also widely separated and balanced. Many weaker correlations have emerged in the H, S and L space, features like peaks and total distance are positively correlated with the ΔH, ΔS and ΔL.

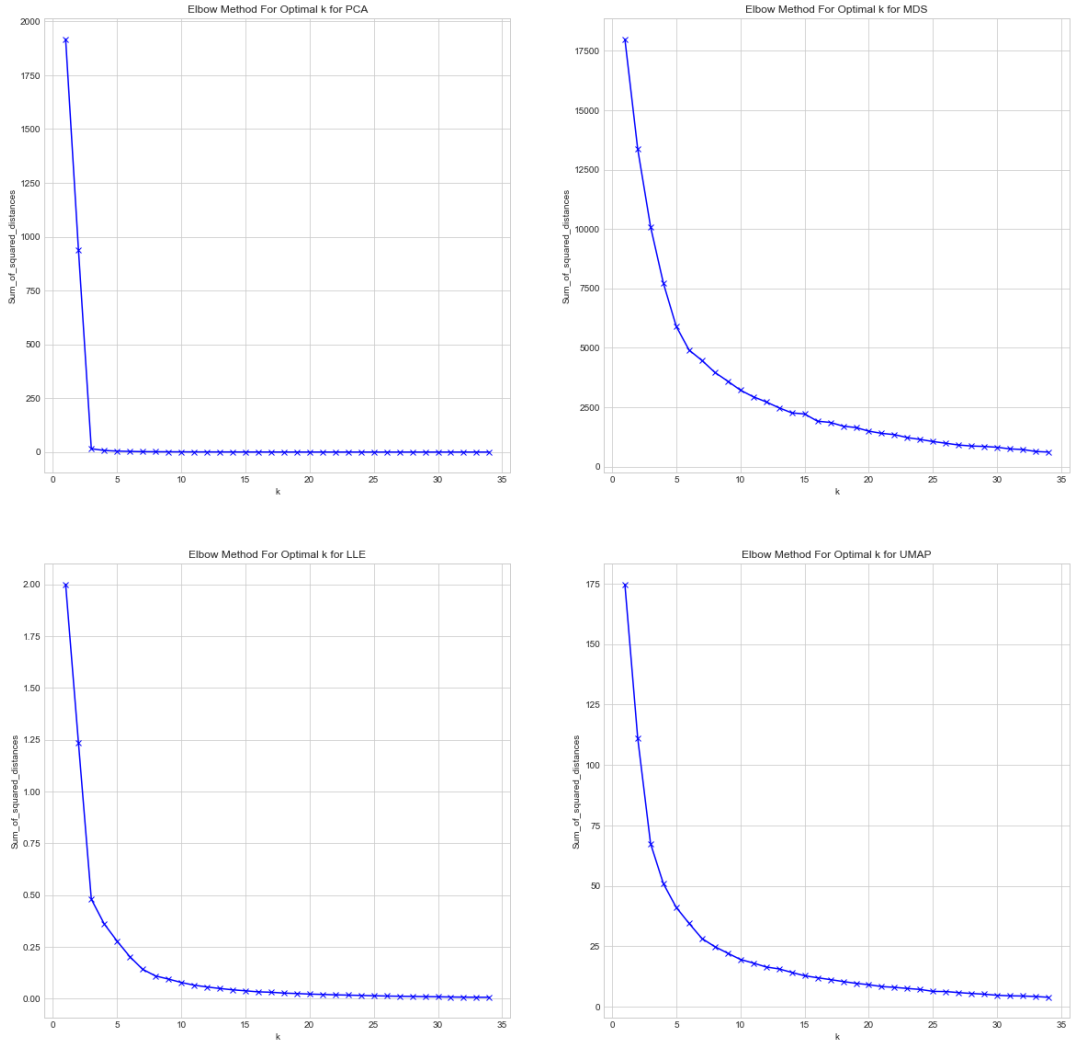


Figure 16: Elbow plots for finding optimal K value for clustering, audio/verbal response (simultaneous pen and audio response), U1

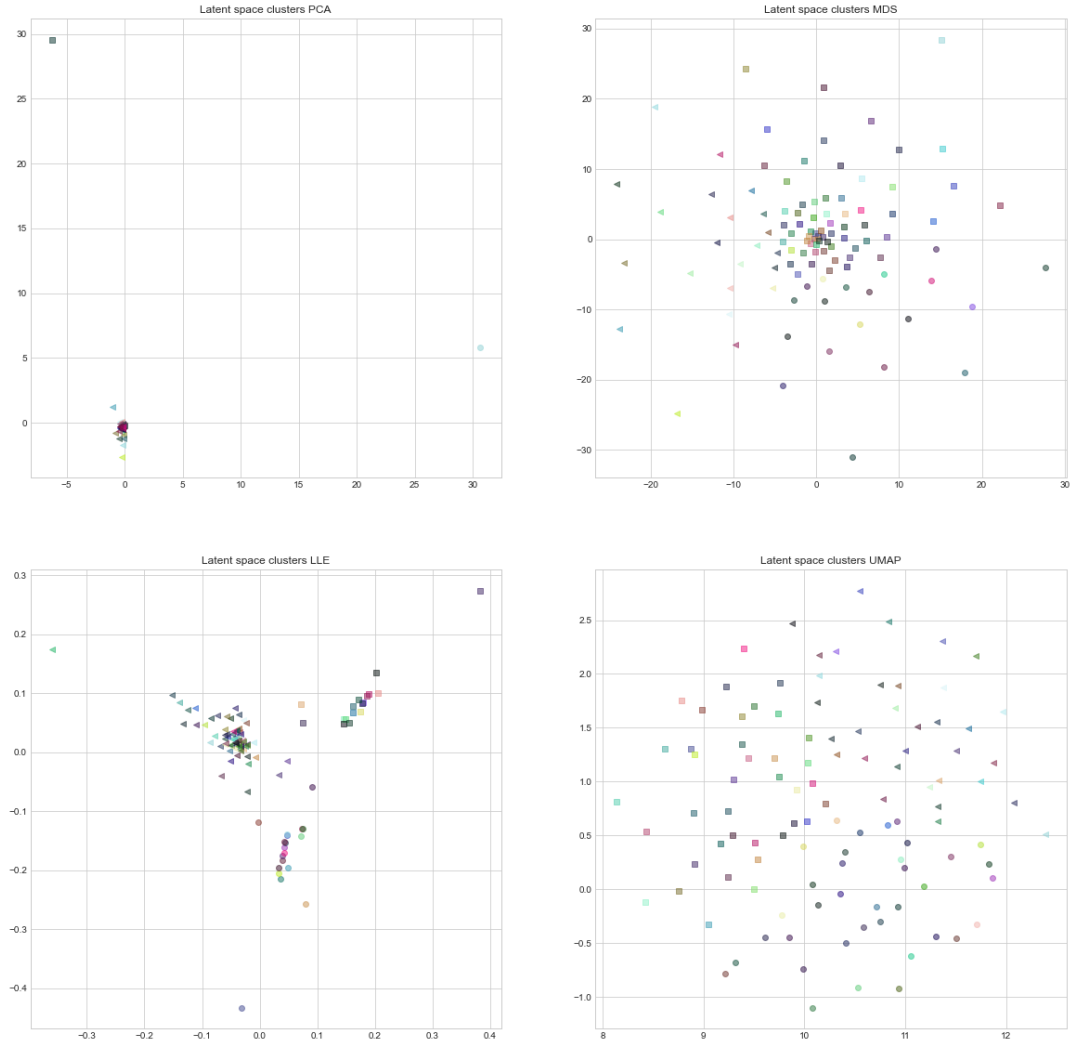


Figure 17: Latent Space Cluster Visualization of latent feature space, audio/verbal response (simultaneous pen and audio response), U1

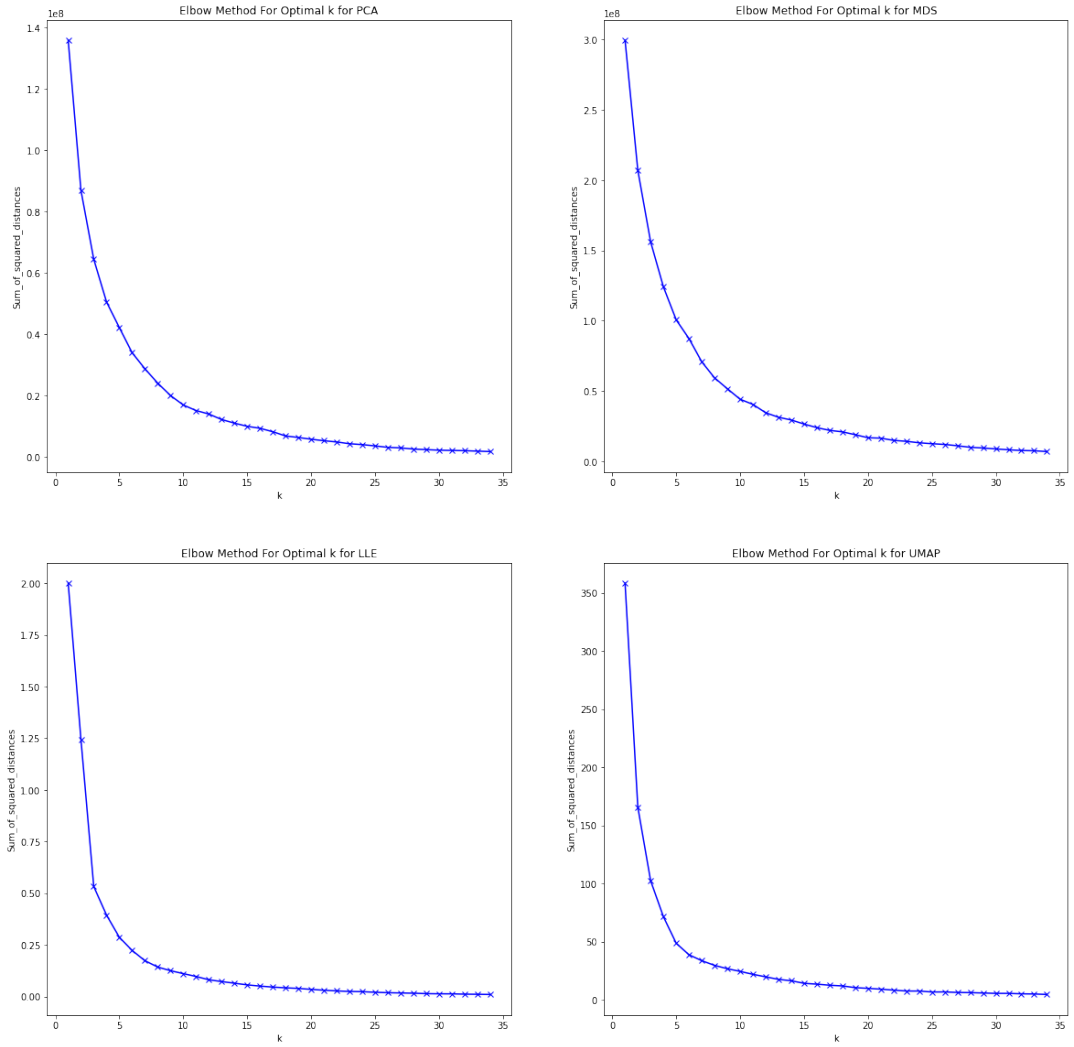


Figure 18: Elbow plots for finding optimal K value for clustering, pen response (simultaneous pen and audio response), U1

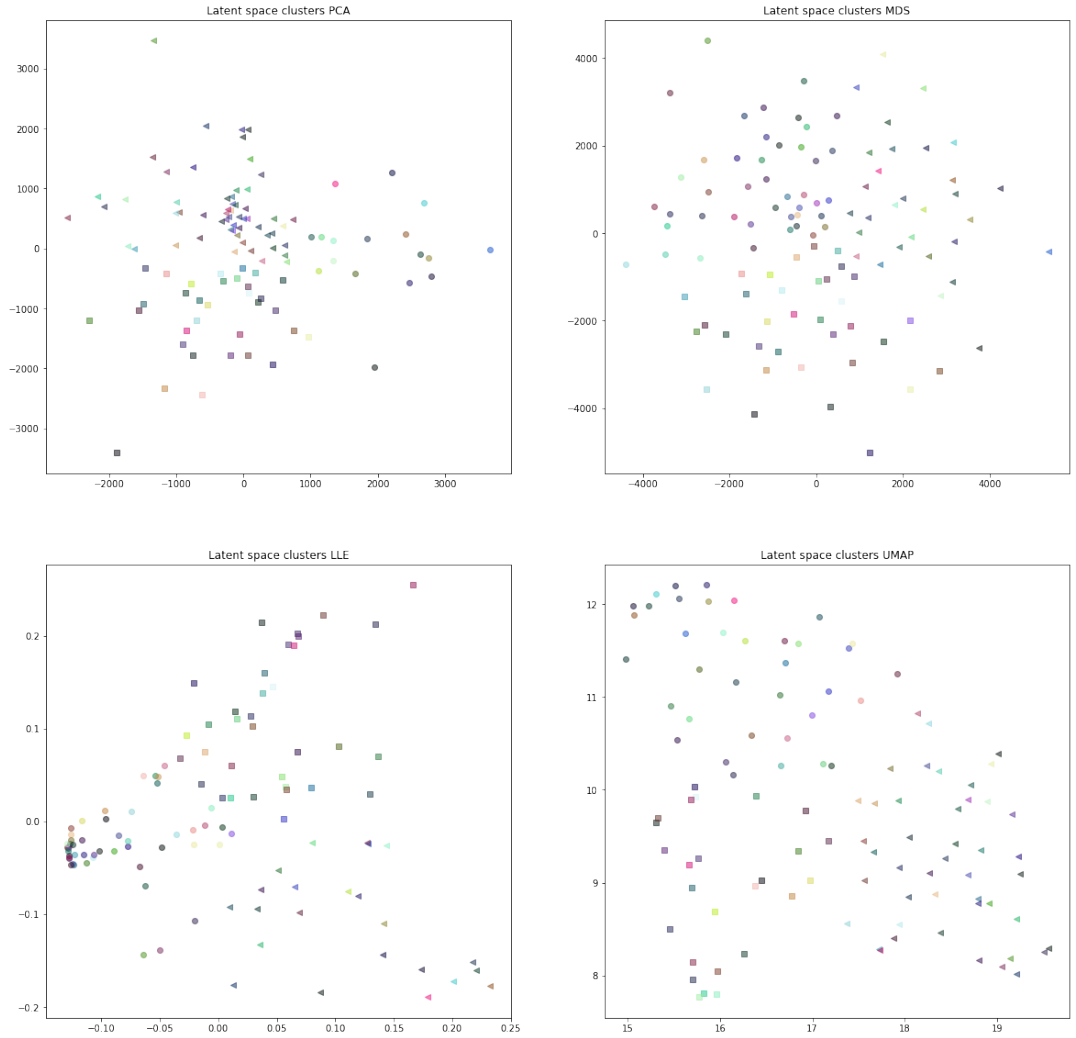


Figure 19: Latent Space Cluster Visualization of latent feature space, pen response (simultaneous pen and audio response), U1

No.	DR	C.Comp	ΔC	Feature	Corr. Type	Corr.coeff	$\rho-$value
1	LLE	H	25.65	PV	Pearsons	0.29	0.031
2	LLE	H	28.44	Peaks	Spearmans	0.38	0.031
3	UMAP	H	26.88	Power	Pearsons	0.39	0.03
4	LLE	S	28.62	Pressure	Spearmans	0.32	0.03
5	UMAP	S	32.14	PV	Pearsons	0.33	0.043
6	UMAP	S	33.68	Peaks	Spearmans	0.43	0.033
7	UMAP	S	36.4	Peaks	Pearsons	0.35	0.037
8	LLE	S	35.33	PV	Pearsons	0.41	0.021
9	LLE	S	35.33	PV	Spearmans	0.39	0.029
10	LLE	S	35.33	Pressure	Spearmans	0.39	787
11	LLE	L	28.4	PV	Pearsons	0.364	0.033
12	LLE	L	32.6	Pressure	Spearmans	0.325	0.011
13	UMAP	L	43.4	Peaks	Spearmans	0.27	0.032

Table 7: Table to Simultaneous Sound and Pen Gestures correlation analysis, U1

Figure 17 and 19 show the 2-D representation of the latent space of audio and pen responses respectively in the simultaneous response setting 3. In the graphs, LLE and UMAP seem to create good clusters in the latent space. According to the test of PEMC, some correlations have emerged in the HSL space in both the audio and pen gestures space. For instance, in the H space, the pitch and power of audio signals are positively correlated to ΔH. The same happens with peaks in the pen gestures response in the H space. Similarly in the S space, Peaks and pressure of the pen gestures are positively correlated to ΔS. In the verbal audio space, the pitch is positively correlated with ΔS. Furthermore, in the L space, pitch (audio response), pressure and peaks (pen response) are positively correlated. According to the PEMC framework, possible correlations may exist in the multimodal spaces of audio and pen gestures and future controlled experiments need to be done to verify the following :

1. Pitch and Power (audio space) might be positively correlated to peaks (pen responses).

2. Pitch (audio space) might be positively correlated to Peaks and Pressure (pen responses).

For user 1, the Correlation Analyser finds out many weaker correlations and features like pitch, power, peaks and pressure were found to correlate to the colour space features, satisfying the PEMC tests 1, 2 and 3.

Please refer to the A section to see the elbow criterion plots and the 2-D latent space cluster plots for user 2 to user 8. The following sections will discuss the results that emerged from the correlation analysis in the 3 test settings for user 2 to user 8. The tables will show any emerging multimodal correlations in the tests.

6.2.2 User 2

No.	DR	C.Comp	ΔC	Feature	Corr. Type	Corr.coeff	$\rho-$value
1	LLE	H	26.88	Power	Pearsons	0.34	0.046
2	UMAP	H	15.28	PV	Spearmans	0.38	0.032
3	UMAP	H	24.4	Power	Spearmans	0.26	0.039
4	UMAP	S	18.44	PV	Pearsons	0.32	0.032
5	UMAP	S	21.46	PV	Pearsons	0.39	0.045
6	LLE	L	28.6	PV	Spearmans	0.26	0.033
7	LLE	L	24.35	PV	Pearsons	0.38	0.045

Table 8: Table to Sound only correlation analysis, U2

Figure A2 shows the 2-D representation of latent space for verbal responses only for user 2 with K = 3. The algorithms LLE and UMAP seem to create the best possible representation with clear, well-separated groups. The correlation analysis also shows weaker correlations emerging with LLE and UMAP dimensionality reduction techniques. In the H space, power and pitch are positively correlated to ΔH. The pitch gets correlated positively to both ΔL and ΔS.

No.	DR	C.Comp	ΔC	Feature	Corr. Type	Corr.coeff	$\rho-$value
1	LLE	H	47.72	Peaks	Spearmans	0.48	0.042
2	LLE	H	47.72	Average Velocity	Spearmans	0.38	0.052
3	LLE	H	25.33	Peaks	Pearsons	0.24	0.04
4	UMAP	H	19.84	Peaks	Pearsons	0.38	0.02
5	LLE	S	32.5	Total Distance	Spearmans	0.29	0.056
6	UMAP	S	25.33	Peaks	Pearsons	0.36	0.044
7	PCA	S	35.44	Pressure	Spearmans	0.46	0.041
8	LLE	L	38.45	Velocity	Spearmans	0.44	0.04
9	LLE	L	38.45	Velocity	Pearsons	0.39	0.045
10	UMAP	L	45.22	Peaks	Pearsons	0.33	0.049

Table 9: Table to Pen only correlation analysis, U2

For the pen-only responses for U2, LLE and UMAP again perform a good job separating the clusters. Peaks and average velocity get positively correlated to ΔH. Total distance, Peaks and pressure gets positively correlated with the ΔS. Finally, velocity and Peaks are positively correlated with ΔL.

No.	DR	C.Comp	ΔC	Feature	Corr. Type	Corr.coeff	$\rho-$value
1	LLE	H	28.89	Peaks	Pearsons	0.41	0.031
2	LLE	H	28.45	PV	Spearmans	0.34	0.022
3	UMAP	H	35.22	Pressure	Spearmans	0.37	0.036
4	LLE	S	25.54	Pressure	Spearmans	0.38	0.04
5	LLE	S	25.54	Total Distance	Pearsons	0.39	0.02
6	LLE	S	25.54	PV	Pearsons	0.36	0.034
7	UMAP	L	38.5	Power	Spearmans	0.32	0.027
8	LLE	L	44.36	Peaks	Pearsons	0.38	0.036
9	LLE	L	42.8	Power	Pearsons	0.31	0.025

Table 10: Table to Simultaneous Sound and Pen responses correlation analysis, U2

Peaks, Pressure (pen gesture response) and pitch (audio response) are positively correlated to the ΔH. According to the PEMC framework, possible correlations may exist in the multimodal spaces of audio and pen gestures and future controlled experiments need to be done to verify the following :

1. Peaks, Pressure (pen gesture response) and pitch (audio response) are positively correlated.

2. Pressure, total distance (pen response) and pitch (audio response) are positively correlated.

3. Peaks (pen response) and power (audio response) are positively correlated.

It is interesting to note that velocity, which was correlated to ΔL in pen-only responses, was not correlated in test 3. Hence, it is possible that the presence of additional response space in a task can affect the responses of a user and features which emerge as having weaker correlations might not be present in a dual response setting in test 3.

Please refer to A.1 to see the elbow criterion plots and the 2-D latent space cluster plots for user 2.

6.2.3 User 3

No.	DR	C.Comp	ΔC	Feature	Corr. Type	Corr.coeff	$\rho-$value
1	UMAP	H	55.8	Spec Centroid	Spearmans	0.31	0.06
2	UMAP	H	56.54	RMS	Spearmans	0.35	0.03
3	LLE	S	10.95	PV	Spearmans	0.38	0.08
4	LLE	S	10.95	Chroma Energy	Spearmans	-0.37	0.09
5	LLE	S	10.95	MFCC	Spearmans	-0.39	0.07
6	UMAP	S	17.16	PV	Spearmans	0.27	0.09
7	UMAP	S	17.16	PVFT	Spearmans	-0.28	0.08
8	UMAP	S	17.16	Chroma Energy	Spearmans	-0.36	0.02

Table 11: Table to Sound only correlation analysis, U3

Many weaker correlations have emerged in the audio-only responses for user 3. UMAP and LLE have performed well to cluster the latent space representation. Spectral centroid (brightness of the spectral content) and RMS (average power or amplitude of audio) was positively correlated to ΔH. The pitch was positively correlated, Chroma energy (average value of chroma of the audio signal), MFCC coefficients and PVFT (pitch of musical octave) were negatively correlated to the ΔS.

No.	DR	C.Comp	ΔC	Feature	Corr. Type	Corr.coeff	$\rho-$value
1	PCA	H	60.5	Average Velocity	Pearsons	-0.36	0.03
2	PCA	H	60.5	Total Distance	Pearsons	-0.29	0.08
3	PCA	H	48.52	Average Velocity	Pearsons	-0.33	0.08
4	LLE	S	14.7	Average Velocity	Spearmans	-0.4	0.05
5	LLE	S	14.7	Total Distance	Spearmans	-0.34	0.09
6	LLE	S	13.87	Average Velocity	Spearmans	-0.47	0.02
7	MDS	S	13.87	Average Velocity	Pearsons	-0.36	0.08
8	PCA	S	15.96	Total Distance	Spearmans	-0.39	0.02
9	MDS	S	13.87	Angles	Pearsons	-0.43	0.03
10	PCA	L	13.56	Pressure	Spearmans	0.31	0.03
11	MDS	L	13.48	Pressure	Spearmans	0.29	0.04
12	UMAP	L	13.47	Pressure	Spearmans	0.43	0.004
13	PCA	L	14.2	Average Velocity	Pearsons	-0.36	0.015
14	LLE	L	12.66	Average Velocity	Pearsons	-0.38	0.003
15	UMAP	L	13.45	Average Velocity	Pearsons	-0.48	0.001
16	UMAP	L	13.45	Pressure	Pearsons	0.3	0.05

Table 12: Table to Pen only correlation analysis, U3

Many weaker correlations have emerged in the pen gesture responses for user 3 as well. The latent representation of features shown in figure A12 shows all 4

dimensionality reduction algorithms performing well to create distant clusters. Average velocity and total distance are negatively correlated to ΔH, unlike users 1 and 2. This shows that these subjective experiences of different modalities might vary from person to person and the sample size of the experiments need to be increased to get the realization of universal correlations. This concurs with the arguments of Dennet that there is a presence of possible inter-subjective correlations between different qualia for different individuals. Similarly, in the S space, total distance, average velocity and angles emerged as negatively correlated to ΔS. Furthermore, Pressure is positively correlated and average velocity is negatively correlated to ΔL.

No.	DR	C.Comp	ΔC	Feature	Corr. Type	Corr.coeff	$\rho-$value
1	PCA	H	13.97	Average Velocity	Spearmans	0.36	0.02
2	PCA	H	13.97	Total Distance	Spearmans	0.45	0.004
3	MDS	H	2.77	Average Velocity	Spearmans	0.28	0.004
4	MDS	H	2.77	Total Distance	Spearmans	0.38	0.01
5	UMAP	H	13.97	Average Velocity	Pearsons	0.36	0.02
6	LLE	H	13.97	Total Distance	Pearsons	0.39	0.01
7	PCA	H	2.77	Average Velocity	Pearsons	0.29	0.08
8	PCA	H	2.77	Total Distance	Pearsons	0.4	0.01
9	LLE	H	18.66	PV	Spearmans	0.98	0.001
10	LLE	H	18.66	Chroma Energy	Spearmans	0.90	0.001
11	UMAP	H	5	Spec Contrast	Spearmans	0.38	0.001
12	LLE	S	1.51	PV	Spearmans	0.28	0.01
13	LLE	S	1.51	PVFT	Spearmans	0.28	0.002
14	LLE	S	1.51	Chroma Energy	Spearmans	-0.93	0.003
15	UMAP	S	1.2	PV	Spearmans	-0.89	0.01
16	UMAP	S	1.2	PVFT	Spearmans	-0.8	0.007
17	MDS	L	18.33	MFCC	Spearmans	-0.98	0.002
18	MDS	L	18.33	RMS	Spearmans	-0.99	0.04
19	MDS	L	18.33	Spec Centroid	Spearmans	0.93	0.01
20	MDS	L	18.33	Spec Bandwidth	Spearmans	0.99	0.04
21	MDS	L	18.33	Spec Contrast	Spearmans	-0.89	0.03

Table 13: Table to Simultaneously Sound and Pen correlation analysis, U3

In test 3, No correlation emerged between pen responses and pitch, PVFT, and Chroma energy (audio responses) are positively correlated to ΔS. No correlation emerged from pen responses in L space. Furthermore, in audio responses, RMS values, and spectral contrast are negatively correlated and spectral centroid and spectral bandwidth are positively correlated with very high correlation coefficients to ΔL. According to the PEMC framework, possible correlations may exist in the multimodal spaces of audio and pen gestures and future controlled experiments need to be done to verify the following :

1. Average Velocity, total distance (pen gesture response) and pitch, chroma Energy,

and spectral contrast (audio response) might be correlated.

Although very high correlations emerged in ΔL and audio responses, there were no correlations present in the pen gesture response. Hence, for user 3, PEMC states that there are no strong correlations between the pen gestures space and audio space. But further experiments may still be concluded for testing correlations between pitch and average velocity total distance on the basis of weaker correlations. On the contrary, further experiments might be conducted to verify the strong correlation between the ΔL and audio space.

Please refer to A.2 to see the elbow criterion plots and the 2-D latent space cluster plots for user 3.

6.2.4 User 4

No.	DR	C.Comp	ΔC	Feature	Corr. Type	Corr.coeff	$\rho-$value
1	LLE	H	26.88	Power	Pearsons	0.34	0.046
2	UMAP	H	15.28	PV	Spearmans	0.38	0.032
3	UMAP	H	24.48	Power	Spearmans	0.26	0.039
4	UMAP	S	18.44	PV	Pearsons	0.32	0.032
5	UMAP	S	21.46	PV	Pearsons	0.39	0.045
6	LLE	L	28.6	PV	Spearmans	0.26	0.033
7	LLE	L	24.35	PV	Pearsons	0.38	0.045

Table 14: Table to Sound only correlation analysis, U4

For user 4, figure A18 shows the 2-D latent space representation of audio-only responses using dimensionality reduction techniques. It can be seen that LLE and UMAP have been able to create better clusters and that shows in the table of correlations. In the H space, power and pitch are weakly correlated to ΔH. Pitch is positively correlated to change to ΔS and ΔL.

No.	DR	C.Comp	ΔC	Feature	Corr. Type	Corr.coeff	$\rho-$value
1	PCA	H	5.92	Pressure	Spearmans	-0.4	0.04
2	UMAP	H	3.67	Pressure	Spearmans	-0.33	0.04
3	LLE	H	5.92	Pressure	Pearsons	-0.36	0.02
4	UMAP	H	3.67	Pressure	Pearsons	-0.4	0.01
5	PCA	S	4.9	Angles	Spearmans	-0.76	0.01
6	LLE	S	7.9	Angles	Spearmans	-0.59	0.007
7	PCA	S	4.9	Angles	Pearsons	-0.7	0.02
8	LLE	S	7.58	Angles	Pearsons	-0.51	0.02
9	LLE	L	2.7	Pressure	Spearmans	-0.46	0.04
10	LLE	L	2.7	Pressure	Pearsons	-0.47	0.03
11	LLE	L	5.12	Average Velocity	Pearsons	0.3	0.03
12	LLE	L	5.12	Total Distance	Pearsons	0.37	0.01
13	UMAP	L	5.7	Average Velocity	Pearsons	0.38	0.01

Table 15: Table to Pen only correlation analysis, U4

For pen-only responses, UMAP, LLE and PCA have shown correlated features in the clusters. Pressure and angles are negatively correlated to ΔS and ΔH. Multiple correlations emerged with the change in L space, where Pressure is negatively correlated and average velocity and total distance are positively correlated.

No.	DR	C.Comp	ΔC	Feature	Corr. Type	Corr.coeff	$\rho-$value
1	PCA	S	13	Average Velocity	Spearmans	1	0.03
2	PCA	S	13	Total Distance	Spearmans	1	0.02
3	MDS	S	3.4	Angles	Spearmans	-0.3	0.03
4	MDS	S	3.6	Angles	Spearmans	-0.35	0.03
5	UMAP	S	3.3	Angles	Pearsons	-0.3	0.04
6	LLE	S	2.5	Pressure	Spearmans	-1	0.02
7	PCA	L	6.6	Average Velocity	Spearmans	-1	0.02
8	PCA	L	6	Total Distance	Spearmans	-0.9	0.03
9	PCA	L	6.6	Average Velocity	Pearsons	-0.8	0.04
10	PCA	L	0.8	Spec Contrast	Spearmans	-0.8	0.04
11	PCA	L	4.2	Spec Contrast	Spearmans	-0.8	0.04
12	MDS	L	0.8	RMS	Spearmans	-0.8	0.04
13	LLE	L	0.8	Spec Contrast	Spearmans	-0.54	0.024
14	LLE	L	0.8	PVFT	Spearmans	-0.98	0.031
15	LLE	L	0.8	Spec Bandwidth	Spearmans	-0.99	0.004
16	UMAP	L	4.8	Spec Contrast	Spearmans	-0.34	0.034
17	UMAP	L	4.8	MFCC	Spearmans	0.44	0.004

Table 16: Table to simultaneous Sound and pen responses correlation analysis, U4

Many interesting results were observed for U4 simultaneous audio and pen gesture

responses to changes in colour space. No correlations emerged in the H space in both audio and pen responses, which were present in test 1 and test 2 for the same user. In the S space, Average velocity and total distance are strongly positively correlated to ΔS. Angles and rressure were found to be negatively correlated to ΔS. No features from the audio responses were found to correlate to the S space. Furthermore, in the L space, many new correlations emerged following test 1, spectral contrast, spectral bandwidth, RMS, PVFT, and spec contrast were negatively correlated and MFCC coefficients were positively correlated to ΔL. For the pen responses. Average velocity and total distance were negatively correlated to ΔL.

According to the PEMC framework, possible correlations may exist in the multi-modal spaces of audio and pen gestures and future controlled experiments need to be done to verify the following :

1. Average velocity, total distance (pen responses) and MFCC, Spec Bandwidth, PVFT, spec contrast, RMS (audio response)

Please refer to A.3 to see the elbow criterion plots and the 2-D latent space cluster plots for user 4.

6.2.5 User 5

No.	DR	C.Comp	ΔC	Feature	Corr. Type	Corr.coeff	$\rho-$value
1	UMAP	H	19.63	PV	Spearmans	0.364	0.03
2	UMAP	H	19.63	MFCC	Spearmans	0.34	0.04
3	LLE	S	0.52	PV	Spearmans	0.22	0.02
4	LLE	S	0.52	Spec Contrast	Spearmans	-0.178	0.09
5	LLE	S	7.06	PV	Spearmans	0.22	0.02
6	UMAP	S	7.06	Spec Contrast	Spearmans	-0.478	0.004
7	UMAP	S	7.06	RMS	Spearmans	-0.33	0.048
8	LLE	L	7	PVFT	Spearmans	-0.89	0.03

Table 17: Table to Sound only correlation analysis, U5

For the audio-only responses for user 5, UMAP and LLE resulted in the best possible clusters. Pitch and MFCC coefficients were positively correlated to ΔH. Pitch and RMS values were positively correlated while spectral contrast was negatively correlated to ΔS. PVFT was strongly negatively correlated to ΔL.

No.	DR	C.Comp	ΔC	Feature	Corr. Type	Corr.coeff	$\rho-$value
1	UMAP	H	69.97	Angles	Spearmans	0.385	0.008
2	UMAP	H	70.76	Angles	Spearmans	0.378	0.001
3	LLE	H	80.93	Angles	Spearmans	0.327	0.001
4	LLE	H	15.1	Pressure	Spearmans	1	0.003
5	LLE	H	15	Angles	Spearmans	1	0.004
6	UMAP	S	72	Angles	Spearmans	-0.44	0.001
7	LLE	S	26.1	Pressure	Spearmans	1	0.002
8	LLE	S	26.1	Angles	Spearmans	-1	0.003
9	LLE	S	16.89	Average velocity	Spearmans	-0.268	0.034
10	LLE	S	16.89	Total Distance	Spearmans	-0.29	0.021
11	LLE	L	22.21	Pressure	Spearmans	1	0.001

Table 18: Table to Pen only correlation analysis, U5

For the pen-only responses, UMAP and LLE created the best possible clusters in the latent space. Angles and pressure were positively correlated to ΔH. Angles, average velocity, and total distance were negatively correlated and pressure was positively correlated to ΔS. Furthermore, pressure was positively correlated to ΔL.

No.	DR	C.Comp	ΔC	Feature	Corr. Type	Corr.coeff	$\rho-$value
1	PCA	H	99.26	Pressure	Spearmans	-0.36	0.001
2	MDS	H	100.4	Pressure	Spearmans	-0.36	0.002
3	LLE	H	101.6	Pressure	Spearmans	-0.39	0.009
4	PCA	H	99.2	Pressure	Pearsons	-0.34	0.002
5	PCA	H	21.06	Average Velocity	Pearsons	0.52	0.04
6	PCA	H	21.06	Total Distance	Pearsons	0.53	0.03
7	MDS	H	100	Pressure	Pearsons	-0.35	0.02
8	LLE	H	101	Pressure	Pearsons	-0.404	0.007
9	PCA	L	19.12	Pressure	Pearsons	-0.31	0.03
10	PCA	L	15.75	Average Velocity	Pearsons	-0.51	0.04
11	MDS	H	9.46	MFCC	Spearmans	-0.468	0.03
12	MDS	L	2.98	Chroma Energy	Spearmans	0.38	0.01

Table 19: Table to Sound only correlation analysis, U5

Interesting correlations emerged in test 3 for user 5 where no correlations emerged in the S space. Also, pressure which was positively correlated in the pen-only response test 2 was found to be negatively correlated in the simultaneous task. It can be inferred that the presence of an additional modality might change the behavioural pattern in a certain modality. The pressure was found to be negatively correlated to ΔH while average velocity and total distance were positively correlated. In the audio responses, MFCC coefficients were negatively correlated to ΔH as well. Pressure,

average velocity (pen response) and chroma energy were negatively and positively correlated respectively to ΔL.

According to the PEMC framework, possible correlations may exist in the multi-modal spaces of audio and pen gestures and future controlled experiments need to be done to verify the following :

1. Pressure, Average velocity, total distance (pen responses) and MFCC coefficients (audio responses)

2. Pressure, average velocity (pen response) and chroma energy (audio responses)

Please refer to A.4 to see the elbow criterion plots and the 2-D latent space cluster plots for user 5.

6.2.6 User 6

No.	DR	C.Comp	ΔC	Feature	Corr. Type	Corr.coeff	$\rho-$value
1	MDS	H	6.63	PV	Spearmans	-0.482	0.04
2	MDS	H	6.63	Spec Contrast	Spearmans	0.473	0.03
3	MDS	S	5.18	Spec Centroid	Spearmans	-0.33	0.007
4	UMAP	S	6.63	PVFT	Spearmans	-1.0	0.03
5	MDS	L	6.63	Chroma Energy	Spearmans	0.26	0.03

Table 20: Table to Sound only correlation analysis, U6

For the sound only responses, MDS and UMAP algorithms create the most distant clusters. Pitch and spectral contrast are positively and negatively correlated respectively to ΔH. PVFT is strongly negatively correlated and spec centroid is weakly negatively correlated to ΔS. Only chroma energy emerged to be weakly positively correlated to ΔL.

No.	DR	C.Comp	ΔC	Feature	Corr. Type	Corr.coeff	$\rho-$value
1	PCA	H	71.04	Angles	Spearmans	-0.429	0.02
2	LLE	S	16.45	Angles	Spearmans	0.33	0.019
3	LLE	S	7.95	Angles	Spearmans	0.39	0.016
4	LLE	S	17.64	Angles	Pearsons	0.33	0.04
5	LLE	L	23.04	Angles	Pearsons	0.59	0.003

Table 21: Table to Pen only correlation analysis, U6

Although all the dimensionality reduction techniques performed well to create clusters in latent space, correlations emerged from PCA and LLE-based K-Means clustering. Angles are negatively correlated to the H space, positively correlated to the S space and positively correlated to ΔL.

No.	DR	C.Comp	ΔC	Feature	Corr. Type	Corr.coeff	$\rho-$value
1	PCA	H	99.26	Pressure	Spearmans	-0.36	0.01
2	LLE	H	101.23	Pressure	Spearmans	-0.39	0.009
3	PCA	H	99.26	Pressure	Spearmans	-0.34	0.02
4	LLE	H	101.12	Pressure	Spearmans	-0.404	0.007
5	UMAP	L	20.82	Pressure	Spearmans	-0.3	0.04
6	PCA	L	20.82	Pressure	Spearmans	-0.51	0.04
7	MDS	L	19.56	Pressure	Spearmans	-0.3	0.03
8	LLE	L	20.82	Pressure	Spearmans	-0.37	0.01
9	UMAP	L	19.62	Pressure	Spearmans	-0.34	0.02
10	UMAP	H	0.63	Spec Bandwidth	Spearmans	-0.25	0.04
11	UMAP	S	4.66	RMS	Spearmans	0.52	0.008
12	UMAP	L	0.6	RMS	Spearmans	0.8	0.03
13	UMAP	L	2.6	PV	Spearmans	-0.34	0.03
14	UMAP	L	2.6	PVFT	Spearmans	-0.4	0.01

Table 22: Table to Simultaneous Pen and Sound responses correlation analysis, U6

For user 6 as well, new correlations emerged in the simultaneous response task for test 3 which did not emerge earlier in test 1 and test 2. Pressure (pen responses) and spectral bandwidth (audio responses) is negatively correlated to ΔH. In the S space, only RMS values were positively correlated and no features were correlated in pen responses. Pressure (pen responses) is negatively correlated to ΔL. RMS values are positively correlated, and pitch and PVFT are both negatively correlated to ΔL. This result concurs with the findings presented in [19] where loudness (RMS equivalent) was found to be correlated with brightness (Lightness). According to the PEMC framework, no possible correlations exist in the multimodal spaces audio and pen gestures as each of the tests had different multimodal correlations emerging but not together in the simultaneous test setting 3.

Please refer to A.5 to see the elbow criterion plots and the 2-D latent space cluster plots for user 6.

6.2.7 User 7

No.	DR	C.Comp	ΔC	Feature	Corr. Type	Corr.coeff	$\rho-$value
1	None	H,S,L	None	None	No Correlations	None	None

Table 23: Table to Sound only correlation analysis, U7

For the audio/verbal responses by user 7, no correlations emerged from the data.

No.	DR	C.Comp	ΔC	Feature	Corr. Type	Corr.coeff	$\rho-$value
1	UMAP	S	16.75	Average velocity	Spearmans	-0.38	0.02
2	UMAP	S	12.36	Average velocity	Spearmans	-0.47	0.036
3	UMAP	S	17.26	Average velocity	Spearmans	-0.45	0.04
4	UMAP	S	12.36	Average velocity	Spearmans	-0.45	0.04

Table 24: Table to Pen only correlation analysis, U7

For the pen-only responses by user 7, the average velocity was the only feature showing correlations to ΔS. No other correlations emerged in H and L space.

No.	DR	C.Comp	ΔC	Feature	Corr. Type	Corr.coeff	$\rho-$value
1	LLE	S	16.1	Average velocity	Spearmans	-0.34	0.02
2	LLE	S	16.1	Total Distance	Spearmans	-0.33	0.02
2	MDS	S	14.84	PV	Spearmans	0.53	0.021
2	MDS	S	14.84	PVFT	Spearmans	0.46	0.001
2	MDS	S	14.84	Spec Contrast	Spearmans	-0.54	0.03
2	MDS	L	4.8	PV	Spearmans	-0.63	0.024
2	MDS	L	4.8	PVFT	Spearmans	-0.65	0.021
2	MDS	L	4.8	Chroma Energy	Spearmans	-0.69	0.034
2	MDS	L	4.8	RMS	Spearmans	0.5	0.031
2	MDS	L	4.8	Spec Bandwidth	Spearmans	-0.58	0.03
2	MDS	L	4.8	Spec Contrast	Spearmans	-0.60	0.01

Table 25: Table to simultaneous Sound and Pen responses only correlation analysis, U7

For the test setting 3, all the dimensionality reduction techniques perform quite similarly for pen responses and UMAP performs well in the audio responses space. Many weaker and moderate correlations emerge in S and L space. In the S space, Average velocity and total distance (pen response) is negatively correlated to ΔS. Pitch and PVFT are positively correlated and spectral contrast (audio response)is negatively correlated to ΔS. In the L space, no correlations emerge in the pen responses. Hence, according to the PEMC framework, correlations in audio and pen space are non-existant as similar correlations in stimuli and response space haven't emerged in all the 3 test settings.

Please refer to A.6 to see the elbow criterion plots and the 2-D latent space cluster plots for user 7.

6.2.8 User 8

No.	DR	C.Comp	ΔC	Feature	Corr. Type	Corr.coeff	$\rho-$value
1	No Correlation Found						

Table 26: Table to Sound only correlation analysis, U8

For the audio/verbal responses by user 8, none of the dimensionality reduction algorithms creates a good representation of latent space and fails to create any well-defined clusters. Furthermore, no correlations emerged from the data.

No.	DR	C.Comp	ΔC	Feature	Corr. Type	Corr.coeff	$\rho-$value
1	PCA	H	55.45	Average velocity	Spearmans	-0.33	0.037
2	PCA	H	55.45	Angles	Spearmans	0.32	0.044
3	PCA	H	54.45	Average velocity	Spearmans	-0.34	0.032
4	PCA	H	106.54	Angles	Spearmans	0.98	0.03
5	LLE	H	55.41	Average velocity	Spearmans	-0.33	0.02
6	UMAP	H	54.44	Average velocity	Spearmans	-0.35	0.02
7	UMAP	H	106.54	Angles	Spearmans	0.93	0.006
8	UMAP	H	54.78	Average velocity	Spearmans	-0.37	0.01
9	UMAP	L	26.78	Angles	Spearmans	0.82	0.04
10	UMAP	S	17.92	Average velocity	Spearmans	-0.28	0.03
11	UMAP	S	17.92	Total Distance	Spearmans	-0.29	0.03
12	LLE	S	17.92	Angles	Spearmans	-0.32	0.018
13	LLE	S	17.92	Average velocity	Spearmans	-0.32	0.017
14	LLE	S	17.92	Total Distance	Spearmans	-0.3	0.03

Table 27: Table to Pen only responses correlation analysis, U8

For the audio-only responses for user 8, PCA, LLE and UMAP create good representations of the latent space of the audio data. In the H space, average velocity emerges as positively correlated and angles are negatively correlated toΔH. Angles are also positively correlated to ΔL. Finally, average velocity, total distance and angles are negatively correlated to ΔS.

No.	DR	C.Comp	ΔC	Feature	Corr. Type	Corr.coeff	$\rho-$value
1	PCA	H	67.49	Total Distance	Spearmans	-0.38	0.04
2	PCA	H	67.49	Average Velocity	Spearmans	-0.4	0.02
3	MDS	H	67.49	Total Distance	Spearmans	-0.4	0.02
4	MDS	H	4.3	Total Distance	Spearmans	-1	0.01
5	UMAP	H	65.2	Average Velocity	Spearmans	0.42	0.02
6	UMAP	H	65.2	Total Distance	Spearmans	-0.43	0.01
7	LLE	S	1.24	Average Velocity	Spearmans	-1	0.01
8	LLE	L	14.38	Pressure	Spearmans	1	0.01

Table 28: Table to Sound only correlation analysis, U8

In the simultaneous task, no correlations emerged for the audio responses. Only pen responses were found to correlate to the colour space. Total distance and average velocity are negatively correlated to ΔH. Average Velocity was also negatively correlated to ΔS. The pressure was found to be positively correlated to ΔL with a strong correlation coefficient.

According to the PEMC framework, tests 1, 2 and 3 are not satisfied by any modalities as audio responses are not correlated in test 1 and test 3.

Please refer to A.7 to see the elbow criterion plots and the 2-D latent space cluster plots for user 8.

7 Discussions and Analysis

7.1 Key Findings

1. From the elbow plots of the user data and the visualization of the latent space of the responses, it could be inferred that there exists a relationship between the responses, i.e, the number of clusters formed in the audio space and the pen gesture space, as the ideal number of clusters corresponds to K = 3 (axes of H, S and L colour model). The K value was chosen where the WCSS began to level off following the elbow criterion.

2. Subjective experiences of different modalities might vary from person to person and the sample size of the experiments need to be increased to get the realization of universal correlations. For example, average velocity and total distance are negatively correlated to ΔH in user 3, unlike users 1 and 2. This concurs with the arguments of Dennet that there is a presence of possible inter-subjective comparisons between different qualia for different individuals.

3. Different clustering techniques are effective against different modalities, for example, LLE or UMAP for audio responses and PCA or LLE for pen gestures, depending on the nature of the data and the linear/non-linear relationships.

4. Many Weaker Correlations emerged (corr. coeff. < 0.3), moderate correlations (0.3 < corr. coeff.< 0.6) and high correlations (0.8 < corr. coeff.) emerged from the 10-person user study and the subsequent correlation analysis of the color to audio, color to pen gestures and color to both audio and pen responses. Future controlled experiments could be planned to verify the presence of these multimodal correlations closely.

5. The Correlation Analyser does a good job of finding the correlations between 2 modalities of data as numerous correlations emerged for each user. This shows the efficiency of EDA-based techniques in analysing a large amount of high-dimensional multimodal data. The found correlations could be used in designing creative and artistic experiments for use cases of human-AI co-creation of arts.

Thus, it can be concluded that PEMC is a good starting point to analyze multimodal data in high dimensional spaces and future experiments could use more controlled pattern-matching techniques followed by user surveys to understand the complexity of the tasks and data. More controlled experiments might have led to stronger correlations but the goal of PEMC is primarily to give indications or primary evaluations for the presence of correlation in highly complex data. Further experiments should be planned with limited stimulus and response space to verify the results shown in the thesis and to continue future work in the field. These findings answer the RQ 1 and RQ 2 as well as establish a foundation for future experiments.

7.2 Limitations and Challenges

1. The experiments might have a steep learning curve for the users, hence most users don't exactly know how to express themselves to a certain stimulus and slow learning might affect the quality of data.

2. Considering the uncontrolled data collection methodology, there is the possibility that performing a highly intensive task such as responding to multiple stimuli simultaneously leads to increased mental effort and thus, creates a sense of confusion for the user. For example, U2 velocity was correlated to ΔL in pen-only responses but was not correlated in test 3. There were some instances where a different set of correlations emerges in the response space in test 1 or 2 and test 3. For instance, for user 5 pressure which was positively correlated in test 2 (pen-only response) was found to be negatively correlated in simultaneous task 3.

It is recommended that future experiments should be planned with simpler modalities or ensure that the users might be given some preliminary training before the data collection. A qualitative study might also follow the quantitative nature of experiments on how the users could learn to express themselves better, this will help to design better experiments in the future.

8 Use Cases and Applications

The use cases discussed in this section are not the direct applications and usefulness of the proposed PEMC framework but a brief commentary (based on RQ 3) on the future application areas and benefits of finding better multimodal correlation in complex high-dimensional data.

8.1 Multimodal Correlations and Improved Machine Attention

Multimodal correlations provide a unique advantage that can enhance performance and lead to more accurate predictions by enabling machine learning systems to find more complex patterns from diverse sources of data. The development of algorithms that can learn from a large amount of data has advanced significantly recently in the field of machine learning. While this limits their ability to comprehend complex relationships and patterns, the majority of machine learning systems rely on unimodal data, such as images or text. Through the application of multimodal correlations, which require the analysis of data from various sources, machine learning systems have a unique opportunity to improve their accuracy and performance.

1. Increased Accuracy: Multimodal correlations enable machine learning algorithms to combine data from several sources, which increases accuracy. For example, a self-driving car that uses both visual and aural data is likely to be more accurate than a system that simply uses optical data. By combining many data sources, models can get a more detailed understanding of the underlying patterns and structures. For instance, while looking at images of items, integrating visual information with textual descriptions aids in improving object identification.

2. Robustness: By utilizing multimodal correlations, machine learning systems can be made more resilient to noise and outliers. By combining many data sources, the system may eliminate irrelevant information and focus on important details, reducing the impact of noise on the system's predictions.

3. Increased Generalization: Multimodal learning can aid in the generalization of machine learning systems. By learning from a range of data sources, models can get a deeper understanding of the underlying concepts and patterns, improving generalization performance. The models may then be better equipped to deal with unique situations and make precise predictions based on as-of-yet-unknown data.

4. Improved Interpretability: Multimodal learning can also make it simpler to comprehend machine learning models. By combining several data modalities, models can provide a more complete view of the data, which makes it easier to understand the relationships between the different parts. As a result, it will be simpler to identify the relevant traits and factors that affect the model's judgment. This can result in models that are more clear and understandable.

5. Better Human-Machine Interaction: Multimodal machine learning enables more intuitive and organic communication between humans and machines. For instance, speech recognition technology may improve the accuracy and authenticity of human-machine communication by reading both spoken words and facial expressions.

6. Scalability: Multimodal correlations enable machine learning algorithms to scale to larger datasets by using a variety of data. With more data sources available, the system may pick up linkages and patterns that are more complex, which enhances performance.

7. More Human-like perception: Machines can recognize patterns in multimodal data with high accuracy but human perception is complex and subjective. Human perception is influenced by past experiences, emotions, and attentional focus. Multimodal data used to train machines is often preprocessed to remove noise and bias. While machines may not fully replicate human perception, advances in AI may lead to more human-like models in the future.

In conclusion, the future of multimodal machine learning appears bright due to potential improvements in data sources, deep learning techniques, human-machine interaction, autonomous systems, and ethical issues. As technology develops, we could expect to see increasingly powerful and accurate models that handle and analyze diverse sorts of data more effectively.

8.2 Multimodal Correlations in Artistic Creations

In creative studies, the multimodal correlation between different types of data could expand the use of technology and interaction techniques. For example, our experiment which included colour data (HSL space), sound data, and pen gestures can be useful in several ways. For example:

1. Artistic Expression: Multimodal linkage can give artists additional methods to express themselves artistically, enhancing artistic expression. For instance, combining colour information in the HSL (Hue, Saturation, Lightness) colour space with music and pen gesture information can enable artists to produce works of art that harmoniously and synergistically mix visual, aural, and kinesthetic elements. This may lead to original and cutting-edge artistic expressions that challenge the limitations of conventional artistic mediums.

2. Creating Art Inspired by Synesthesia: Synesthesia is a condition where one sensory perception sets off another. Seeing colours when listening to music is one example, as is matching certain colours with certain sounds. Artists can create synesthesia-inspired art that blurs the lines between multiple senses and gives the spectator a multi-sensory experience by combining colour data, sound data, and pen motion data. This may present fresh opportunities for innovation and inquiry in the arts.

3. Interactive Installations: In interactive installations or immersive spaces, designers can use sound and colour correlations to produce more interesting and interactive experiences. As people travel around the area, a sound installation in a museum, for instance, might use colours to activate various sounds or musical compositions. By altering the room's colour scheme, visitors could engage with the work by altering the sounds being played.

4. Audio-Visual Performances: Multimodal correlation may also be employed in audio-visual presentations like live performances or concerts. Performers can give the audience a more immersive and synesthetic experience by matching colours to similar sounds. A visual artist could, for instance, use the HSL colour space to map the visualization's colours to the music's accompanying sounds as a musician performs a live concert.

5. Augmented Reality: Multimodal correlation can also be utilized in augmented reality apps to make encounters more interesting and interactive. The HSL colour space, for instance, might be used by augmented reality software to map colours in the real world to corresponding sounds or music. The user would hear a musical composition that reflected the colours all around them as they moved around the room.

In creative studies, the multimodal correlation may incorporate individual aesthetic preferences and subjective interpretation. Depending on their artistic techniques, tastes, and goals, different artists may interpret the correlation in different ways and produce various works of art. Because of this subjectivity, it can be difficult to build reliable and objective correlations between various data modalities, and the outcomes may differ depending on the artist's point of view. The use of data in creative research, such as colour, sound, and pen gesture data, may give rise to ethical questions about privacy, consent, and data exploitation. The ethical ramifications of using data from multiple sources must be considered by artists and researchers, and they must make sure they abide by all applicable laws, rules, and ethical principles. There should be a serious examination of ethical issues. In conclusion, discovering multimodal associations in colour, sound, and pen data or other modalities might be advantageous for creative studies in a number of ways, including boosting artistic expression, enabling interactive artworks, and generating synesthesia-inspired art. In order to successfully integrate many data modalities in creative research, it also presents difficulties in terms of technological implementation, subjective interpretation, and ethical considerations.

9 Future Work and Conclusion

The preliminary correlations that emerged during the experiment shown in the thesis could be followed up with a more controlled experiment in the future. An example of such a controlled experiment is the following:

1. Goal: Design an experiment to find out multimodal correlation in human sensory data, specifically between audio responses and pen gestures

2. Select participants: Participants should be individuals who are comfortable with using a pen and have no hearing or motor impairments that may interfere with the experiment.

3. Create stimuli: Create a set of audio prompts that participants will respond to using a pen on a tablet. The prompts can be questions, statements, or even music. Ensure that the audio stimuli are clear and consistent across all participants.

4. Record data: Record both the audio responses and the pen gestures using appropriate equipment. This could include microphones to capture the audio, and tablets or touchscreens to capture the pen gestures.

5. Analyze data Using Correlation Analyser: The correlation analyser could be used to find multimodal correlations. Since the modalities are controlled, more concrete correlations should emerge.

6. Interpret the results: Look at the results obtained and interpret the correlation values. If the correlation is high, it means that there is a strong relationship between the audio responses and the pen gestures, while a low correlation means that the two modalities are not closely related.

7. Evaluate the experiment: Evaluate the experiment and see if there are any limitations or biases that may have affected the results. For instance, the experiment may have a small sample size or limited audio prompts, which could impact the correlation values.

8. Realization of Universal Correlation: To uncover generalized multimodal correlations from human data, a large dataset of multimodal human data from diverse individuals and settings is necessary. The dataset should have sufficient variation to account for individual differences and contextual factors. A sample size of at least 30 is typically recommended for statistical power. Validation of the correlations across multiple populations and settings is crucial to ensure the existence of universal correlations.

9. Other advanced algorithms or variants of proposed techniques could be used to used in the CA method. For example, Variants of LLE could be tried for dimensionality reduction. Other correlation techniques such as CCA could also be used along with Spearman's and Pearson's methods.

Based on the results from the 6 section, future experiments might be focused on the following features in the stimulus and response spaces:

1. PV and Power (audio space) and peaks (pen responses).
2. PV (audio space) and peaks, total distance and pressure (pen responses).
3. PV, Chroma energy, spectral contrast (audio response) and average velocity, total distance (pen gesture response)
4. MFCC, Spec Bandwidth, PVFT, spec contrast, RMS (audio response) and average velocity, total distance (pen responses)
5. MFCC coefficients (audio responses) and Pressure, Average velocity, and total distance (pen responses)

Deciphering multimodal correlation and understanding its impact on artistic and creative generation has the potential to unlock new insights into the human brain's workings and its relationship with artistic and creative processes. This thesis proposed a framework to find multimodal correlations in complex sensory data collected from humans and studies how to efficiently find the intrinsic correlations from the data. Many weaker and stronger correlations emerged in the response and stimulus spaces in the data collected from human participants and potential future experiments were suggested to test for universal correlations based on human behaviour and response to given stimuli.

10 Acknowledgement

This work was done in collaboration with MFA. Jaana Okulov, Doctoral Candidate, Aalto University. The goal of the experiments was to find correlations from human behavioural data using statistical and machine learning techniques that could be further used in Jaana's work. I would like to thank Jaana for the help with data collection from the participants. Finally, I would like to thank Tassu for his support during these years and guidance throughout the process.

References

[1] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022.

[2] Midjourney. https://www.midjourney.com/home/.

[3] Tadas Baltrusaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *CoRR*, abs/1705.09406, 2017.

[4] Charles Spence. Crossmodal correspondences: A tutorial review. *Attention, perception psychophysics*, 73:971–95, 05 2011.

[5] Paul Bertelson, Beatrice Gelder, Charles Spence, and Jon Driver. *The Psychology of Multimodal Perception*, pages 141–177. 04 2004.

[6] Odbert H. S. Osgood C. E. Karwoski, T. F. Studies in synesthetic thinking: Ii. the rôle of form in visual responses to music. *Journal of General Psychology*, 26:199–222, 1942.

[7] Lawrence Marks. On associations of light and sound: The mediation of brightness, pitch, and loudness. *The American journal of psychology*, 87:173–88, 03 1974.

[8] Wolfgang Köhler. Gestalt psychology. *New York: Liveright*, 1929.

[9] Brown V. LaBerge, D. Theory of attentional operations in shape identification. *Psychological Review*, 96(1), 1989.

[10] Yu-Han Chen, J Christopher Edgar, Tom Holroyd, Jürgen Dammers, Heike Thönnessen, Timothy P L Roberts, and Klaus Mathiak. Neuromagnetic oscillations to emotional faces and prosody. *The European journal of neuroscience*, 31(10):1818—1827, May 2010.

[11] Störmer VS. Orienting spatial attention to sounds enhances visual processing. *Curr Opin Psychol.*, 2019 Oct.

[12] B. Zhou, A. Khosla, Lapedriza. A., A. Oliva, and A. Torralba. Learning Deep Features for Discriminative Localization. *CVPR*, 2016.

[13] Jaana Okulov. Quantifying qualia. *C. Misselhorn & Klein M. (eds.) Emotional Machines: Perspectives from Affective Computing and Emotional Human-Machine Interaction, Futures of Technology, Science and Society (In Press)*, 2023.

[14] Jaana Okulov. Artificial aesthetics and aesthetic machine attention. *AM Journal of Art and Media Studies*, (29):13–28, October 2022.

[15] Burr D Alais D1. The ventriloquist effect results from near-optimal bimodal integration. *Curr Biol*, 14(3):257–62, 2004.

[16] Bülthoff HH Ernst MO. Merging the senses into a robust percept. *Trends Cogn Sci.*, 2004:162–9, 2004.

[17] Hubbard E. M. Ramachandran, V. S. Synaesthesia–a window into perception, thought and language. *Journal of Consciousness Studies*, 8(12), 2001.

[18] Quinn M. Ausubel D. P. Simpson, R. H. Synesthesia in children: Association of colors with pure tone frequencies. *The Journal of Genetic Psychology: Research and Theory on Human Development*, 89, 1956.

[19] J C Stevens and L E Marks. Cross-modality matching of brightness and loudness. *Proceedings of the National Academy of Sciences*, 54(2):407–411, 1965.

[20] J. T. Cowles. An experimental study of the pairing of certain auditory and visual stimuli. *Journal of Experimental Psychology*, 18(4):461–469, 1935.

[21] Michail N. Giannakos, Kshitij Sharma, Ilias O. Pappas, Vassilis Kostakos, and Eduardo Velloso. Multimodal data as a means to understand the learning experience. *International Journal of Information Management*, 48:108–119, 2019.

[22] Dr. Cheryl Olman. Introduction to sensation and perception. *Introduction to Sensation and Perception*, 1, 2022.

[23] G Rajesh and Ashvini Chaturvedi. Correlation analysis of multimodal sensor data in environmental sensor networks. In *2019 IEEE International Conference on Advanced Networks and Telecommunications Systems (ANTS)*. IEEE, December 2019.

[24] Ernst M. O. Helbig, H. B. Knowledge about a common source can promote visual–haptic integration. *Perception*, 36(10), 2007.

[25] Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, Aug 2013.

[26] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y. Ng. Multimodal deep learning. In *ICML*, pages 689–696, 2011.

[27] Wanli Ouyang, Xiao Chu, and Xiaogang Wang. Multi-source deep learning for human pose estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[28] Daixin Wang, Peng Cui, Mingdong Ou, and Wenwu Zhu. Deep multimodal hashing with orthogonal regularization. In Qiang Yang and Michael J. Wooldridge, editors, *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25-31, 2015*, pages 2291–2297. AAAI Press, 2015.

[29] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[30] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate, 2014.

[31] Qing-Yuan Jiang and Wu-Jun Li. Deep cross-modal hashing. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[32] H. Hotelling. Relations Between Two Sets of Variates. *Biometrika*, 28(3-4):321–377, December 1936.

[33] S. Szedmak D. R. Hardoon and J. Shawe-Taylor. Canonical correlation analysis; an overview with application to learning methods. *Neural computation*, pages 2639–64, 2016.

[34] Benjamin Klein, Guy Lev, Gil Sadeh, and Lior Wolf. Fisher vectors derived from hybrid gaussian-laplacian mixture models for image annotation, 2014.

[35] E. Coviello G. Doyle G. R. Lanckriet R. Levy N. Rasiwasia, J. Costa Pereira and N. Vasconcelos. A new approach to cross-modal multimedia retrieval. *ACMMM*, 2010.

[36] M. E. Sargin, Y. Yemez, E. Erzin, and A. M. Tekalp. Audiovisual synchronization and fusion using canonical correlation analysis. *Trans. Multi.*, 9(7):1396–1403, November 2007.

[37] Malcolm Slaney and Michele Covell. Facesync: A linear operator for measuring synchronization of video facial images and audio tracks. In T. Leen, T. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems*, volume 13. MIT Press, 2000.

[38] A. J. Hunt and A. W. Black. Unit selection in a concatenative speech synthesis system using a large speech database. In *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, volume 1, pages 373–376 vol. 1, 1996.

[39] Takashi Masuko, Takao Kobayashi, Masatsune Tamura, Jun Masubuchi, and Keiichi Tokuda. Text-to-visual speech synthesis based on parameter generation from hmm. *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP '98 (Cat. No.98CH36181)*, 6:3745–3748 vol.6, 1998.

[40] T. Tamura A. Kojima and K. Fukunaga. Natural language description of human activities from video images based on concept hierarchy of actions. *IJCV*, 2002.

[41] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3156–3164, 2015.

[42] Vicente Ordonez, Girish Kulkarni, and Tamara L. Berg. Im2text: Describing images using 1 million captioned photographs. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 1143–1151. Curran Associates, Inc., 2011.

[43] P. Young M. Hodosh and J. Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *JAIR*, 47, 2013.

[44] Andrej Karpathy, Armand Joulin, and Li Fei-Fei. Deep fragment embeddings for bidirectional image sentence mapping. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, page 1889–1897, Cambridge, MA, USA, 2014. MIT Press.

[45] Andrew Owens, Phillip Isola, Josh McDermott, Antonio Torralba, Edward H. Adelson, and William T. Freeman. Visually indicated sounds, 2016.

[46] Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. WaveNet: A Generative Model for Raw Audio. In *Proc. 9th ISCA Workshop on Speech Synthesis Workshop (SSW 9)*, page 125, 2016.

[47] Elman Mansimov, Emilio Parisotto, Jimmy Ba, and Ruslan Salakhutdinov. Generating images from captions with attention. 11 2015.

[48] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1060–1069, New York, New York, USA, 20–22 Jun 2016. PMLR.

[49] William Chan, Navdeep Jaitly, Quoc V. Le, and Oriol Vinyals. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *ICASSP*, 2016.

[50] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6):82–97, 2012.

[51] Yuchen Fan, Yao Qian, Feng-Long Xie, and Frank K. Soong. Tts synthesis with bidirectional lstm based recurrent neural networks. In *INTERSPEECH*, 2014.

[52] Anna Rohrbach, Marcus Rohrbach, and Bernt Schiele. The long-short story of movie description. In Juergen Gall, Peter Gehler, and Bastian Leibe, editors, *Pattern Recognition*, pages 209–221, Cham, 2015. Springer International Publishing.

[53] Patel P. Rau M. A. Mason B. Nowak R. Rogers T. T. Zhu J. Sen, A. For teaching perceptual fluency, machines beat human experts. *Cogn Sci.*, 2018, 2018.

[54] Christina M. Funke, Judy Borowski, Karolina Stosio, Wieland Brendel, Thomas S. A. Wallis, and Matthias Bethge. The notorious difficulty of comparing human and machine perception. *CoRR*, abs/2004.09406, 2020.

[55] Thomas Nagel. What is it like to be a bat? *The Philosophical Review*, 83(4):435–450, 1974.

[56] Pentti O. A. Haikonen. QUALIA AND CONSCIOUS MACHINES. *International Journal of Machine Consciousness*, 01(02):225–234, December 2009.

[57] Philipos C. Loizou, Yi Hu, Ruth Litovsky, Gongqiang Yu, Robert Peters, Jennifer Lake, and Peter Roland. Speech recognition by bilateral cochlear implant users in a cocktail-party setting. *The Journal of the Acoustical Society of America*, 125(1):372–383, January 2009.

[58] Roman V. Yampolskiy. Detecting Qualia in Natural and Artificial Agents. *arXiv e-prints*, page arXiv:1712.04020, December 2017.

[59] Giulio Tononi. *BMC Neuroscience*, 5(1):42, 2004.

[60] Eric Langford, Neil Schwertman, and Margaret Owens. Is the property of being positively correlated transitive? *The American Statistician*, 55(4):322–325, 2001.

[61] Ian T. Jolliffe and Jorge Cadima. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202, April 2016.

[62] Sam T. Roweis and Lawrence K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, December 2000.

[63] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction, 2020.

[64] Trevor Cox and Michael Cox. *Multidimensional Scaling*. Chapman and Hall/CRC, September 2000.

[65] George H. Joblove and Donald Greenberg. Color spaces for computer graphics. In *Proceedings of the 5th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '78, page 20–25, New York, NY, USA, 1978. Association for Computing Machinery.

A Appendix

This section shows the elbow criterion plots and the 2-D latent space representation of audio and pen data for 3 test scenarios according to the PEMC framework.

A.1 User 2

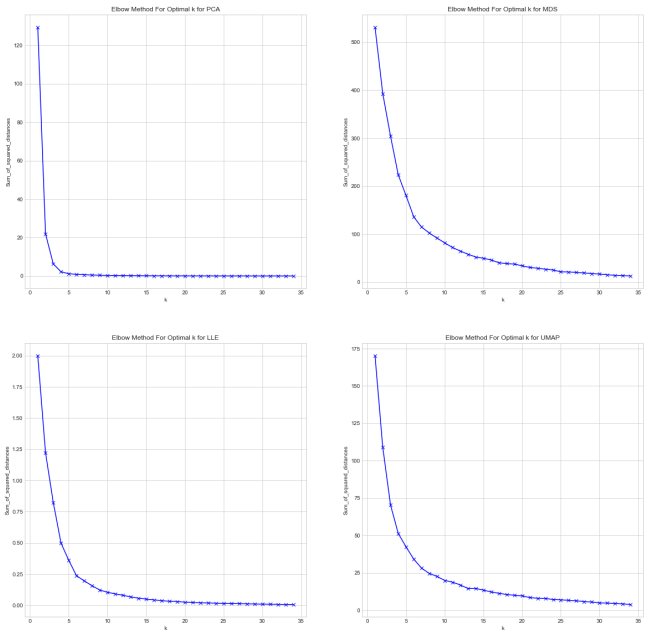


Figure A1: Elbow plots for finding optimal K value for clustering, audio/verbal only response, U2. The x axis and y axis shows the k value and the sum of the squared distance (between each point and the centroid in a cluster) respectively for each subplot.

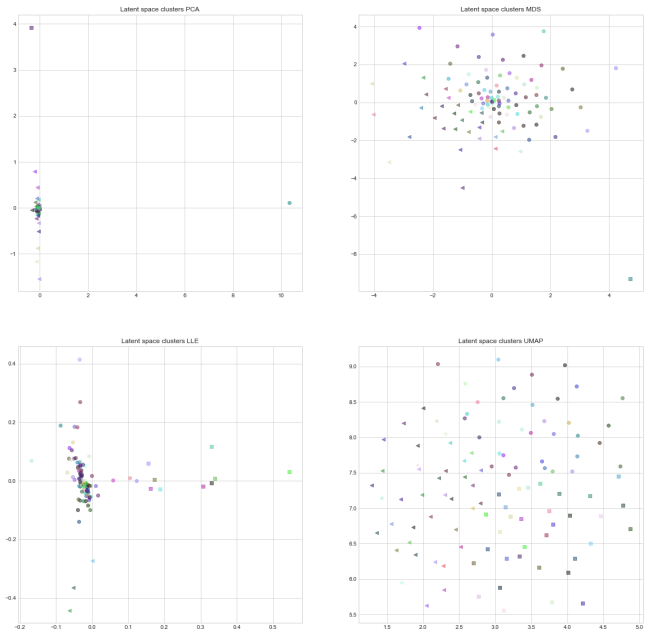


Figure A2: Latent Space Cluster Visualization of latent feature space, audio/verbal only response, U2

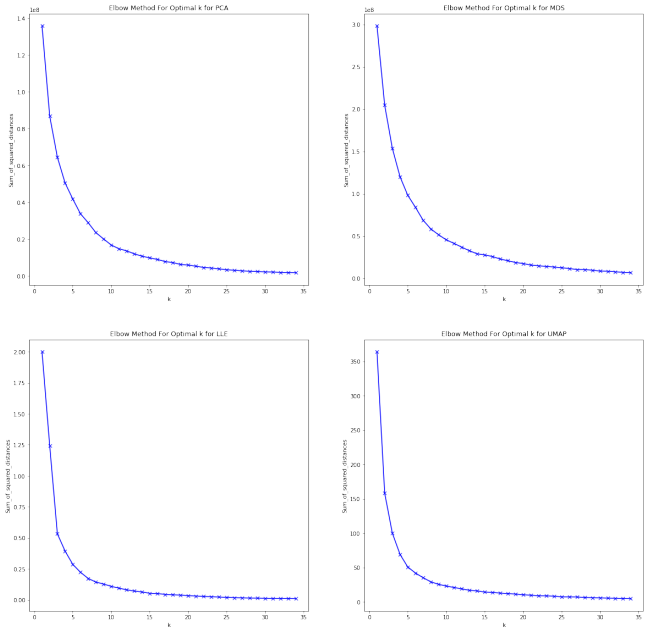


Figure A3: Elbow plots for finding optimal K value for clustering, pen gesture only response, U2. The x axis and y axis shows the k value and the sum of the squared distance (between each point and the centroid in a cluster) respectively for each subplot.

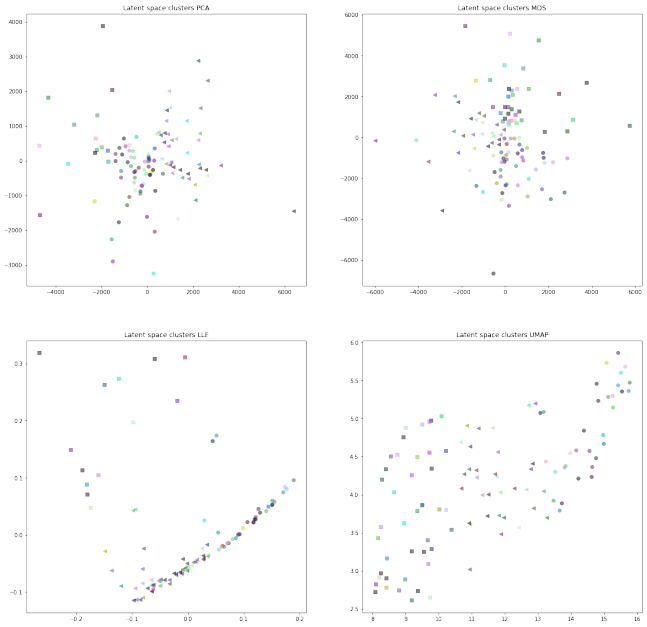


Figure A4: Latent Space Cluster Visualization of latent feature space, pen gesture-only response, U2

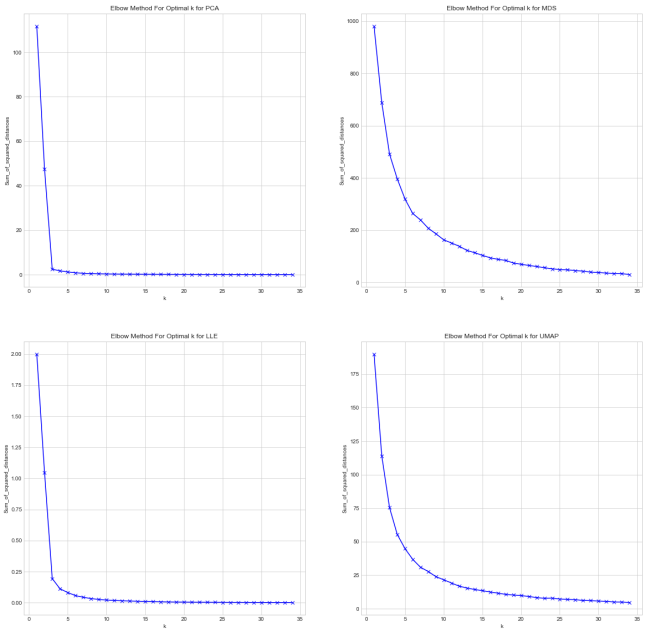


Figure A5: Elbow plots for finding optimal K value for clustering, audio/verbal response (simultaneous pen and audio response), U2. The x axis and y axis shows the k value and the sum of the squared distance (between each point and the centroid in a cluster) respectively for each subplot.

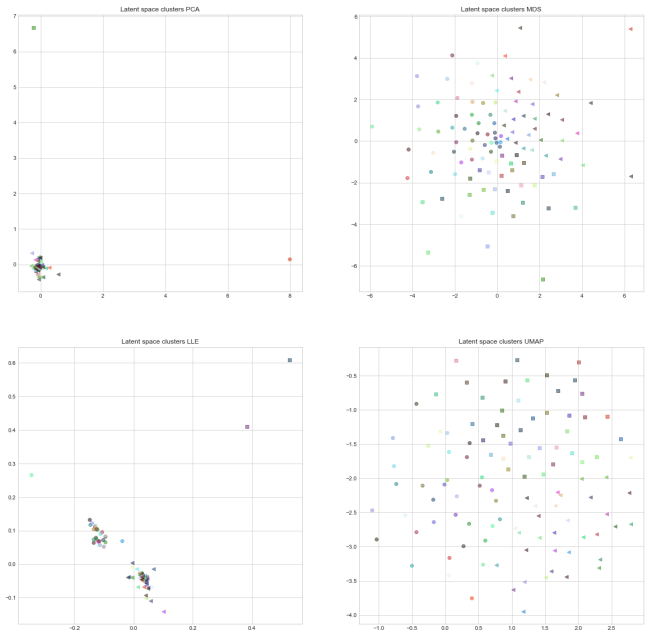


Figure A6: Latent Space Cluster Visualization of latent feature space, audio/verbal response (simultaneous pen and audio response), U2

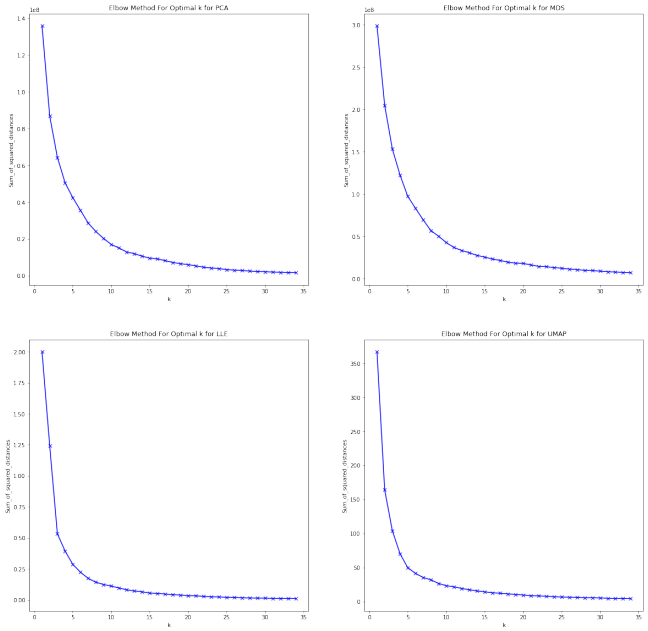


Figure A7: Elbow plots for finding optimal K value for clustering, pen response (simultaneous pen and audio response), U2. The x axis and y axis shows the k value and the sum of the squared distance (between each point and the centroid in a cluster) respectively for each subplot.

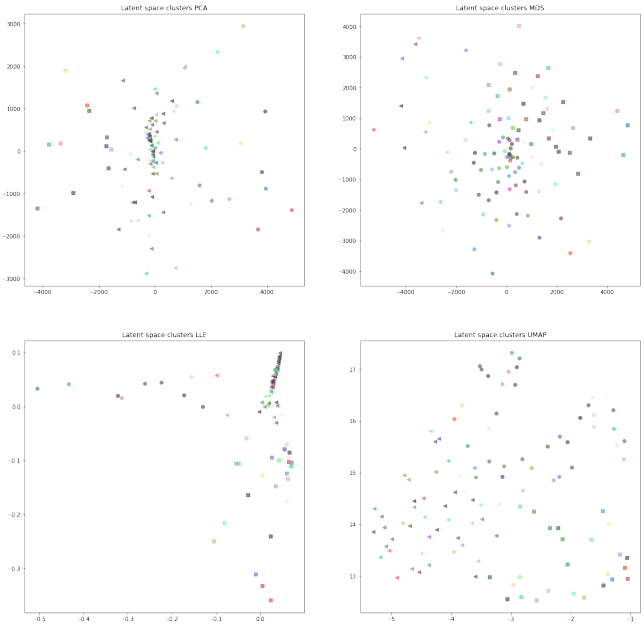


Figure A8: Latent Space Cluster Visualization of latent feature space, pen response (simultaneous pen and audio response), U2

A.2 User 3

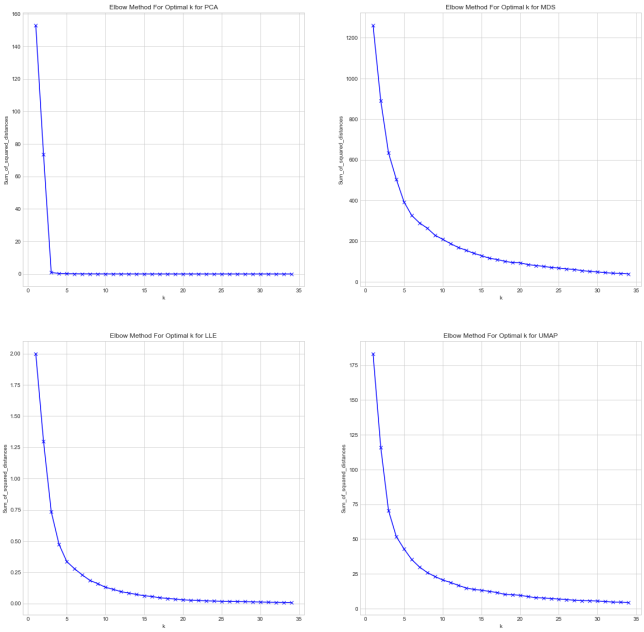


Figure A9: Elbow plots for finding optimal K value for clustering, audio/verbal only response, U3. The x axis and y axis shows the k value and the sum of the squared distance (between each point and the centroid in a cluster) respectively for each subplot.

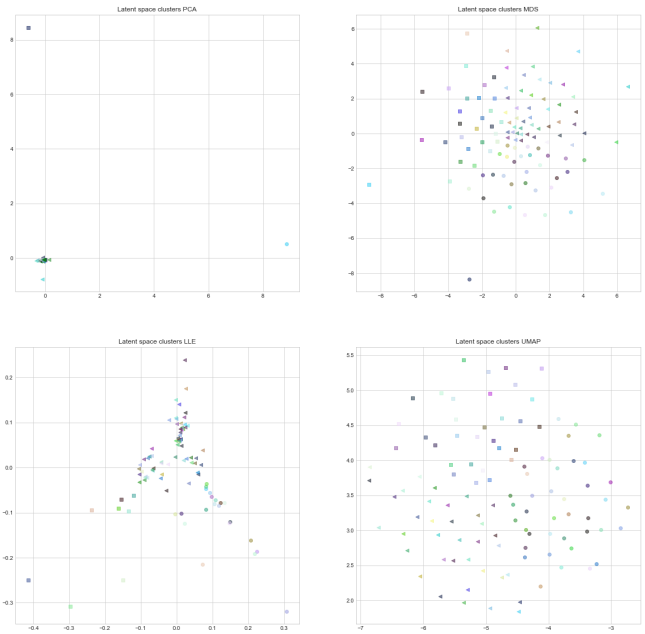


Figure A10: Latent Space Cluster Visualization of latent feature space, audio/verbal only response, U3

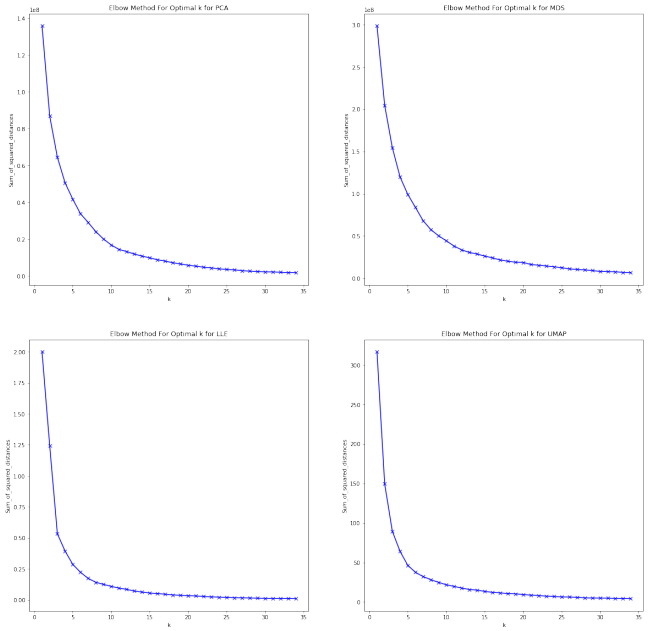


Figure A11: Elbow plots for finding optimal K value for clustering, pen gesture only response, U3. The x axis and y axis shows the k value and the sum of the squared distance (between each point and the centroid in a cluster) respectively for each subplot.

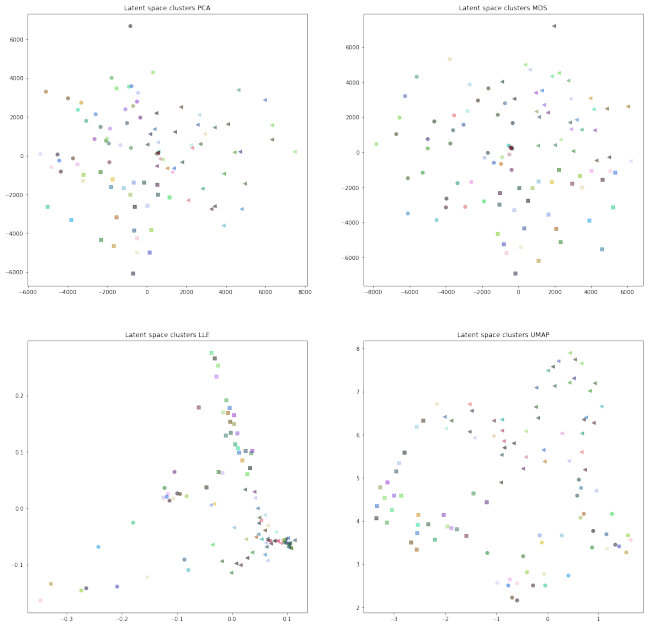


Figure A12: Latent Space Cluster Visualization of latent feature space, pen gesture-only response, U3

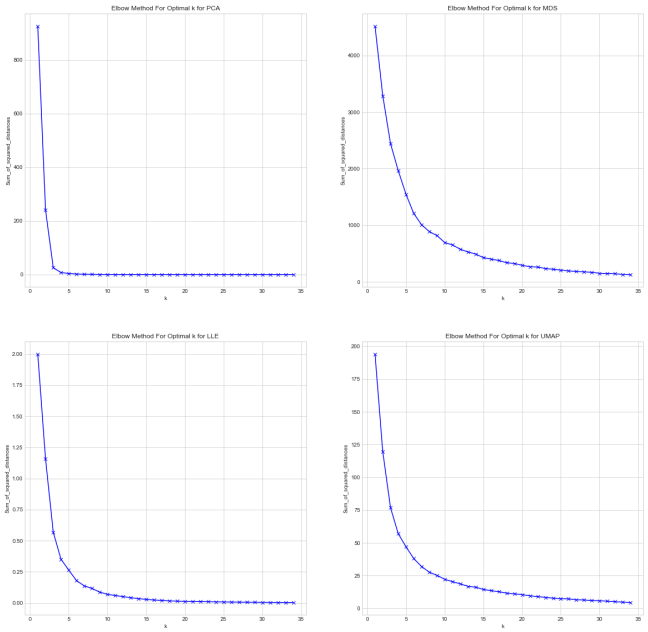


Figure A13: Elbow plots for finding optimal K value for clustering, audio/verbal response (simultaneous pen and audio response), U3. The x axis and y axis shows the k value and the sum of the squared distance (between each point and the centroid in a cluster) respectively for each subplot.

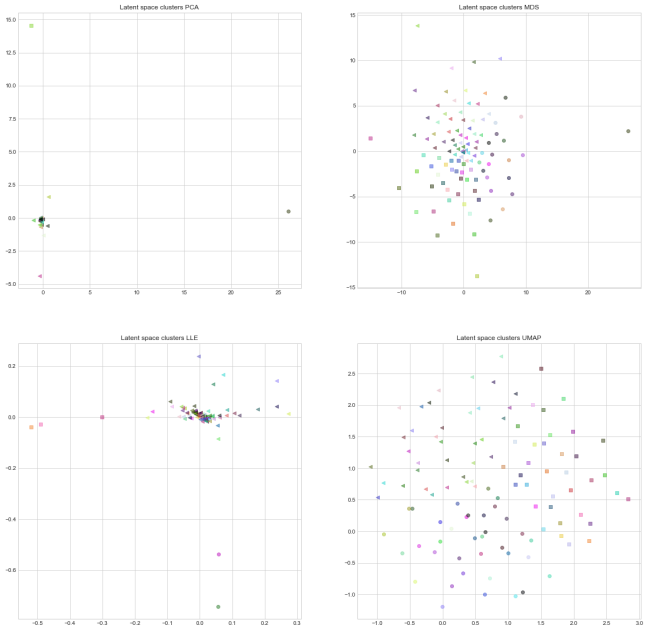


Figure A14: Latent Space Cluster Visualization of latent feature space, audio/verbal response (simultaneous pen and audio response), U3

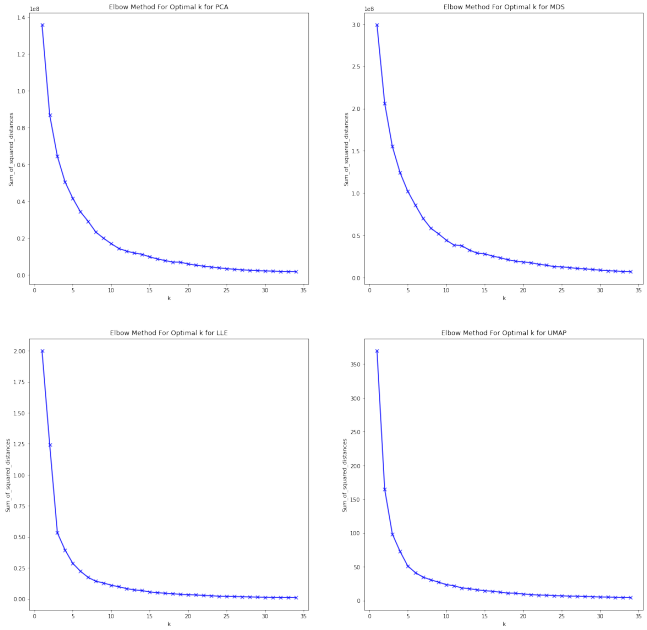


Figure A15: Elbow plots for finding optimal K value for clustering, pen response (simultaneous pen and audio response), U3. The x axis and y axis shows the k value and the sum of the squared distance (between each point and the centroid in a cluster) respectively for each subplot.

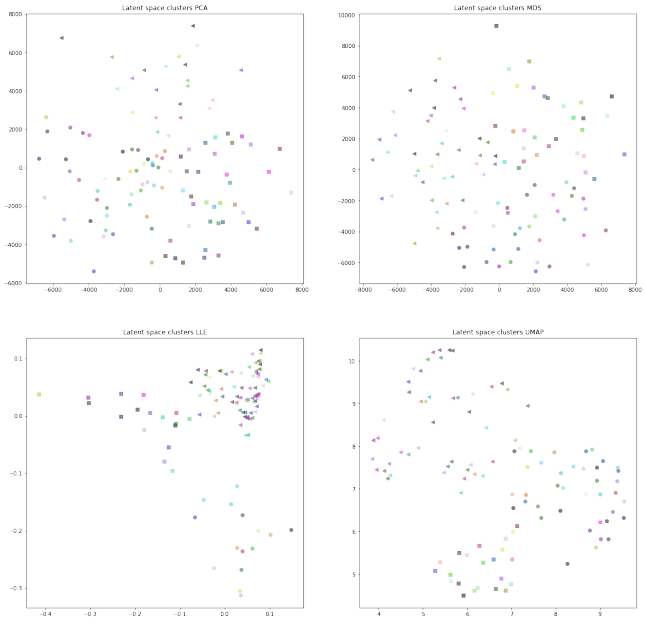


Figure A16: Latent Space Cluster Visualization of latent feature space, pen response (simultaneous pen and audio response), U3

A.3 User 4

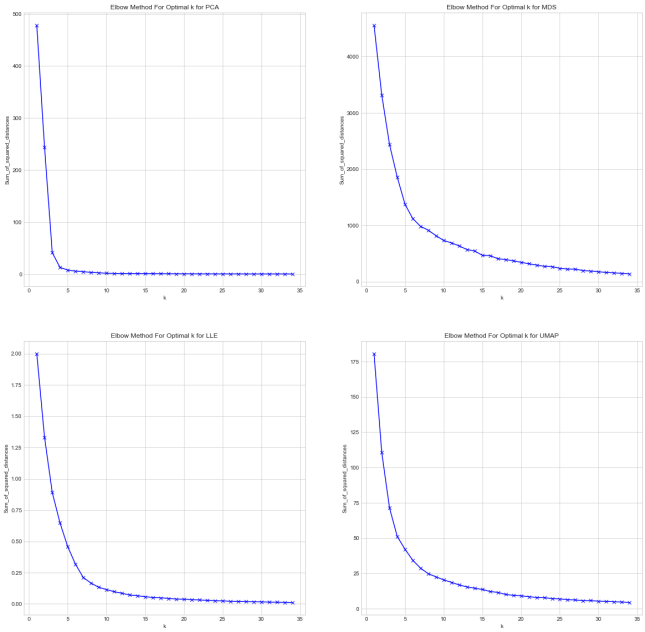


Figure A17: Elbow plots for finding optimal K value for clustering, audio/verbal only response, U4. The x axis and y axis shows the k value and the sum of the squared distance (between each point and the centroid in a cluster) respectively for each subplot.

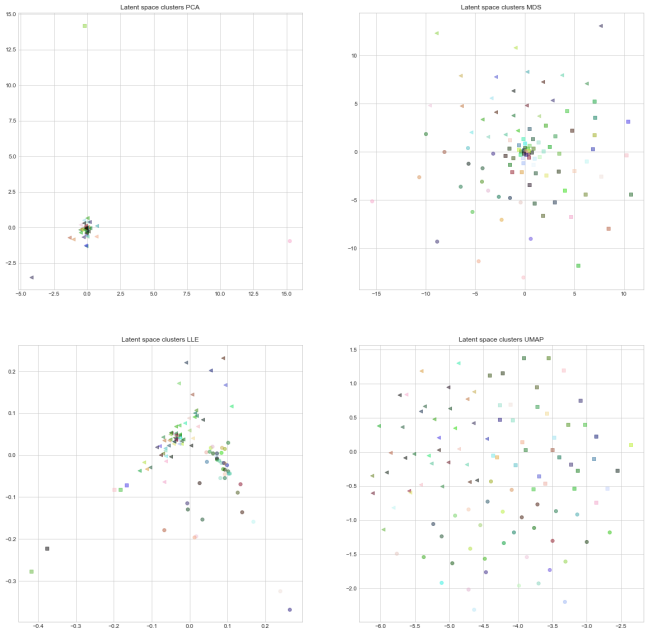


Figure A18: Latent Space Cluster Visualization of latent feature space, audio/verbal only response, U4

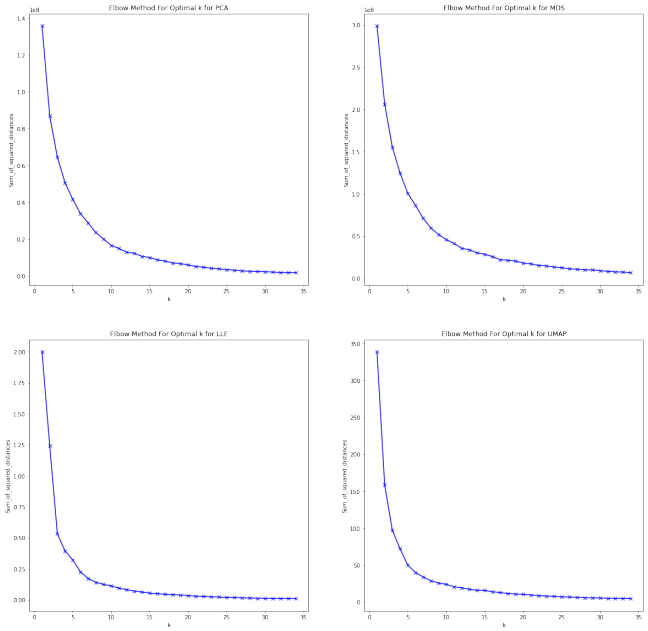


Figure A19: Elbow plots for finding optimal K value for clustering, pen gesture only response, U4. The x axis and y axis shows the k value and the sum of the squared distance (between each point and the centroid in a cluster) respectively for each subplot.

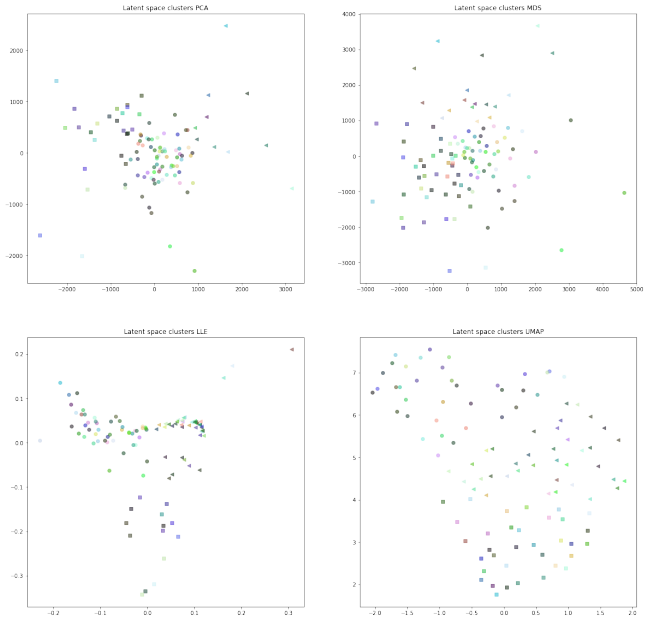


Figure A20: Latent Space Cluster Visualization of latent feature space, pen gesture-only response, U4

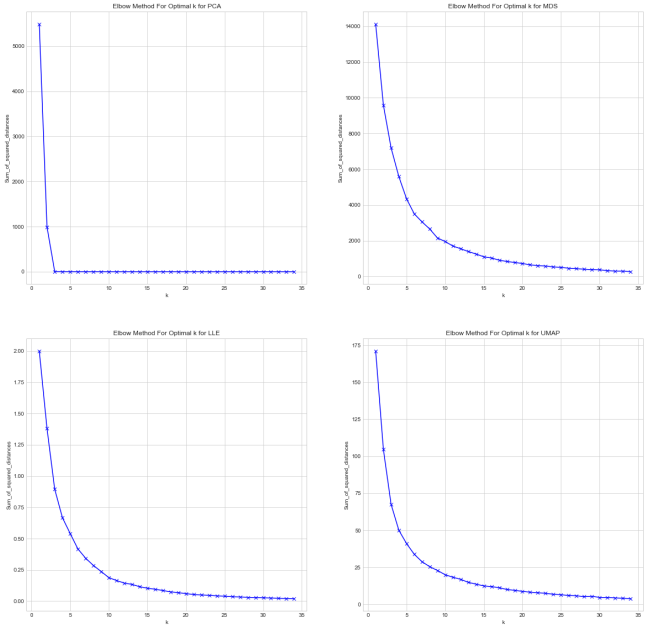


Figure A21: Elbow plots for finding optimal K value for clustering, audio/verbal response (simultaneous pen and audio response), U4. The x axis and y axis shows the k value and the sum of the squared distance (between each point and the centroid in a cluster) respectively for each subplot.

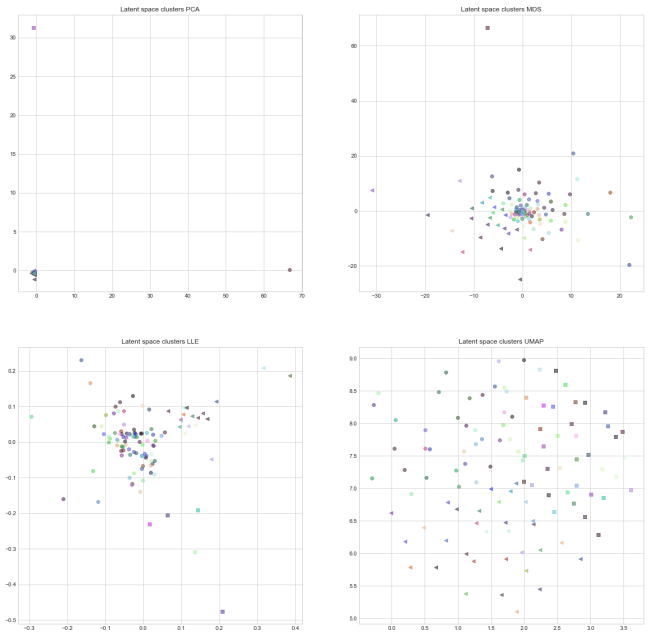


Figure A22: Latent Space Cluster Visualization of latent feature space, audio/verbal response (simultaneous pen and audio response), U4

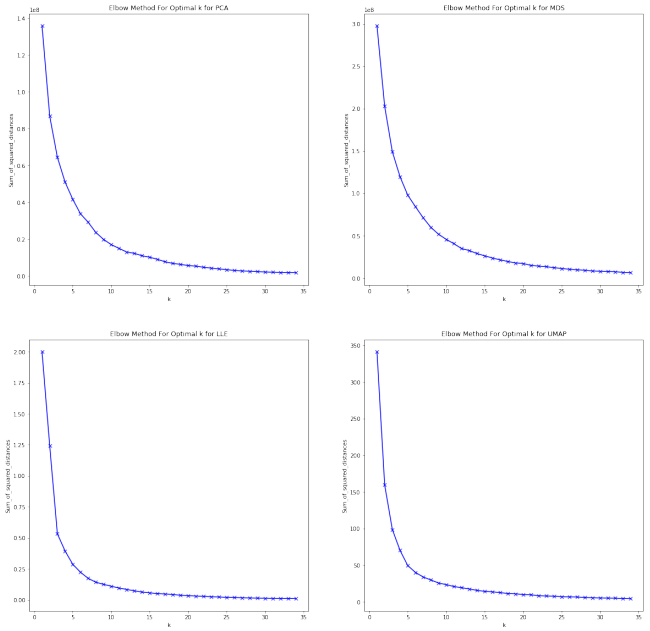


Figure A23: Elbow plots for finding optimal K value for clustering, pen response (simultaneous pen and audio response), U4. The x axis and y axis shows the k value and the sum of the squared distance (between each point and the centroid in a cluster) respectively for each subplot.

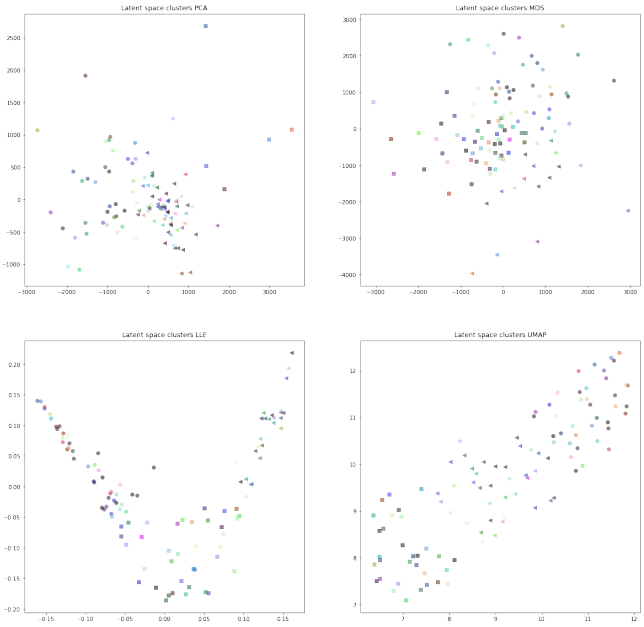


Figure A24: Latent Space Cluster Visualization of latent feature space, pen response (simultaneous pen and audio response), U4

A.4 User 5

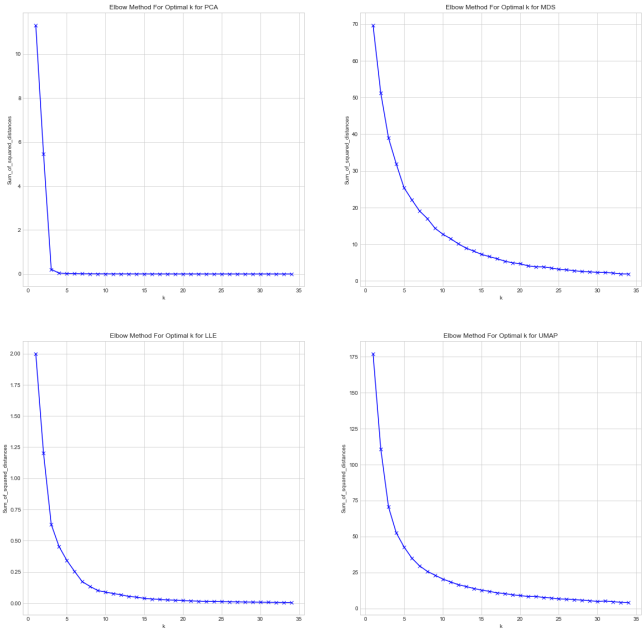


Figure A25: Elbow plots for finding optimal K value for clustering, audio/verbal only response, U5. The x axis and y axis shows the k value and the sum of the squared distance (between each point and the centroid in a cluster) respectively for each subplot.

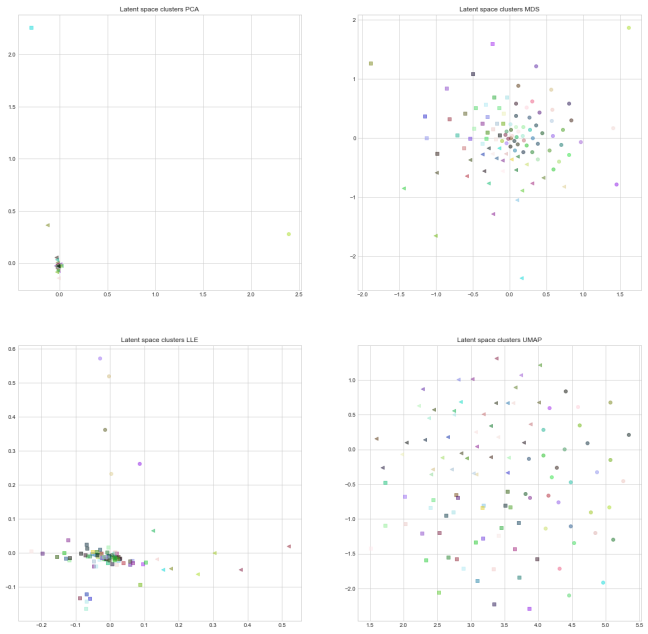


Figure A26: Latent Space Cluster Visualization of latent feature space, audio/verbal only response, U5

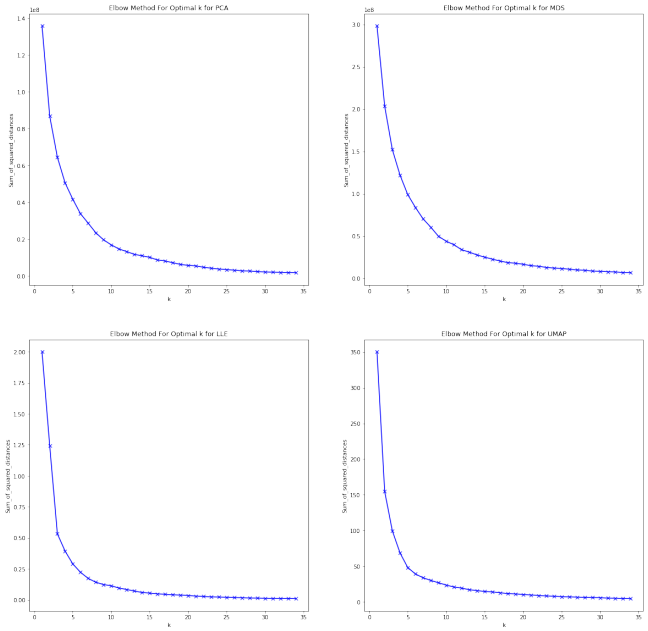


Figure A27: Elbow plots for finding optimal K value for clustering, pen gesture only response, U5. The x axis and y axis shows the k value and the sum of the squared distance (between each point and the centroid in a cluster) respectively for each subplot.

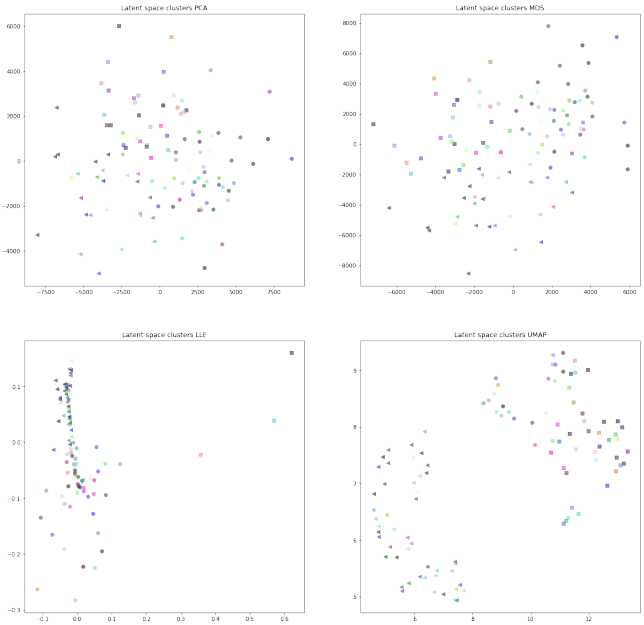


Figure A28: Latent Space Cluster Visualization of latent feature space, pen gesture-only response, U5

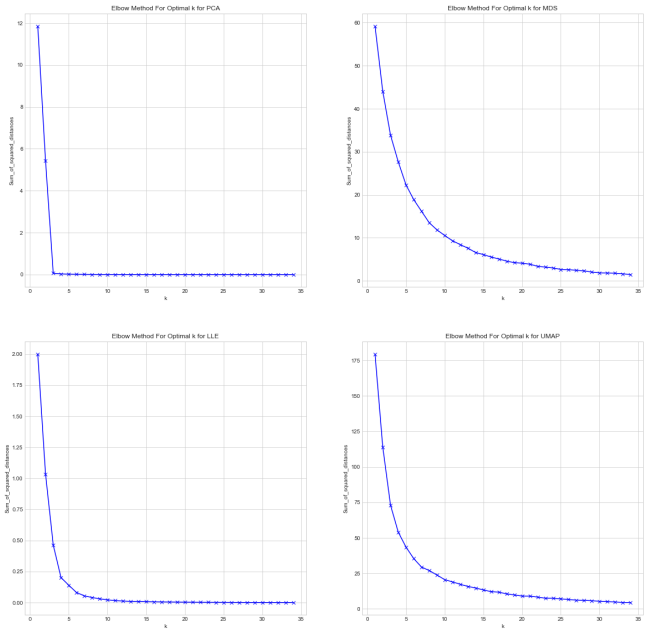


Figure A29: Elbow plots for finding optimal K value for clustering, audio/verbal response (simultaneous pen and audio response), U5. The x axis and y axis shows the k value and the sum of the squared distance (between each point and the centroid in a cluster) respectively for each subplot.

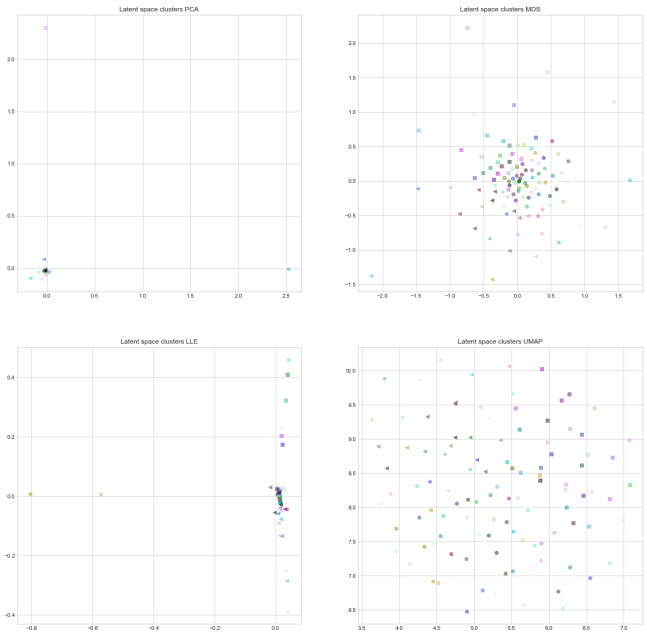


Figure A30: Latent Space Cluster Visualization of latent feature space, audio/verbal response (simultaneous pen and audio response), U5

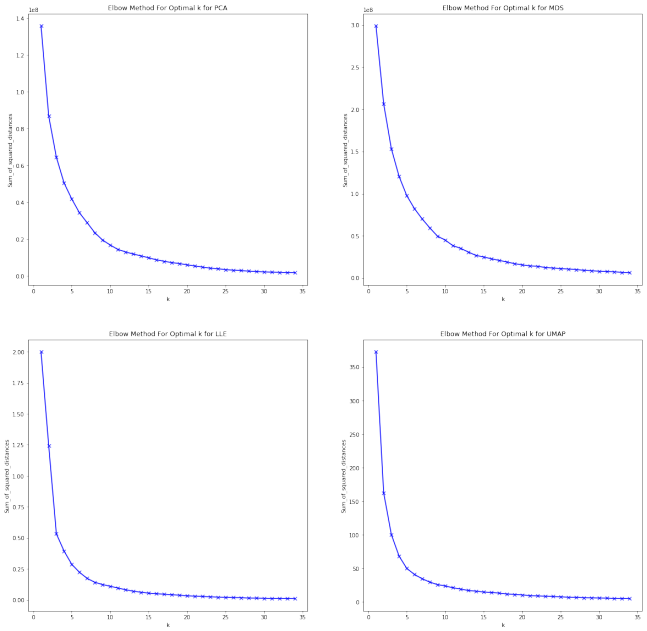


Figure A31: Elbow plots for finding optimal K value for clustering, pen response (simultaneous pen and audio response), U5. The x axis and y axis shows the k value and the sum of the squared distance (between each point and the centroid in a cluster) respectively for each subplot.

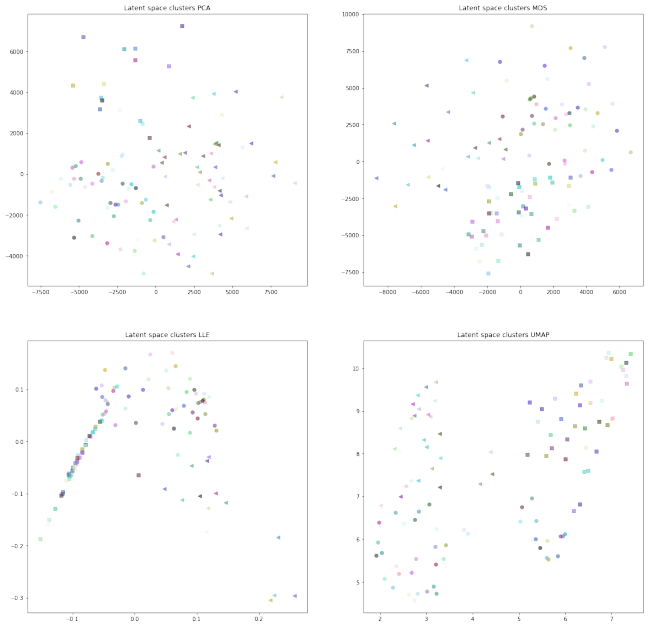


Figure A32: Latent Space Cluster Visualization of latent feature space, pen response (simultaneous pen and audio response), U5

A.5 User 6

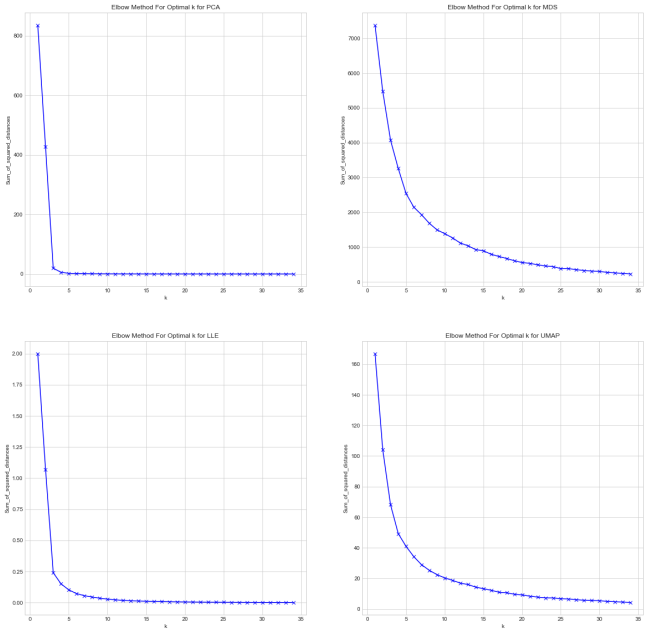


Figure A33: Elbow plots for finding optimal K value for clustering, audio/verbal only response, U6. The x axis and y axis shows the k value and the sum of the squared distance (between each point and the centroid in a cluster) respectively for each subplot.

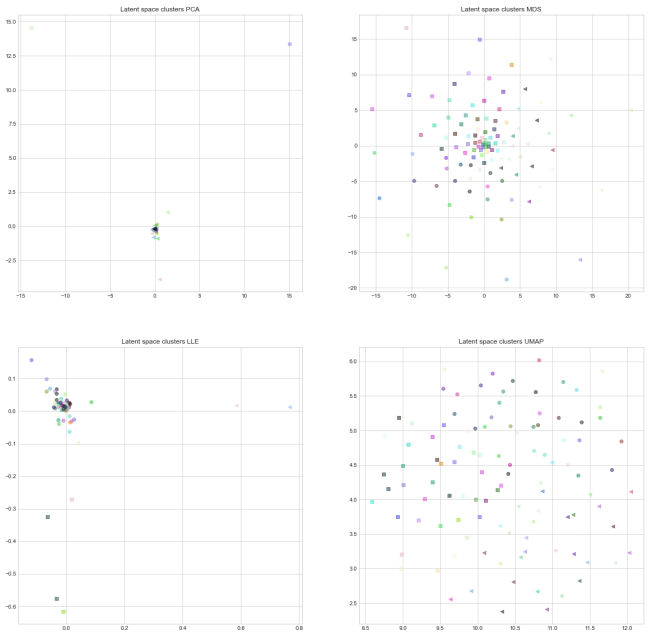


Figure A34: Latent Space Cluster Visualization of latent feature space, audio/verbal only response, U6

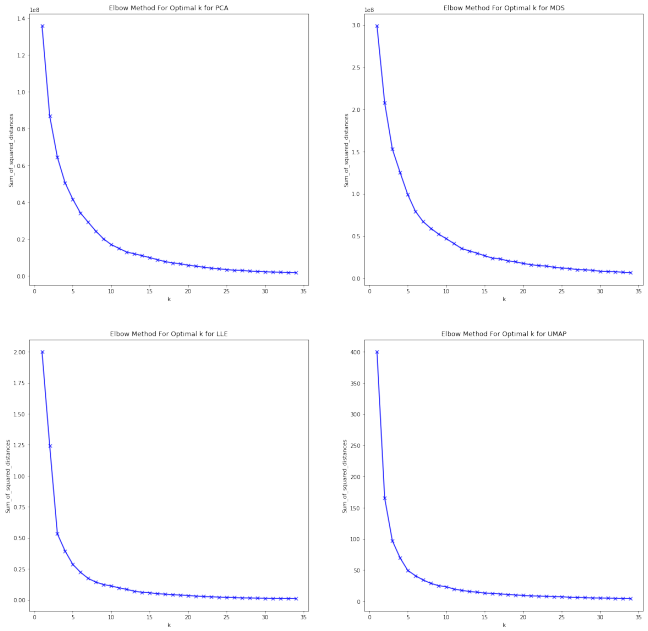


Figure A35: Elbow plots for finding optimal K value for clustering, pen gesture only response, U6. The x axis and y axis shows the k value and the sum of the squared distance (between each point and the centroid in a cluster) respectively for each subplot.

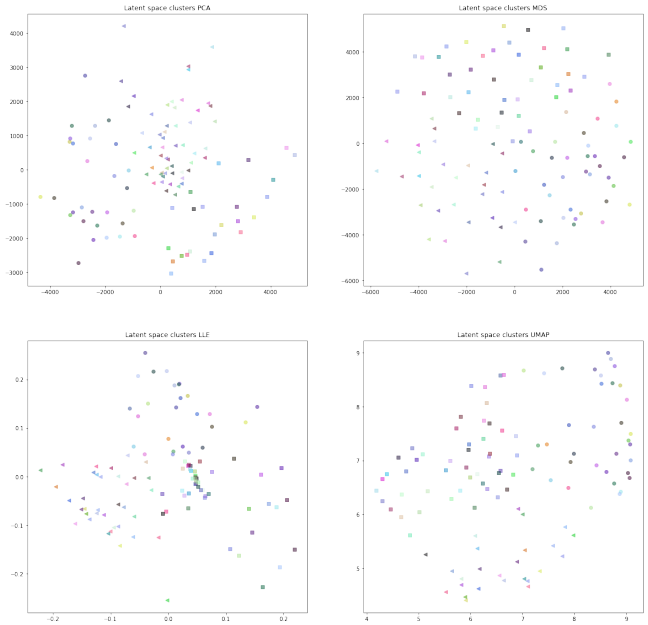


Figure A36: Latent Space Cluster Visualization of latent feature space, pen gesture-only response, U6

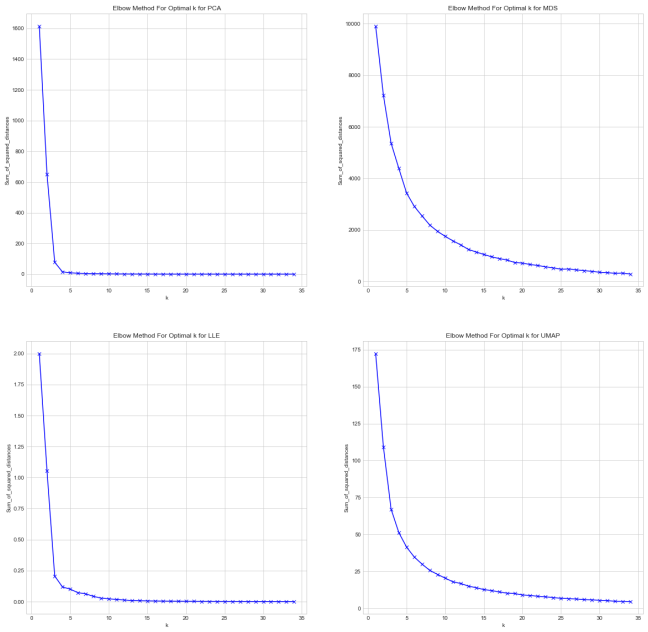


Figure A37: Elbow plots for finding optimal K value for clustering, audio/verbal response (simultaneous pen and audio response), U6. The x axis and y axis shows the k value and the sum of the squared distance (between each point and the centroid in a cluster) respectively for each subplot.

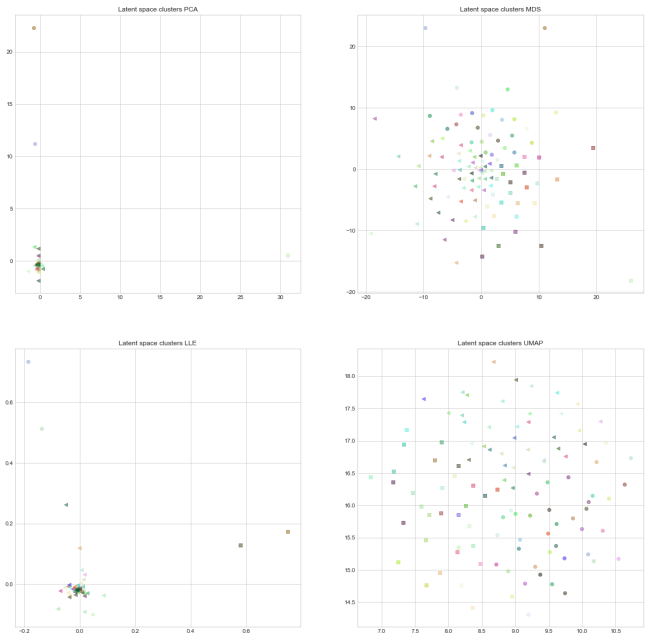


Figure A38: Latent Space Cluster Visualization of latent feature space, audio/verbal response (simultaneous pen and audio response), U6

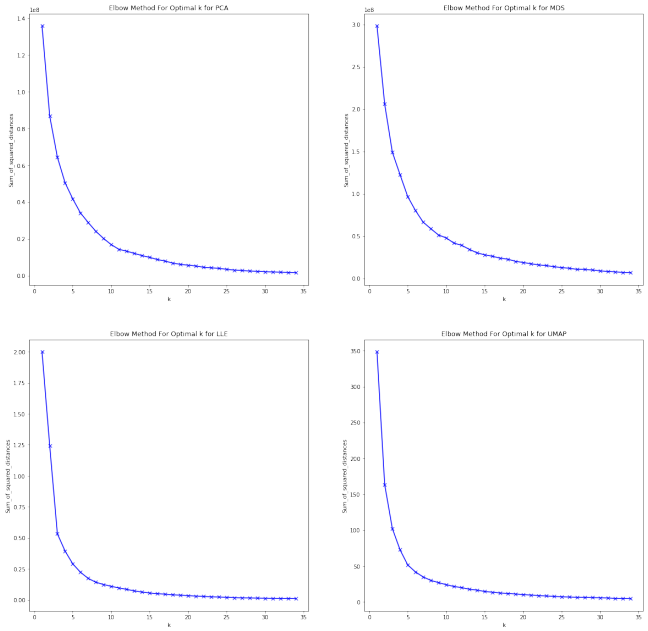


Figure A39: Elbow plots for finding optimal K value for clustering, pen response (simultaneous pen and audio response), U6. The x axis and y axis shows the k value and the sum of the squared distance (between each point and the centroid in a cluster) respectively for each subplot.

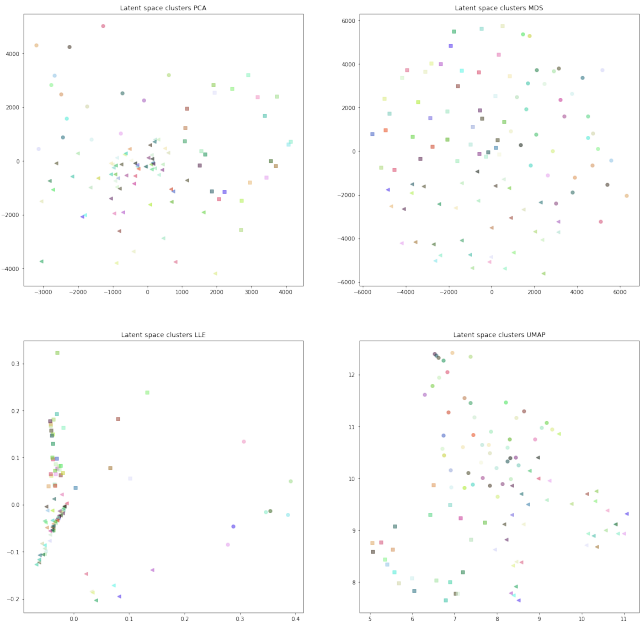


Figure A40: Latent Space Cluster Visualization of latent feature space, pen response (simultaneous pen and audio response), U6

A.6 User 7

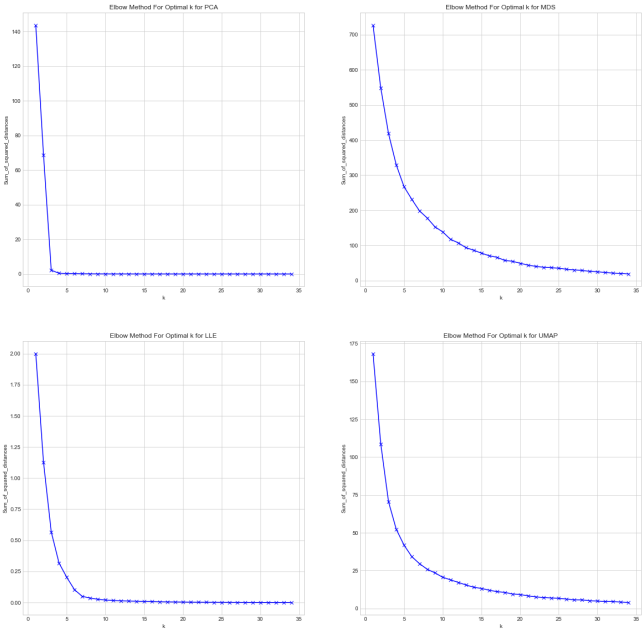


Figure A41: Elbow plots for finding optimal K value for clustering, audio/verbal only response, U7. The x axis and y axis shows the k value and the sum of the squared distance (between each point and the centroid in a cluster) respectively for each subplot.

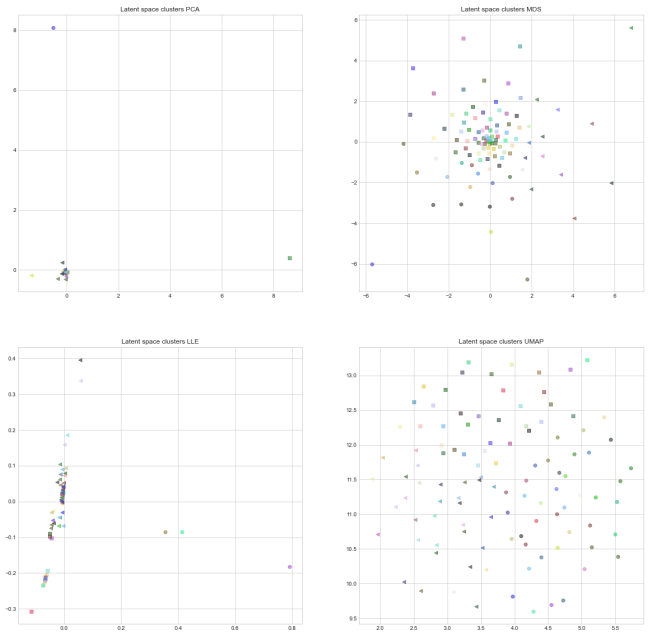


Figure A42: Latent Space Cluster Visualization of latent feature space, audio/verbal only response, U7

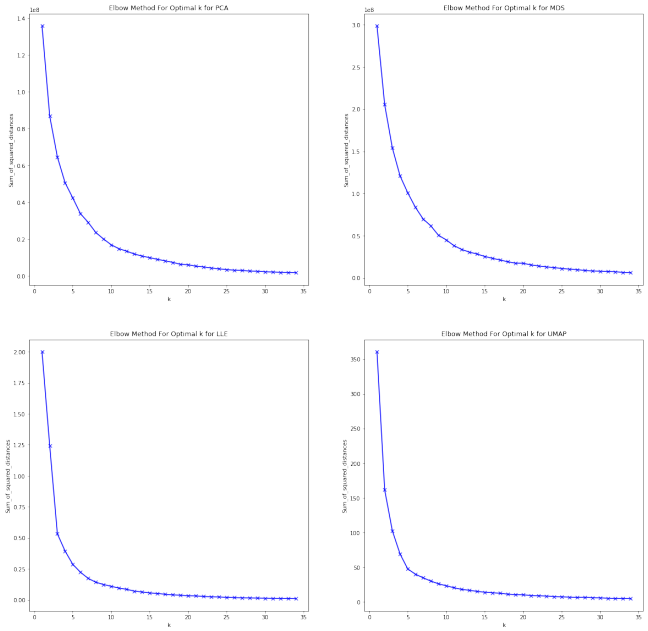


Figure A43: Elbow plots for finding optimal K value for clustering, pen gesture only response, U7. The x axis and y axis shows the k value and the sum of the squared distance (between each point and the centroid in a cluster) respectively for each subplot.

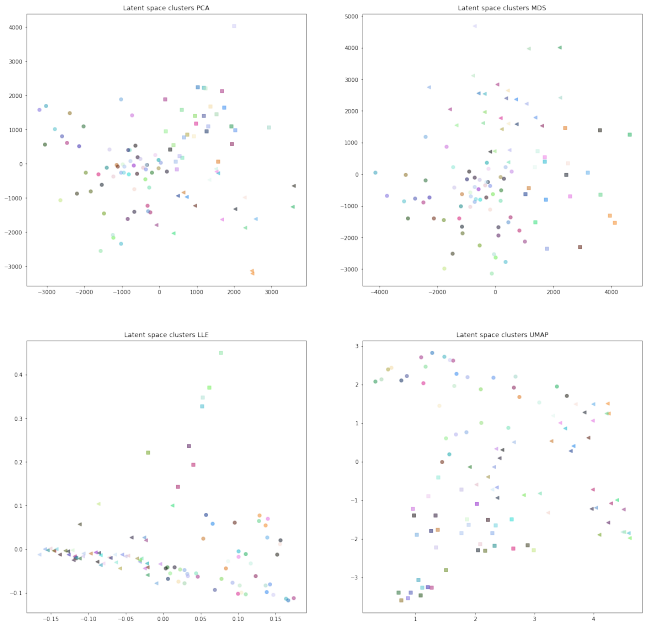


Figure A44: Latent Space Cluster Visualization of latent feature space, pen gesture-only response, U7

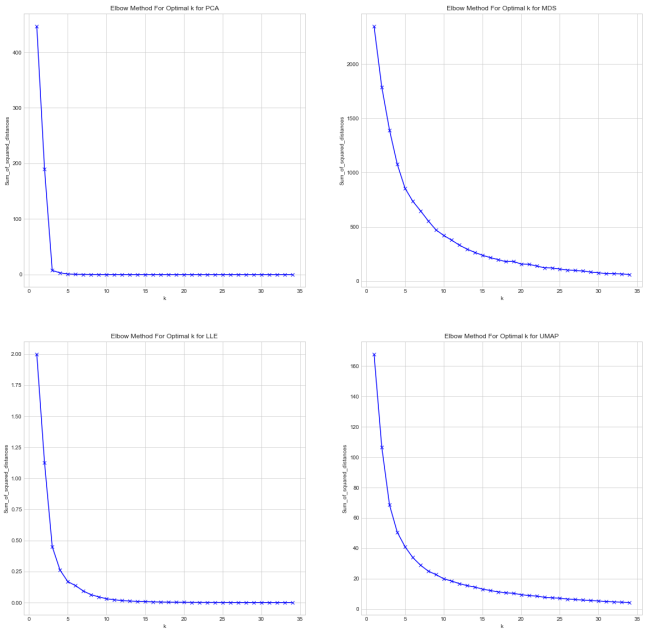


Figure A45: Elbow plots for finding optimal K value for clustering, audio/verbal response (simultaneous pen and audio response), U7. The x axis and y axis shows the k value and the sum of the squared distance (between each point and the centroid in a cluster) respectively for each subplot.

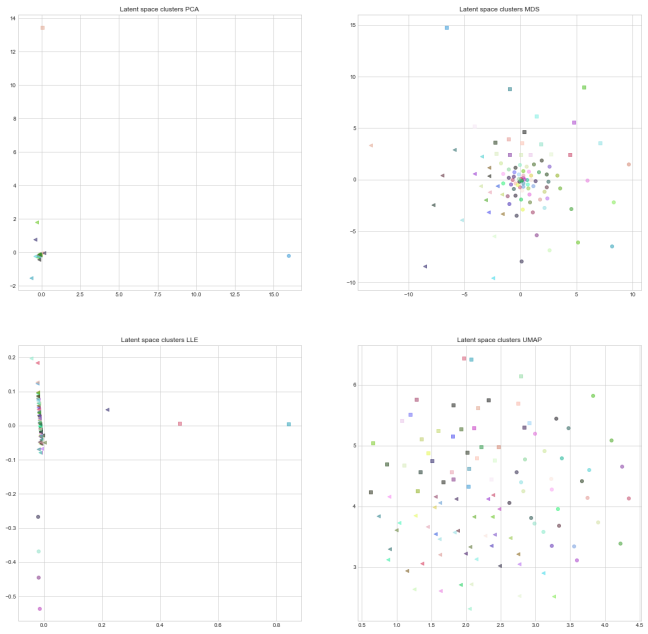


Figure A46: Latent Space Cluster Visualization of latent feature space, audio/verbal response (simultaneous pen and audio response), U7

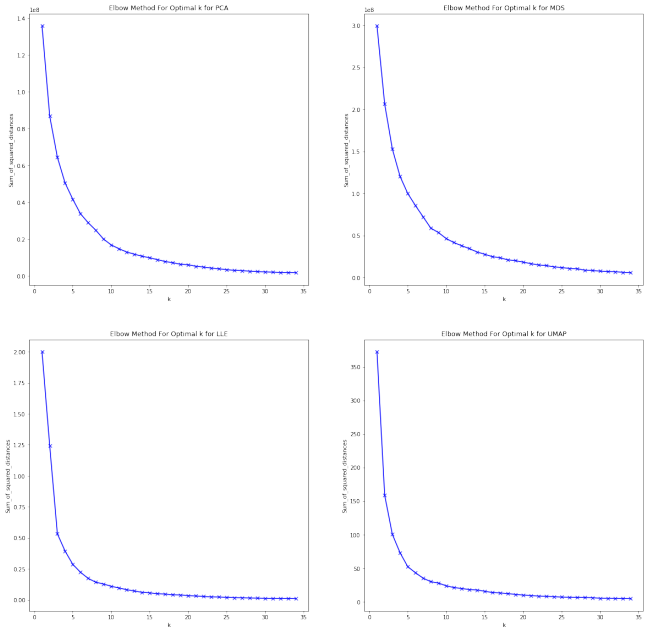


Figure A47: Elbow plots for finding optimal K value for clustering, pen response (simultaneous pen and audio response), U7. The x axis and y axis shows the k value and the sum of the squared distance (between each point and the centroid in a cluster) respectively for each subplot.

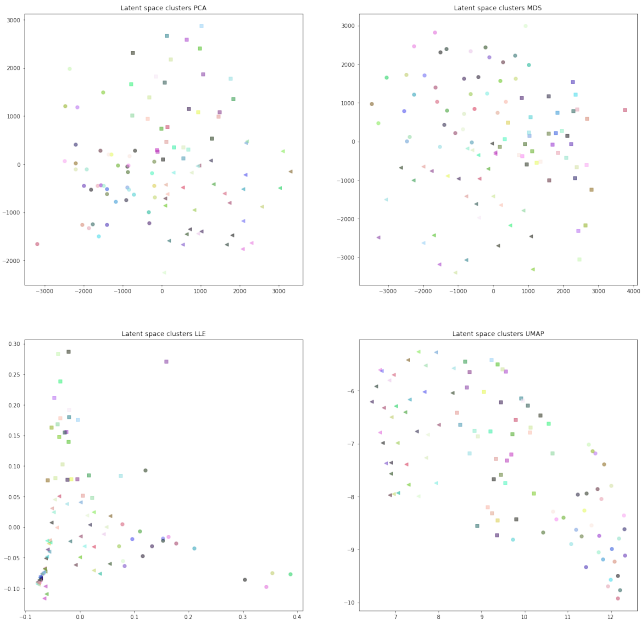


Figure A48: Latent Space Cluster Visualization of latent feature space, pen response (simultaneous pen and audio response), U7

A.7 User 8

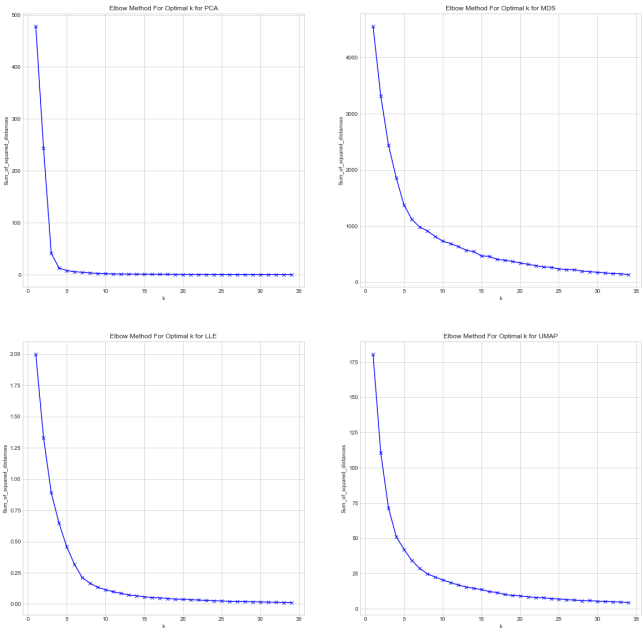


Figure A49: Elbow plots for finding optimal K value for clustering, audio/verbal only response, U8. The x axis and y axis shows the k value and the sum of the squared distance (between each point and the centroid in a cluster) respectively for each subplot.

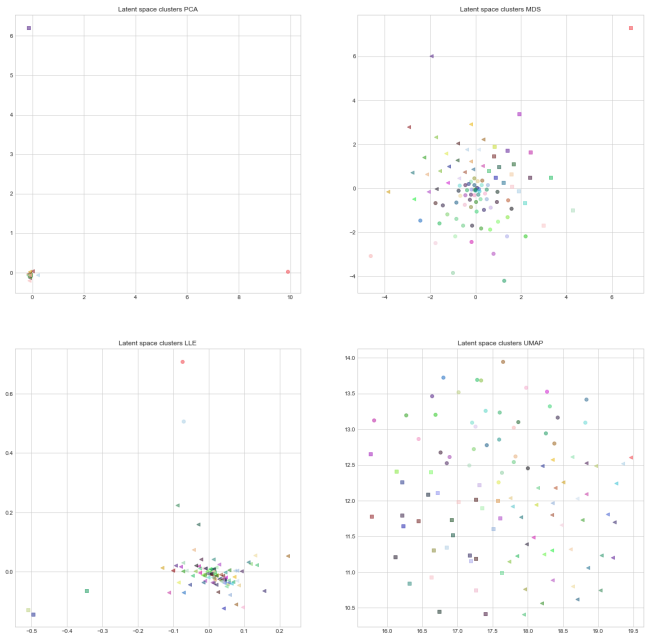


Figure A50: Latent Space Cluster Visualization of latent feature space, audio/verbal only response, U8

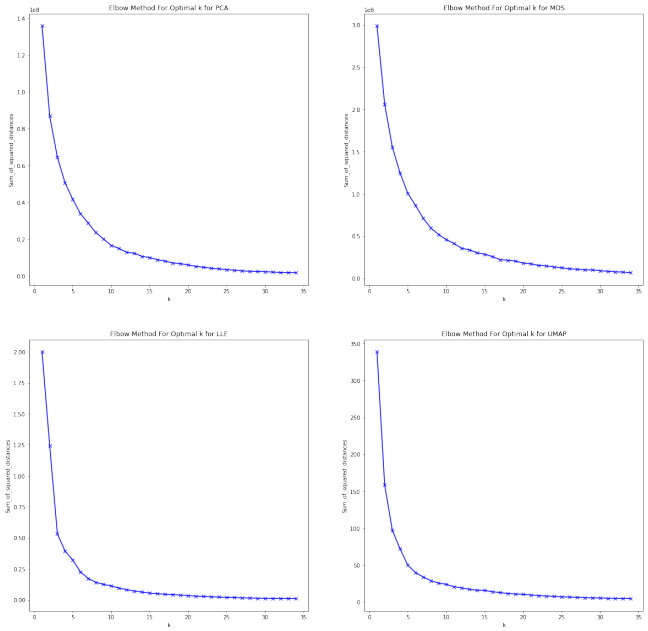


Figure A51: Elbow plots for finding optimal K value for clustering, pen gesture only response, U8. The x axis and y axis shows the k value and the sum of the squared distance (between each point and the centroid in a cluster) respectively for each subplot.

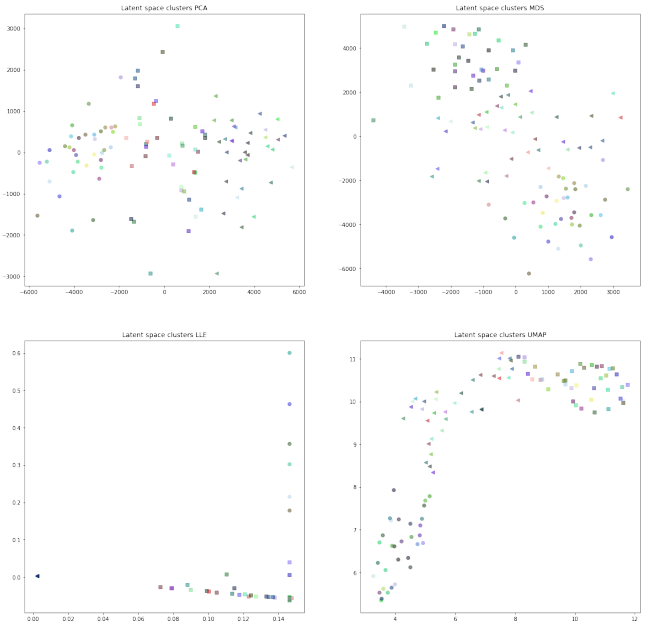


Figure A52: Latent Space Cluster Visualization of latent feature space, pen gesture only response, U8

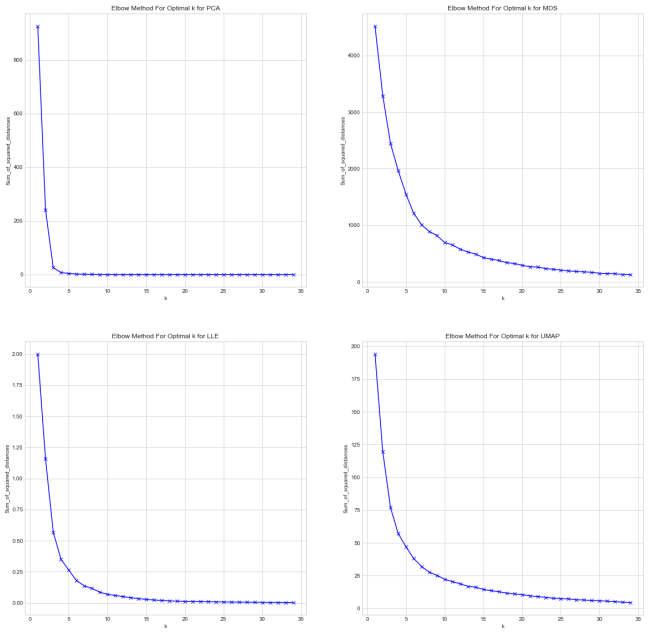


Figure A53: Elbow plots for finding optimal K value for clustering, audio/verbal response (simultaneous pen and audio response), U8. The x axis and y axis shows the k value and the sum of the squared distance (between each point and the centroid in a cluster) respectively for each subplot.

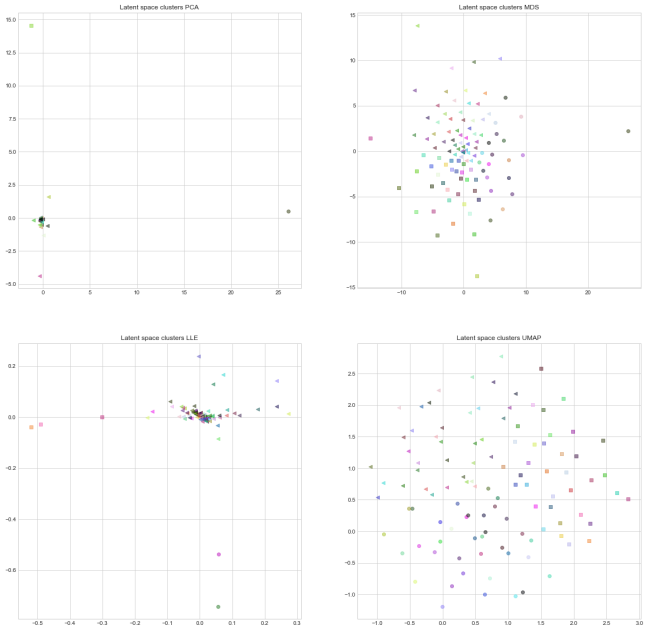


Figure A54: Latent Space Cluster Visualization of latent feature space, audio/verbal response (simultaneous pen and audio response), U8

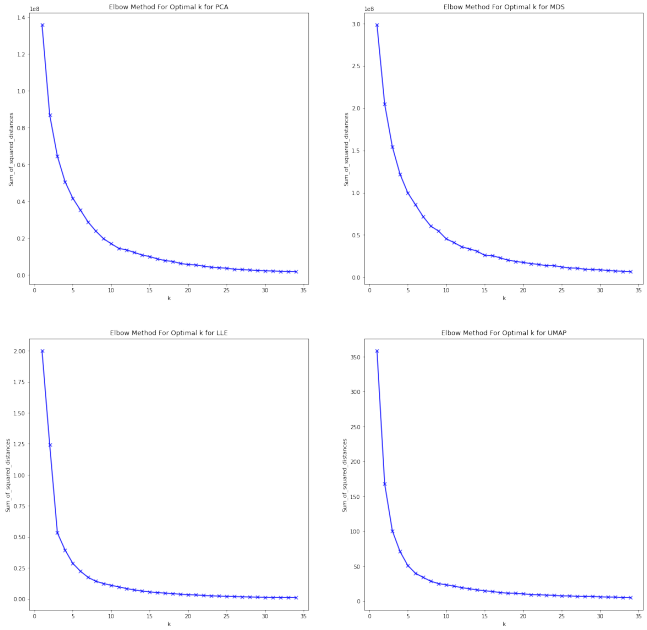


Figure A55: Elbow plots for finding optimal K value for clustering, pen response (simultaneous pen and audio response), U8. The x axis and y axis shows the k value and the sum of the squared distance (between each point and the centroid in a cluster) respectively for each subplot.

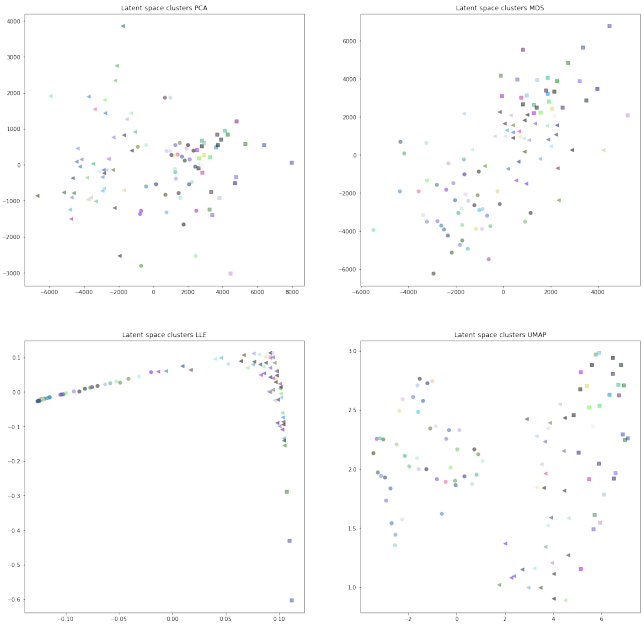


Figure A56: Latent Space Cluster Visualization of latent feature space, pen response (simultaneous pen and audio response), U8