Affective Analysis of Multimodal Dialogue Systems for Preschoolers

Vassiliki F. Kouloumenta

Department of Electronic and Computer Engineering

Technical University of Crete

Thesis committee:
Alexandros Potamianos, Supervisor
Vasileios Digalakis
Aikaterini Mania

A thesis submitted in partial fulfilment for the M.Sc. degree of

Electronic and Computer Engineering

Chania, Spring 2013

ΠΟΛΥΤΕΧΝΕΙΟ ΚΡΗΤΗΣ

Συναισθηματική Ανάλυση Πολυτροπικών Συστημάτων Διαλόγου για παιδιά προσχολικής ηλικίας

Βασιλική Φ. Κουλουμέντα

Τμήμα Ηλεκτρονικών Μηχανικών και Μηχανικών Υπολογιστών

Χανιά, Άνοιξη 2013

This work is dedicated to my parents, that made my choices come true by supporting me every step of the way.

Acknowledgements

This work took almost one and half year to complete, and a lot of things happened in my life during that time, things that made working on this thesis harder than I expected. I would like to take some time here and thank the people that made completion of this work possible despite of all that happened.

I'd like to thank all my colleagues and friends in TU Crete, who made these years in Chania what is possibly the best years in my life so far. Their help during my studies was invaluable to me, and I will always be grateful for it.

Of course it goes without saying that tremendous help was provided by my supervisor professor Potamianos. From the undergraduate courses all the way to the diploma thesis, he was always open to new ideas, providing his knowledge and experience to make what we had in mind possible. Also I would like to thank earnestly my colleagues, for their priceless assistance during those years.

Most importantly, I would like to thank my family, my parents and brother, for their help, support, guidance and love that made my life and my choices clear, easier, and most importantly: possible.

Finally, I would like to thank the person that made these past years truly the best years of my life; my friend Vassilis. Thank you for bearing with me during my good times and my bad times, I don't know if I could have made it without you!

Abstract

In this thesis, we investigate how the three Malone factors (fantasy, curiosity, challenge) influence the interaction patterns in preschool ages, and how emotion can be a significant indicator for adaptation. For this reason, we designed and implemented a multimodal dialogue system with five incorporated task oriented games. These five tasks were designed to be suitable for preschool children and they assert various levels in fantasy, curiosity and challenge. The implemented platform was used in an experimental process, where fifteen children four to six years old participated in. Each child was asked to play different versions of the application, one for each factor level. The results indicated that fantasy and curiosity are highly correlated with the user's entertainment, while the preferred level of challenge seems to be very user dependent. In addition various objective features, as well as audio features were measured during the game play, in order to study which are the interaction patterns formed by infants. The comprehension of such interaction patterns are expected to lead us to the definition of the parameters, that are good for adaptation.

Also, by using speech characteristics, we try to predict the user's emotional state, as an additional factor for the system's adaptation. Results indicate that emotion recognition, in early ages, is at least as easy as emotion recognition for adult users. This is attributed to the fact that infants tend to be more expressive when talking. Moreover the prediction of optimal level of fantasy, curiosity and challenge is attempted.

Finally, we attempt to investigate how arousal, engagement and meditation are evolute during interaction, as well as how they are affected

by input modality, by using the NeuroSky device. NeuroSky is a device that used for measuring engagement and meditation levels as well as various encephalographic signals. For this reason we conducted a different experiment where both adults and children took part.

Results indicated that fantasy and challenge are easily predictable (97% accuracy for the challenge), while curiosity recognition demonstrated to be a hard task. Four classifiers are trained and tested for both factor and emotion prediction, along with feature selection techniques. With this work, we attempt to bring forward the basis for the development of adaptive systems for preschool ages, by associating optimal levels of the three Malone factors along with emotion automatic recognition.

Abstract (in Greek)

Σε αυτή την εργασία, μελετάμε πώς οι τρείς παράγοντες του Malone (φαντασία, περιέργεια και δυσκολία) επηρεάζουν τα πρότυπα αλληλεπίδρασης σε παιδιά προσχολικής ηλικίας, και πώς το συναίσθημα μπορεί να αποτελέσει έναν σημαντικό δείκτη για την προσαρμογή υπολογιστικών συστημάτων στα δεδομένα του χρήστη.

Για αυτό τον σκοπό, σχεδιάστηκε και υλοποιήθηκε ένα πολυτροπικό σύστημα διαλόγου, με πέντε ενσωματωμένα παιχνίδια. Αυτά τα παιχνίδια σχεδιάστηκαν ώστε να είναι κατάλληλα για παιδιά προσχολικής ηλικίας, και υποστηρίζουν ποικίλα επίπεδα φαντασίας, περιέργειας και δυσκολίας. Η πλατφόρμα που υλοποιήθηκε χρησιμοποιήθηκε στην πειραματική διαδικασία, όπου έλαβαν μέρος δεκαπέντε παιδία ηλικίας τεσσάρων έως έξι ετών. Σε κάθε παιδί ζητήθηκε να παίξει διάφορες εκδόσεις της εφαρμογής, μία για κάθε παράγοντα.

Τα αποτελέσματα έδειξαν ότι η φαντασία και η περιέργεια είναι στενά συνδεδεμένες με το επίπεδο ψυχαγωγίας του χρήστη, ενώ το επιθυμητό επίπεδο δυσκολίας δείχνει να εξαρτάται από τον εκάστοτε χρήστη. Επιπροσθέτως διάφορα 'αντικειμενικά' χαρακτηριστικά, όπως επίσης και ηχητικά χαρακτηριστικά μετρήθηκαν κατά την διάρκεια του παιχνιδιού, ώστε να μελετήσουμε ποιά είναι τα πρότυπα αλληλεπίδρασης που σχηματίζονται από τα παιδιά. Η κατανόηση των προτύπων αλληλεπίδρασης αναμένουμε να μας οδηγήσει στον ορισμό των παραμέτρων που είναι κατάλληλες για την προσαρμογή των υπολογιστικών συστημάτων.

Επιπλέον, χρησιμοποιώντας χαρακτηριστικά της φωνής, προσπαθούμε να προβλέψουμε την συναισθηματική κατάσταση του χρήστη, ως έναν επιπλέον παράγοντα για την προσαρμογή συστημάτων. Τα αποτελέσματα αποδεικνύουν ότι η αναγνώριση συναισθήματος σε πρώιμες ηλικίες είναι το ίδιο εύκολη με την αναγνώριση συναισθήματος σε ενήλικες.

Αυτό οφείλεται στο γεγονός ότι τα παιδιά τείνουν να είναι περισσότερο εκφραστικά όταν μιλάνε. Επιπλέον η πρόβλεψη του βέλτιστου επιπέδου της φαντασίας, της περιέργειας και της δυσκολίας επιχειρείται.

Τέλος, επιχειρούμε να διερευνήσουμε χρησιμοποιώντας την συσκευή NeuroSky πώς τα arousal, engagement και meditation εξελίσσονται κατά την διάρκεια της αλληλεπίδρασης, καθώς επίσης πώς επηρεάζονται από την είσοδο. Η συσκευή NeuroSky χρησιμοποιείται για να μετράμε τα επίπεδα ενγαγεμεντ και μεδιτατιον καθώς επίσης και ποικίλα εγκεφαλικά σήματα. Για αυτό το λόγο κάναμε ένα διαφορετικό πείραμα όπου ενήλικες και παιδιά έλαβαν μέρος.

Τα αποτελέσματα έδειξαν ότι η φαντασία και η δυσκολία είναι εύκολο να προβλεφθούν (97% επιτυχία πρόβλεψης για την δυσκολία), ενώ η περιέργεια αποδείχθηκε δύσκολη εργασία. Τέσσερεις ταξινομητές εκπαιδεύτηκαν και δοκιμάστηκαν τόσο για αναγνώριση των παραγόντων όσο και για αναγνώριση συναισθήματος. Χρησιμοποιήθηκαν επίσης και τεχνικές επιλογής χαρακτηριστικών.

Με αυτή την δουλειά, επιχειρούμε να θέσουμε τις βάσεις για την ανάπτυξη συστημάτων για παιδιά προσχολικής ηλικίας, συνδυάζοντας τα βέλτιστα επίπεδα των τριών παραγόντων του Malone μαζί με την αυτόματη αναγνώριση συναισθήματος.

Contents

List of Figures

Chapter 1

Introduction

Human communication is easy and efficient due to the fact that, humans when communicate use several modalities, such as speech, gestures, gaze etc. Humans, in their everyday face - to - face communication, use several means in order to express emotions, opinion etc. For example a face expression, a gaze, or a movement may change the whole meaning of a sentence. That's why human communication is so alive and direct.

The humans' ability to communicate under all circumstances in combination with the emergence of powerful computer systems trigger research interest about new, more efficient ways of interaction, that involve combination of more than one modalities. Such examples of modalities include speech, pen, gestures, head and body motions.

Nowadays, multimodal systems are all the more part of our everyday life, e.g. mobile applications. Multimodal applications take advantage of the human, face-to-face, communication characteristics by combining various input and output modalities, such as speech, gestures, touch, multi-touch etc. As a result multimodal applications can lead to a more natural, robust and effective human-computer interaction.

One interesting and open field of research is the design of multimodal dialogue systems for children. As in early ages the learning procedure takes place through gaming, it is important to develop educational computer applications. Multimodal interaction systems for children have become increasingly popular, mainly due to the way children express themselves. Children tend to use more than

one modalities in communication with others (e.g. voice, gestures etc.).So, different input modalities may help in order to increase efficiency in child-computer interaction.

Although children are good adopters of technology, there are several challenges in the design and the implementation of multimodal dialogue systems for preschoolers. As it is already mentioned, children tend to use more than one modalities simultaneously. Additionally automatic speech recognition, in such ages, is not an easy task because of the linguistic variability children display when talking. So the existence of various modalities would help to overcome such recognition problems.

Games in early ages are part of the learning procedure. The Child Computer Interaction (CCI) is a research field, that is concentrated on the design and implementation of computer systems suitable for children. Educational computer games should combine various factors in order to be fun and interesting. According to Malone [78, 77], Fantasy, Curiosity and Challenge are essential components of a game. These components constitute fundamental elements of the gaming procedure, as their mixture defines, in part, whether the game would be interesting or not. So the prediction of these three factors during game play is an important task. On that basis, we also provide classification results, along with feature selection, as an attempt to recognize the optimal levels of fantasy, curiosity and challenge, for preschool ages.

In order to investigate how preschoolers interact with multimodal dialogue systems, we have designed and implemented an on-line, web-based, multimodal platform. By this way, it is easier to quickly prototype, deploy and evaluate multimodal dialogue systems. By using this platform we have implemented five games, that provide both mouse and speech input modalities.

One of the goals of this work is to provide a multimodal platform that can be used in the future, not only for studying child-computer interaction, but for any web-based multimodal application. Furthermore, as we mentioned previously, there is little research in studying how interaction patterns are formed at preschool ages and the child-computer interaction with such systems. This work provides results that could help in further studies towards developing better applications for children. Moreover, by identifying which of the factors are good indicators

towards adapting multimodal dialogue systems for children, the goal for better multimodal interfaces for kids is even closer. In such ages entertainment and learning are intertwined activities, so the existence of an usable and entertaining interface is crucial during the learning procedure.

In addition to optimal factor levels prediction, we attempt to automatically recognize users' emotions and study the mental state during the game. Emotions are important characteristics of human communication, since they help us express ourselves. Computers, able to recognize human emotions, would lead to a more natural and effective interaction. Also, such systems can be easily adaptable to users' needs. For example, if the user feels frustration the system should be able to adapt its behavior respectively. For this reason, we apply several models on speech features, in an attempt to classify children emotions. As well as by using an Electroencephalographer (EEG) device we attempt to study the mental state of a user's.

The remainder of this thesis is organized as follows. In Chapter 2 e establish the context behind this work, by referring to previous research work and the state of the art. In Chapter 3 we represent the system's architecture and the games functionality. Moreover in Chapter 4 we define the three factors that make a computer game "fun" and interesting, as well as we will represent the features and the models used in the optimal factor classification and emotion classification problems. In addition, we describe the whole experimental methodology. Also a representation of the results for the system evaluation, emotion classification and factor classification, is included. In Chapter 5 we represent the affective evaluation results with physiological signals.

Chapter 2

Background and Related Work

2.1 Human Computer Interaction

Human communication is effective and efficient due to the fact that, humans when communicate use several modalities, such as speech, gestures, gaze etc. Humans in their everyday life, when communicating with each other use various means in order to express emotions, opinion etc.

The field of Human Computer Interaction (HCI) studies the interaction between users and computer systems. HCI is subject of study for psychologists and cognitive scientists who concentrate on how humans perceive the world and how the process information from their environment, in order to make decisions and solve problems. Also computer scientists ad engineers are occupied with the study and the design of computer systems, that are capable to simulate the human face-to-face communication.

2.1.1 Multimodal Systems

"During multimodal communication, we speak, shift eye gaze, gesture and move in a powerful flow of communication that bears little resemblance to the discrete keyboard and mouse clicks entered sequentially with a Graphical User Interface (GUI)" [69].

The multimodal interface design tries to achieve an interaction closer to human-human communication, as well as to increase the robustness of the inter-

action by using redundant or complementary information. Reeves in [44] refers interaction paradigms and guidelines for user interface design. Such guidelines are:

- **Requirements Specification:** Specify the user requirements and system capabilities for a given domain. The "*design for broadest range of users and contexts of use*" and "*address privacy and security issues*" are two general considerations for multimodal system requirements specification.
- **Designing Multimodal Input and Output:** Maximize human cognitive and physical abilities. Integrate modalities in a manner compatible with user preferences, context and system functionality.
- **Adaptivity:** Dynamic adaptivity enables the interface to degrade gracefully by leveraging complementary and supplementary modalities according to changes in task and context.
- **Consistency:** Presentation and prompts should share common features as much as possible and should refer to a common task.
- **Feedback:** Users should be aware of their current connectivity and know which modalities are available to them.
- **Error Prevention/Handling:** User errors can be minimized and error handling improved by providing clearly marked exits from a task, modality or entire system and by easily allowing users to undo a previous action or command.

Besides those general design principles, when we build a multimodal dialogue application, data collection and evaluation are important components of the design procedure. According to [12] to design a successful dialogue system, designer should take four steps: *architectural design, application design and data collection, speech and natural language interface design* and finally *user feedback and evaluation.*

Spoken dialogue systems consist of several components, each with a certain functionality [51] speech recognition systems with automatic speech recognizers

(ASRs), natural language understanding, dialogue manager, communication with external system, response generation and speech output either with prerecorded prompts or text-to-speech (TTS) technologies. A spoken dialogue system's architecture is presented in Figure 3.1.

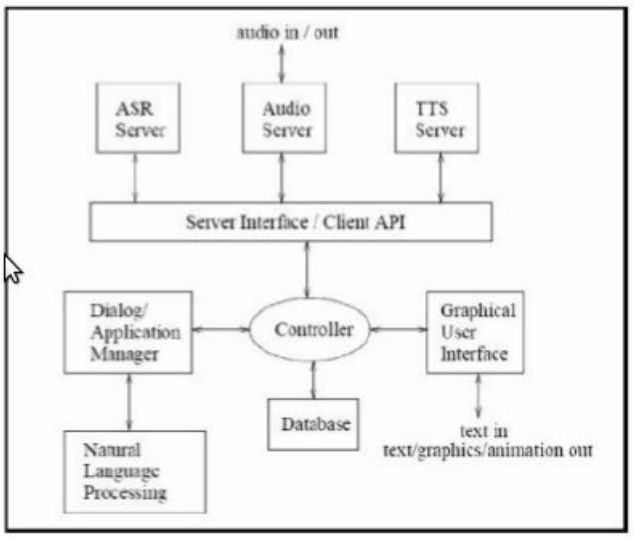


Figure 2.1: A common multimodal system architecture

In some cases a dialogue system may be incomplete, meaning that some of its components are missing. In such cases the usage of a Wizard of Oz system allows the observation of a user operating an apparently fully functioning system. The user is not aware of the presence of a wizard and is led to believe that the computer system is fully operational [24], [23].

Tow famous examples of multimodal systems include the QuickSet speech/pen input system, which had been developed in conjuction with several map-based applications [69], [68], [61]. QuickSet enables users to create and position entities on a map by using speech, pointing gestures with a pen and/or direct manipu-

lation. Also the MATCH system, which provide a mobile multimodal interface for restaurant and subway information [52]. Ancestor of all multimodal systems is the Bolts' Put-that-There system [64], where users could manipulate various objects on a screen by using hand gestures and speech simultaneously.

2.1.2 Multimodal Dialogue Systems for Children

Recently there is a growing interest in the design and development of multimodal dialogue systems for children. As in early ages the learning procedure takes place through gaming it is important to develop educational computer games. Multimodal interactive systems have become increasingly popular, mainly due to the way children express themselves. To be more specific children tend to use more than one modalities in communication with others, (e.g. voice, gestures etc.) [9], [8]. As a result, different input modalities may help in order to increase efficiency in child-computer interaction.

Although children are good adopters of technology, there are several challenges in the design and the implementation of multimodal dialogue systems for children. Automatic speech recognition in such ages is not an easy task because of the linguistic variability children display when talking. So the existence of various modalities should help to overcome such recognition problems.

In recent years lot of research has been done in the design of multimodal dialogue for children. A variety of prototype systems, containing spoken dialogue interfaces and capabilities, have been implemented. In [66], [67] authors describe their efforts in designing and building a prototype multimodal system for children users. The CHildren's Interactive Multimedia Project (agent CHIMP) provides design guidelines for building successful multimodal-input, multimedia-output applications for children users. An important feature of the CHIMP system involves the integration of multiple input and output modalities such as voice, audio, keyboard, mouse, graphics and animation. In [15] researchers make a comprehensive analysis of children's multimodal integration patterns during interaction with an educational software prototype. Moreover, the NICE Fairy-Tale game system, which is described in [36], allows users to interact with various animated characters in a 3D world.

2.1.3 Brain Computer Interfaces

As we previously mentioned, the interaction between humans and computers is done in several ways. Most people use a simple keyboard or mouse to give instructions to the computer. Newer forms of human computer interaction, such as speech, have been invented. In addition to these existing ways of interacting with computers, the use of brain activity is becoming increasingly popular. Brain activity could be used to decipher thoughts, or intent, so that a person could communicate with others or control devices directly, without using the normal channels of peripheral nerves and muscles. This method is called a Brain Computer Interface (BCI).

Zander in [75] classifies BCIs in three categories:

- **Active BCI:** An active BCI derives its outputs from brain activity which is directly consciously controlled by the user, independently from external events, for controlling an application.

- **Reactive BCI:** A reactive BCI derives its outputs from brain activity arising in reaction to external stimulation, which is indirectly modulated by the user for controlling an application.

- **Passive BCI:** A passive BCI derives its outputs from arbitrary brain activity without the purpose of voluntary control, for enriching a human computer interaction with implicit information.

In that paper the author suggests that a BCI could be used as a complementary input modality serving as an action selection device within an eye gaze controlled environment. A two dimensional cursor control is realized by tracking the user's eye gaze and a BCI-detectable mental gesture, an imagination of a two-handed movement, serves as the selection command.

2.2 Fantasy, Curiosity, Challenge and Entertainment

Since the main purpose of the games is to entertain the player, one of the aspects someone should consider is how to measure the entertainment and how multimodal dialogue systems can be adapted, in order to increase the user's satisfaction. Several theoretical studies exist on how we can define fun and what are the elements that should exist in a game in order to be entertaining.

Csikszentmihlyi [46] studied what makes experiences enjoyable to people. He stated that there is an optimal psychological state, namely flow, where people are fully absorbed by an activity. That psychological state is independent of cultural, social and age or gender characteristics. The flow experience is characterized by various elements such as challenge. The two-dimensional flow model is shown in Figure 2.2.

In [78], [77] Malone describes what makes computer games fun. According to Malone "the essential characteristics of good computer games and other intrinsically enjoyable situations can be organized into three categories":

- **Challenge:** A computer game is challenging when provides a goal whose achievement is not obvious and the outcome is uncertain. If the game is too easy and the outcome is likely to be certain, players will be quickly get bored. On the other hand, if it is very difficult to achieve the goal, players will be possibly demotivated. According to his theory there are four ways to make the outcome of a game uncertain, such as various difficulty levels, multiple level goals, hidden information and randomness.

- **Fantasy:** Fantasy makes computer games more interesting. Games that include fantasy show or evoke images of physical objects or social situations not actually present. There are two kinds of fantasies. The first namely extrinsic fantasies, depend only on whether or not the skill is used correctly. The second is intrinsic fantasies, where not only the fantasy depends on the skill, but also the skill depends on the fantasy. It is said that fantasy in games "derive some of their appeal from the emotional needs they help to satisfy in the people who play them".

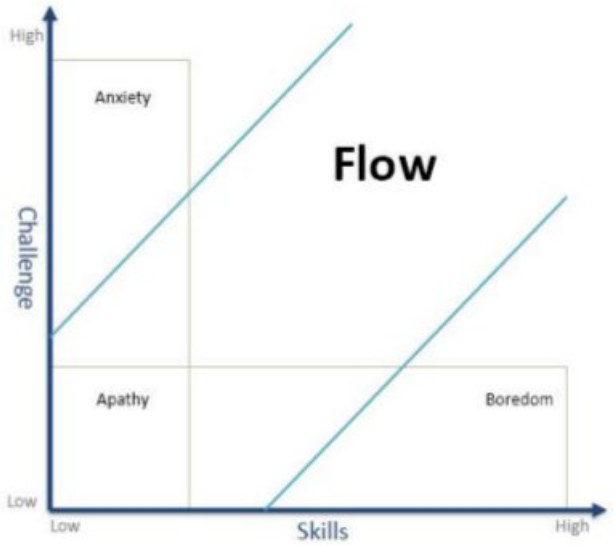


Figure 2.2: The two-dimensional model of flow based on Csikszentmihalyi

- **Curiosity:** Curiosity is the motivation to learn, independent of any goal-seeking or fantasy-fulfillment. Computer games can evoke a learner's curiosity by providing environments that have an optimal level of information complexity. Curiosity can be distinguished between sensory and cognitive curiosity. Sensory curiosity attracts the attention from sensory stimuli, such as audio and visual effects. Cognitive curiosity is achieved by surprising and constuctive feedback. Finally Mallone states that randomness is believed to be an easy way to arise the user's curiosity.

Alternatively, Lazzaro [56] based on Malone's factors and facial expressions data obtained from actual games, identified four relevant categories: *hard fun, easy fun, altered states and the people factor.*

Moreover, adaptation plays a very important role in the quality of the educational experience, as it allows the learning environments to cater to students

with different expectations and objectives.

The more complex the applications become, the harder is to build applications and interfaces satisfying users' needs. Users have different capabilities and preferences, especially when multiple modalities are available to them. As a result the need of adaptation to the specific user characteristics, needs, capabilities and preferences is becoming apparent. According to [10] adaptation can be defined as "an interactive system that adapts its behavior to individual users on the basis of processes of user model acquisition and application that involve some form of learning, inference or decision making" [8]. suggests, technology promotes exciting learning opportunities. Therefore, educational games can be considered as a highly interactive medium and reactive to the actions of the user. From a technical perspective, this makes them the ideal medium and reactive learning experience, meaning that the game can both monitor the activity of the user, as well as change its own behavior accordingly.

Moreover, in the field of entertainment capture Yannakakis [32] indicated that the player-opponent interaction is a major factor in entertainment. Recent studies have also indicated the impact of game content to player experience [34]. Authors, view player experience as the synthesis of affective patterns elicited and cognitive processes generated during game play. In order to create games able to adapt to the children's preferred level of fantasy, curiosity and challenge it is important to investigate the correlation between various objective, metrics and these factors. Additionally, information from other sources such as voice, video and physiological measurements could be used as features. In [35, 32, 33]authors survey how fun can be measured. Physiological measurements such as children's heart rate (HR), blood volume pulse (BVP) and skin conductance (SC) are used as features to predict engagement.

Others such as [71] used EEG measurements in order to study various phenomena during interaction. EEG frequency analysis has shown that during the performance of mental tasks all EEG frequency components change in relation to rest. The most spectacular change is the decrease in alpha power. In general an increase in theta activity has been related with task difficulty and emotional factors.

2.3 Emotion

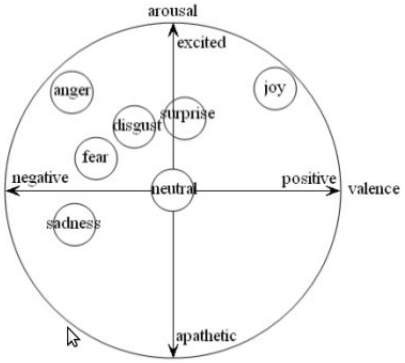
Emotions are very important during human communication and interaction, since they help people express themselves beyond the verbal domain. Recently there is a growing interest in affective gaming on the sensing and recognition of the player's emotions e.g. minimizing frustration, ensuring appropriate challenge. Computers able to understand human emotions will lead in a more natural and consequently more effective interaction. In applications where computers undertake a social role (e.g. an instructor, helper or companion), the ability to recognize human emotions it will enhance their functionality.

The three core areas of affective computing provide methods and techniques directly relevant to affective game design: *emotion sensing and recognition, computational models of emotion, emotion expression by synthetic agents and robots.*

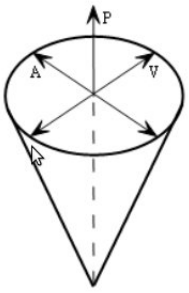
According to psychologists, emotions model the attractiveness or aversiveness in an event, as well as the excitement that same event causes to the subject. In [39, 59, 45] authors suggest that emotions can be projected into a two dimensional space, where the one dimension namely valence represents how pleasant or unpleasant an emotion is, while the second dimension namely arousal is orthogonal to valence and represents how exciting an emotion is. Alternatively, Osgood, Suci and Tannenbaum [20] suggested that emotion computing an be conceptualized as three major dimensions of cognitive meaning: valence, arousal and power. The power dimension refers to the degree of power or sense of control over the emotion. The two dimensional and three dimensional models are shown in Figure 2.3.

To be more specific emotional arousal modifies the allocation of attention and heightens sensitivity to environmental cues related to the motivational state induced by the provoking stimulus [60]. This is particularly true for stimuli with inherent significance for survival. Valence focus is defined as the extent to which an individual incorporates pleasantness or unpleasantness into their conscious affective experience and may be associated with a tendency to attend to the pleasant or unpleasant aspects of a stimulus [43].

In the concept of gaming, Lazzaro states that people play games not as much for the game itself, as for the experience the game creates (e.g. adrenaline rush)



(a)



(b)

Figure 2.3: (a) Distribution of the seven emotions in valence-arousal space (b) Three dimensions of emotion space(V-valence, A-arousal, P-power)

or the structure games provide (e.g. the company of a friend) [56]. Klimmt in [17] also proposed that the anxiety and physical arousal elicited by playing, are interpreted as positive emotions, when gamers are close in wining a game or complete a task.

When it comes to emotion recognition much work has been done in recognizing emotions from speech. It is known that different emotional states can affect a speaker's speech production mechanism and can lead to acoustical variations. People perceive such variations as emotion manifestation [13]. The vocal aspect of communication carries various kinds of information. Traditionally, as well as in more recent studies, emotions can be recognized using features from prosodic characteristics which include pitch, duration and energy statistics extracted from the user's speech input [14, 27, 37, 47, 74, 25, 35, 50, 19, 81]. Facial expressions also have a significant amount of information, regarding emotions. In [58] Ekman and Friesen developed the Facial Action Coding System (FACS) to code facial expressions, where movements of the face are described by a set of action units (AUs). Each facial expression may be described by a combination of AUs. Recent work on emotion recognition using video has used these basic expressions in order to recognize emotions. Also in [16, 80, 76] facial expressions were used in order to estimate emotion in real-time.

Recently great interest is addressed in emotion recognition from physiological features. In [29, 41, 63, 30] Galvanic Skin Response (GSR), blood pressure, temperature sensors and respiration were used in order to collect physiological data during game play. Those data were used in emotion detection.

The electroencephalogram (EEG) also is used in order to detect emotions. Since physiological signals constitute spontaneous reaction of the human body are considered to be more reliable than speech and facial expressions which can easily be faked. Authors in [49] suggest that positive emotions are associated with relatively greater left frontal brain activity, whereas negative emotions are associated with relatively greater right frontal brain activity. Alpha waves are typical for an alert but relaxed mental state and are positively correlated with relaxation and/or meditative states. High alpha activity has also been correlated to brain inactivation. Beta activity, on the other hand is related to an active

state state of mind more prominent during intense focused mental activity. Taking under consideration the characteristics of alpha and beta waves, Bos in [22] states that how relaxed a player is, is determined by the ratio of beta and alpha brainwaves as recorded by the EEG.

In [26] authors focus on the theta rhythm of infants and preschool children. They state that the theta wave response is related to the emotional and attentional processes associated the perception of a stimulation. Children and infants compared to adults experience more intense affective states in the laboratory setting and their attention is to a higher degree.

Finally, emotion recognition is also possible by combining different modalities [18]. Audio, linguistic, pragmatic and visual information can be combined to obtain a good prediction of the child's emotional state [16]. Recently there has been interest in emotion recognition and modeling of children's mood in spoken dialogue and gaming applications [70].

Chapter 3

System Overview

Building web-based systems for multimodal game applications offers a great advantage, since they facilitate data collection and analysis, in order to train language and acoustic models and study users' interaction patterns. Consequently, we can improve and/or adapt the user interface. Below we represent the system implemented by Theofanis Kannetis [72, 73].

3.1 System's Architecture

The system follows the client/server architecture and its structure is shown in Figure 3.1. The system's functionality is based on the collaboration among different modules.

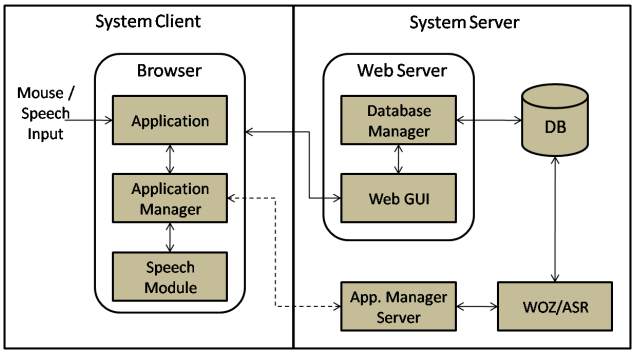


Figure 3.1: System's architecture [72]

The Speech Module is responsible for capturing and streaming the audio, as well as performing Voice Activity Detection (VAD), in order to determine whether the user speaks. Three main functions have been implemented in the module. The *CaptureAudio* function is responsible for capturing and streaming the audio to the Application Manager (server part). The *VoiceDetection* function is responsible for Voice Activity detection (VAD). Voice activity detection is the technique used in speech processing where the presence or absence of human voice is detected in regions of audio (which may also contain music, noise, or other sound). VAD is an important technology for a variety of speech-based applications. Therefore various VAD algorithms have been proposed that provide different compromises between latency, sensitivity, accuracy and computational cost. A very simple algorithm has been implemented, using the energy of the captured frames by computing the Root Mean Square (RMS) for each frame. When RMS is above a specific threshold, the *CaptureAudio* function is called.

The Application Manager Module is the most important part of the system, since it is responsible for the collaboration between the different modules. To be more specific, when a VAD event occurs, the client part of the Application Manager sends a request to the server part, in order to establish the streaming connection, and waits for the appropriate response. If the response is positive, meaning that the connection was established, the Application Manager notifies the Speech Module, to start recording and streaming the audio data. On the other hand, when the user stops talking, the client part of the Application Manager should notify the corresponding server part of the module, that the streaming was over.

The server part of the Application Manager is responsible for receiving the audio data and send them to the ASR/Wizard of Oz (WoZ) module, which is responsible for choosing the appropriate answer according to what the user said. Also it is responsible for receiving and storing the data into log files for further elaboration. The communication between the two parts of the Application manager (client part and server part) occurs through a two-way TCP/IP socket.

The WoZ module creates to the user the "illusion" that is working on a fully functioning system, although some services are missing. The missing services are supplied by a hidden wizard, whose presence is not perceptible by the user.

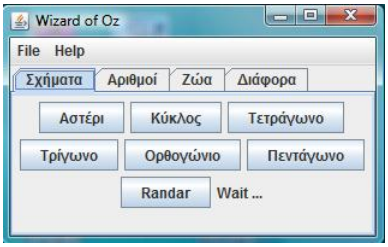


Figure 3.2: The WoZ Graphic User Interface [72]

To be more specific, the wizard observes the user interacting with a computer system, and if the user invokes a function, that is not available, the wizard is called upon to simulate the effect of that missing function. Automatic speech recognition for infantile speech is a quite difficult task, because of infants' acoustic and linguistic variability [67, 65]. So in our system, the WoZ component is used to replace the Automatic Speech Recognition (ASR) component and it is operated by a human transcriber. The WoZ was implemented as a Graphic User Interface (GUI) (Figure 3.2).

On the server side of the system, are also implemented the Web Interface and the Database Communication Modules. Using the web interface, users can register and login to the platform. Functionality such as profile management and preferences configuration, such as microphone calibration are also provided. The Database Communication Module is responsible for all the necessary database queries, including registration and login. The process of a speech request through the system is shown in Figure 3.3.

3.2 Game Functionality

Next we discuss how we implemented the different multimodal games supplied by our system. According to [7] usability is a very important principle that a designer should take under consideration, especially when it is about educational software. If children find it hard to use an educational technology, they will not learn through the process. As far as child users are concerned, age is also an important parameter as it affects significantly the interaction patterns. Younger children display significant attention disruption, compared to that of older ones.

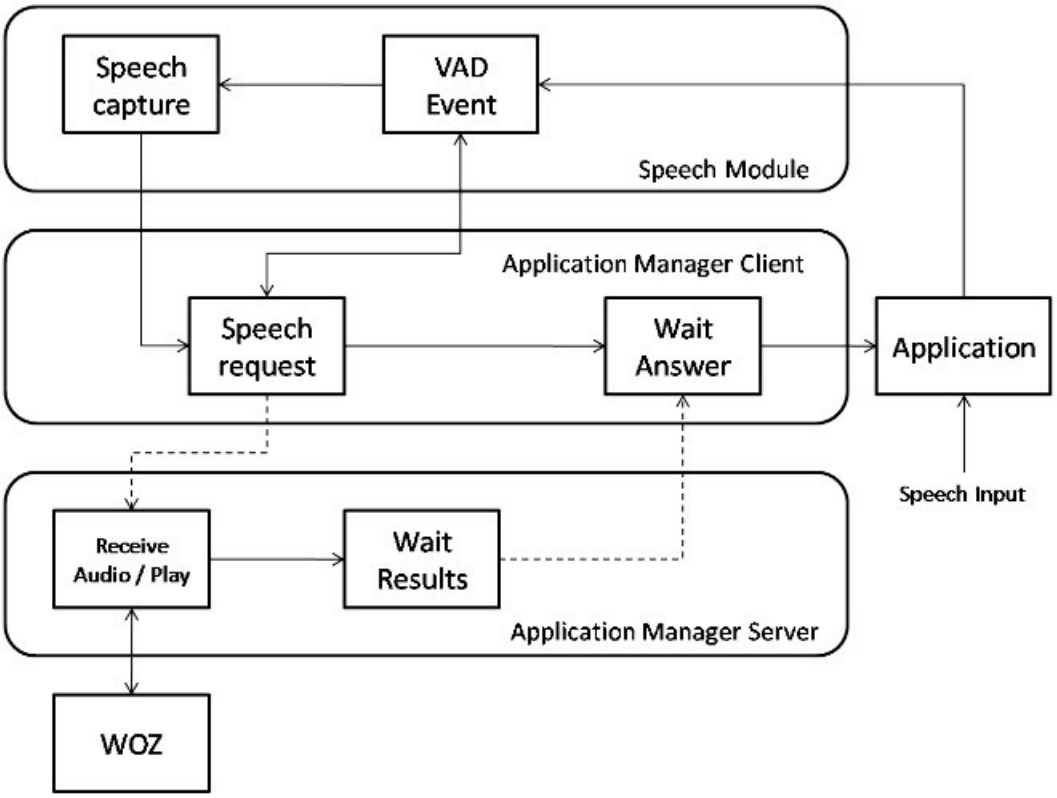


Figure 3.3: The process of a speech request through the system [72]

Consequently the usage of sound and animation helps to attract a child's attention [73]. Another important parameter, that we should have taken under consideration during the system design, was that the target group we study is, in most cases, at a preliterate level. As a result we should avoid using text as an output. Instead, text can be replaced by audio and/or visual effects [73].

As it is described in [73] and [72], there are five different games in our application, based on popular preschool activities. In brief, the selected games are: for children 4-6 years old, animal recognition, shape recognition, number recognition and quantity comparison, while there is one additional task for the 6 year olds, namely additions.

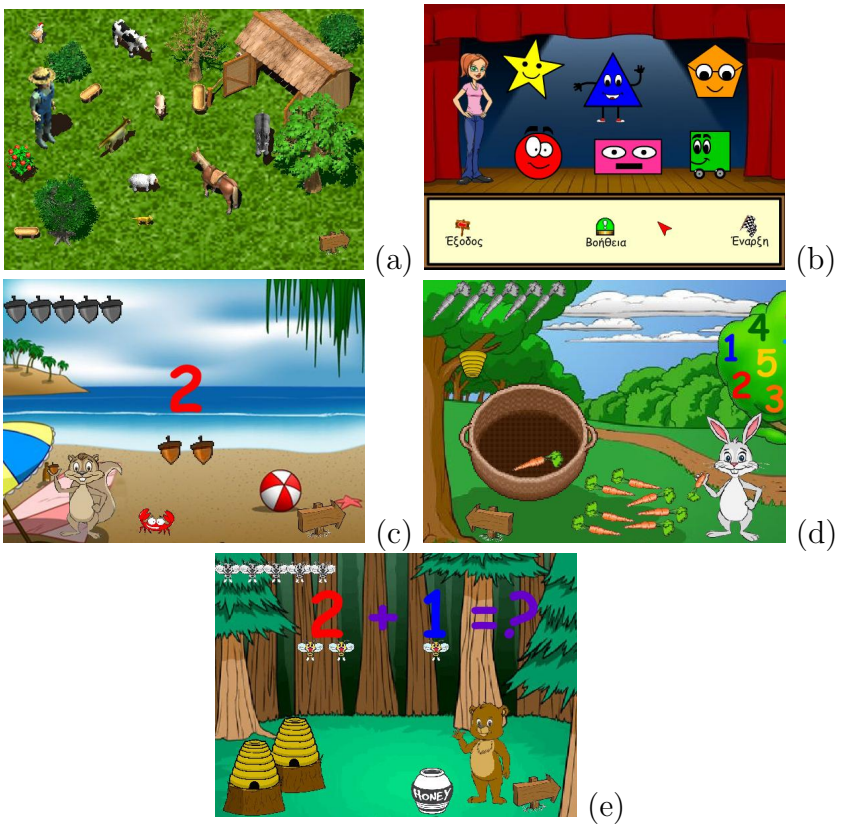
Previous research in multimodal systems, indicates that the existence of an animated character, makes interaction quite enjoyable, especially when this character takes a certain role in the game [67, 36]. So, for each task there is an embodied agent, which plays the role of guidance through the game.

Also children, especially 4 year olds, are not as much familiar with mouse and keyboard devices, as older children are. Instead, speech and touch are ideal input modalities for such ages, since they are closer to the patterns kids use in face-to-face interaction. Such input modalities it is well known that make interaction more natural [53], so our system provides two input modalities: mouse and voice, while as output graphics, animation, prerecorded prompts and synthesized text-to-speech prompts are used. Specifically the implemented tasks are:

- The animal recognition task, takes place in a farm. Here the embodied agent (farmer) urges the child to select the appropriate animal, after hearing its voice, in order to guide it into the stable. The child has to select among nine different animals (Fig. 3.4(a)).

- The shape recognition task, takes place in a theater. Each time one of six, at total, shapes appears on the stage, and the embodied agent (teacher) asks the child to recognize the shape (Figure 3.4(b)).

- The number recognition task, takes place on the beach, where the embodied agent (squirrel) asks the child to recognize the number (1-9), which appears on the screen (Figure 3.4(c)).



(a) (b) (c) (d) (e)

Figure 3.4: Example screen-shots of the five tasks: (a) animal recognition, (b) shape recognition, (c) number recognition, (d) quantity comparison, and (e) additions. [72]

- The quantity comparison task, takes place in the countryside. Here the embodied agent (rabbit), puts some carrots inside and some outside the basket, and then asks the child to determine where are the most/less carrots (Figure 3.4(d)).

- Finally, the addition task takes place in the forest. Here the child is asked to help an embodied agent (bear) to fill his jar with honey from the beehives. A simple addition (up to 9) appears on the screen, and for each correct answer supplied by the child, a bee drops honey into the jar (Figure 3.4(e)).

3.3 Define Fantasy, Curiosity and Challenge

In this section we are going to describe how the different levels of fantasy, curiosity and challenge were implemented for the five tasks. According to Malone [78] there are three different factors, namely fantasy, curiosity and challenge, that make computer games fun. Table 3.1 presents detailed information about the three factor levels as they are implemented in the games application.

3.3.1 Fantasy Definition

Fantasy usually makes games more interesting. This is attributed to the fact that, the main property of fantasy factor is that during game play the players undertake a certain role. According to Malone [78], there are two types of fantasy: the intrinsic and the extrinsic fantasy. As intrinsic fantasy, Malone defines that type of fantasy where there is a two way dependence between fantasy and user skill. On the contrary extrinsic fantasy is defined as a subset of the intrinsic, where only the fantasy depends on the skill, while the opposite is not valid. In our application, we have implemented the intrinsic type of fantasy, by inserting a certain goal into the existing games. The goal is to help an alien crashed on earth return to his home planet (Figure 3.5(a)). Moreover, there are short animations in order to trigger the child's fantasy. For example, for the numbers recognition task (Figure 3.5(b)), the crab starts walking around making noises when the child clicks on the crab. There are three fantasy levels: at fantasy level 0 there is neither story nor fantasy triggers, at fantasy level 1 there is story but no fantasy triggers, while at fantasy level 2 there is story and fantasy triggers.

3.3.2 Challenge Definition

The challenge (or else difficulty) factor is very important in a game. According to Malone [78], a game is challenging if it provides a goal, whose achievement is not an obvious task. If the goal is too obvious, the game becomes predictable and consequently boring. On the other hand, if it is too hard for the players to achieve the goal, it is likely they get disappointed and therefore demotivated. Well-designed computer games, take under consideration the characteristics of

(a)

(b)

Figure 3.5: (a) Intrinsic fantasy (b) Fantasy triggers

challenge, and offer multiple levels of challenge, meaning that a level that seems easy to one player may seem challenging to someone else. So for each of the five tasks we implemented three different levels of challenge [72, 73]. For example in the additions task, at challenge level 0 and level 1 the system asks additions resulting up to five and five to nine, respectively. At both levels 0 and 1 the system provides also helping items underneath the addition. Finally, at challenge level 2 the system asks additions resulting up to nine, without providing any help.

3.3.3 Curiosity

Curiosity is defined as the motivation of learning. According to Malone [78] there are two types of curiosity: sensory and cognitive curiosity. Sensory curiosity is defined as the attraction to the environment. In most computer games sensory curiosity is triggered by using sounds and animations. Cognitive curiosity, on the other hand, is described as the desire of someone to enrich or develop their knowledge. Surprising feedback is an easy way to trigger users' curiosity, but feedback should be also constructive (for example rewards), meaning that it should help

Figure 3.6: (a) No answers bar and randomness, (b) with answers bar but no randomness, (c) with answers bar and ran- domness

users' to elevate and complete their knowledge level. Taking under consideration those principles, we implemented a progress bar at the top of the screen representing the number of correct answers. In our application progress bar is a way to reward the child, since each time they give a correct answer, they earn the item which appears in the bar. Furthermore, when the child completes successfully the game, there is also a "positive" reward (sound of clapping). As Malone emphasizes in [78], an easy way to implement surprising feedback, is by randomness. In our case, according to the curiosity level, the animated characters in each game appear in a random sequence. In brief, there are three levels of curiosity: at level 0 there is neither progress bar nor randomness, at level 1 there is progress bar and finally at level 2 there is progress bar and randomness (Figure 3.6).

By labeling the factors' levels as 0, 1 and 2 is not very sound. In fact the scale among the three factor levels is ordinal, rather than linear. Meaning that the distance from level 0 to level 1 is not the same as from level 1 to level 2. Nevertheless, this fact does not affect the classification among the preferred factor levels, because, as we will explain next, we use only two classes.

Level	Fantasy	Curiosity	Challenge				
			Farm	**More/Less**	**Numbers**	**Addition**	**Shapes**
0	No story or fantasy triggers	No correct answers bar and no randomness	Select from 5 different animals	Item difference is 6-8	Numbers from 1-5 with item help	Add up to 2-5 with item help	Star, circle, square
1	Story but no fantasy triggers	Correct answers bar but no randomness	Select from 7 different animals	Item difference is 3-5	Numbers from 5-9 with item help	Add up to 5-9 with item help	Star, circle, square, triangle
2	Both story and fantasy triggers	Both correct answers bar and randomness	Select from 9 different animals	Item difference is 1-2	Numbers from 1-9 without item help	Add up to 2-9 without item help	Star, circle, square, triangle, rectangle and pentagon

Table 3.1: The three levels of fantasy, curiosity and challenge as implemented in our application. Implementation of challenge is task dependent.

Chapter 4

Modeling Fantasy, Curiosity and Challenge

4.1 Features

In this survey, we investigate a variety of features that can be extracted from the child-machine interaction laps, as well as, audio features extracted from the child's speech input.

4.1.1 Objective Features

During game play the system stores various statistics per task. Such statistics are modality usage, times, number of correct/wrong answers etc, namely objective criteria or objective features. So the objective features used in this work contain user information, such as child's age in years and child's gender, as well as, interaction information. During the game the system asks the user various questions, and it stores the number of questions answered correctly. Also, since the interaction is multimodal, the system stores information about the input modality (mouse or speech) the child used during the interaction. To be more specific the system keeps how many times during a task the child selected speech (Voice Input) and how many times they used the mouse input device (Mouse Input). By using this information the percentage of voice used, during one whole session

(five tasks), is calculated simply as $\frac{VoiceInpout}{VoiceInput+MouseInput} * 100$. Moreover, important time statistics were calculated. Response time is defined as the time that elapses from the end of a system prompt until the child completes their answer (stop talking or clicking the mouse on a valid target). We separated the response time to inactivity time (time form the end of a system prompt until first voice or mouse activity detection) and interaction time (response minus inactivity time) [65]. Finally, the system stores information about whether the user completed a task or not, as a binary value. The list with the objective features follows:

- Child's Age (in years)
- Child's Gender
- Number of Correct Answers
- Voice Use (%)
- Average inactivity time (in ms)
- Average interaction time (in ms)
- Average response time
- Task completion (%)
- Minimum response time (in ms)
- Maximum response time (in ms)

4.1.2 Audio Features

Besides objective features, the system creates and stores audio files, per task, according to the answers given by the child. By these audio files we extracted audio features, using two different feature extraction toolkits (Praat [57] and openSMILE [28]). Table 4.1 shows the audio features calculated by the two feature extraction toolkits.

Praat audio features		openSMILE audio features	
Low-level descriptors	Statistics	Low-level descriptors	Statistics
Pitch	Minimum	rms Energy	Minimum
	Maximum	MFCC[1]	Maximum
	Range	Voicing Probability	Range
	Mean	Fundamental Frequency	Mean
	Median		Standard deviation
	Standard deviation		Kurtosis
	Baseline		Skewness
	Alternative baseline		
Intensity	Minimum		
	Maximum		
	Mean		
	Standard deviation		
Energy			
Duration			
F0 points			

Table 4.1: Audio features

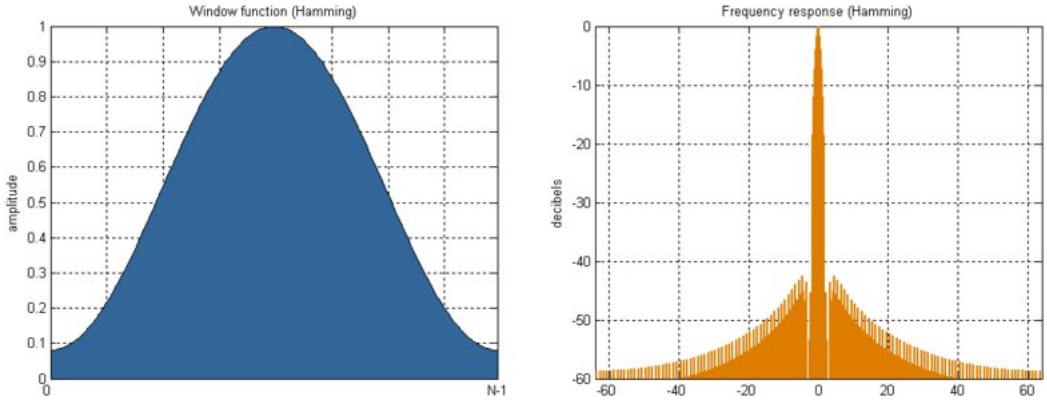


Figure 4.1: Hamming window

4.1.2.1 openSMILE Audio Features

The Munich open Speech and Music Interpretation by Large Space Extraction (openSMILE) toolkit is a modular and flexible feature extractor for signal processing and machine learning applications. It is used to extract low-level audio features or low-level descriptors (llds), such as fundamental frequency, energy, probability of voicing, MFCC etc, for incremental on-line affect analysis.

The openSMILE takes as input an audio file, and creates frames of size 25ms and frame sampling period 10ms. The values of each frame are multiplied with a window function. In this case, a Hamming window: $w_{Ham}[n] = 0.54 + 0.46cos(\frac{2\pi n}{N-1})$ is used (Figure 4.1).

And then, by each windowed frame we generate the MFCC (Mel Frequency Cepstral Coefficients). in general the Mel frequency scale is preferred, because it approximates the human auditory system's response more closely than the linearly-spaced frequency bands (Figure 4.2).

In sound processing, the mel-frequency cepstrum (MFC) is a representation of the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a non-linear mel scale of frequency. Mel-frequency cepstral coefficients (MFCCs) are coefficients that collectively make up an MFC. They are derived from a type of cepstral representation of the audio clip (a non-linear "spectrum-of-a-spectrum"). The difference between the cepstrum and the mel-frequency cepstrum is that in the MFC, the frequency bands are equally spaced

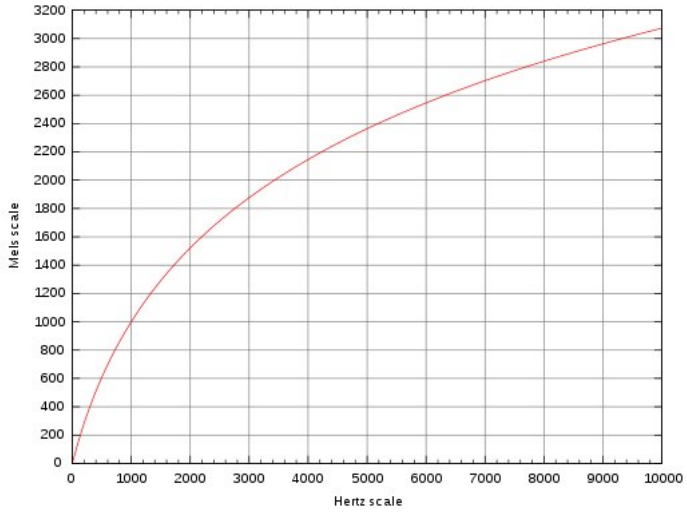


Figure 4.2: Pitch mels versus hertz

on the mel scale, which approximates the human auditory system's response more closely than the linearly-spaced frequency bands used in the normal cepstrum. This frequency warping can allow for better representation of sound, for example, in audio compression.

MFCCs are commonly derived as follows:

- Take the Fourier transform of (a windowed excerpt of) a signal.
- Map the powers of the spectrum obtained above onto the mel scale, using triangular overlapping windows.
- Take the logs of the powers at each of the mel frequencies.
- Take the discrete cosine transform of the list of mel log powers, as if it were a signal.
- The MFCCs are the amplitudes of the resulting spectrum.

We also compute the fundamental frequency and the probability of voicing. The fundamental frequency, often referred to simply as the fundamental and abbreviated as F0, is defined as the lowest frequency of a periodic waveform.

In terms of a superposition of sinusoids (e.g. Fourier series), the fundamental frequency is the lowest frequency sinusoidal in the sum.

Additionally, voicing is a term used in phonetics and phonology to characterize speech sounds, with sounds described as either voiceless (unvoiced) or voiced. The term, however, is used to refer to two separate concepts. Voicing can refer to the articulatory process in which the vocal cords vibrate. This is its primary use in phonetics to describe phones, which are particular speech sounds. It can also refer to a classification of speech sounds that tend to be associated with vocal cord vibration but need not actually be voiced at the articulatory level. This is the term's primary use in phonology when describing phonemes, or in phonetics when describing phones.

At the articulatory level, a voiced sound is one in which the vocal cords vibrate, and a voiceless sound is one in which they do not. For example, voicing accounts for the difference between the pair of sounds associated with the English letters "s" and "z". The two sounds are transcribed as [s] and [z] to distinguish them from the English letters, which have several possible pronunciations depending on context. If one places the fingers on the voice box (i.e. the location of the Adam's apple in the upper throat), one can feel a vibration when one pronounces zzzz, but not when one pronounces ssss.

Finally, we calculate the Root Mean Square (RMS) signal energy: $E_r = \sqrt{\frac{\sum_{n=0}^{N} x_n^2}{N}}$, for a frame duration of 25ms. Each of the llds is smoothed by using a Simple Moving Average (sma) window, of size 3 frames. Meaning that every three frames, we have a replacement in values by their average value. For each smoothed lld, we calculate various statistics such as, maximum value, minimum value, range ($range = Maximum - Minimum$), mean value ($\mu = E\{X\} = \frac{\sum_{n=1}^{N} x_n}{N}$), standard deviation ($\sigma = \sqrt{E\{X^2\} - E\{X\}^2}$), kurtosis and skewness. In probability theory and statistics, kurtosis is any measure of the "peakedness" of the probability distribution of a real-valued random variable. In Figure 4.3 the kurtosis of well-known distributions are shown:

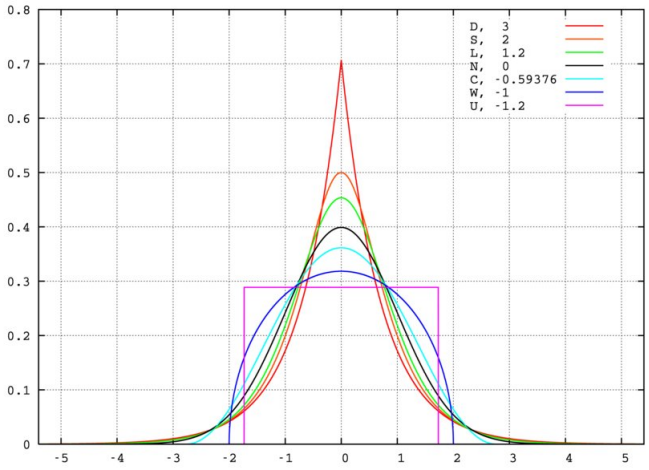


Figure 4.3: Kurtosis of well-known distributions

,where

- D: Laplace distribution, red curve, excess kurtosis = 3
- S: Hyperbolic secant distribution, orange curve, excess kurtosis = 2
- L: Logistic distribution, green curve, excess kurtosis = 1.2
- N: Normal distribution, black curve, excess kurtosis = 0
- C: Raised cosine distribution, cyan curve, excess kurtosis = -0.59
- W: Wigner semicircle distribution, blue curve, excess kurtosis = -1
- U: Uniform distribution, magenta curve, excess kurtosis = -1.2

In probability theory and statistics, skewness is a measure of the asymmetry of the probability distribution of a real-valued random variable. The skewness value can be positive or negative, or even undefined. Qualitatively, a negative skew indicates that the tail on the left side of the probability density function is longer than the right side and the bulk of the values (possibly including the median) lie to the right of the mean. A positive skew indicates that the tail on the right side is longer than the left side and the bulk of the values lie to the left of the mean. A zero value indicates that the values are relatively evenly distributed

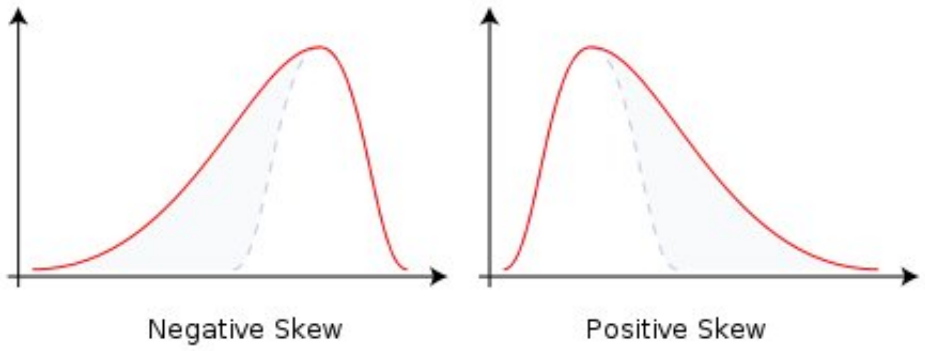


Figure 4.4: Negative and positive skewness

on both sides of the mean, typically but not necessarily implying a symmetric distribution (Figure 4.4).

4.1.2.2 Praat Audio Features

Praat is a free scientific software program for the analysis of speech in phonetics. It has been designed and continuously developed by Paul Boersma and David Weenink of the University of Amsterdam. The program also supports speech synthesis, including articulatory synthesis.

Praat takes an audio file as input and splits it into frames of duration of 10ms each. In order to calculate the pitch in each frame frequency values below 70Hz or above 400Hz are ignored and a periodicity detection algorithm on the basis of an accurate autocorrelation method is performed. As described in [] the algorithmic steps are:

1. Preprocessing: to remove the sidelobe of the Fourier transform of the Hanning window for signal components near the Nyquist frequency, they perform a soft upsampling as follows: do an FFT on the whole signal; filter by multiplication in the frequency domain linearly to zero from 95% of the Nyquist frequency to 100% of the Nyquist frequency; do an inverse FFT of order one higher than the first FFT.

2. Compute the global absolute peak value of the signal.

3. The analysis is performed for a number of small segments (frames) that are taken from the signal in steps given by the TimeStep parameter (default is 0.01 seconds). For every frame, look for at most MaximumNumberOfCandidatesPerFrame (default is 4) lag-height pairs that are good candidates for the periodicity of this frame. This number includes the unvoiced candidate, which is always present. The following steps are taken for each frame:

   (a) Take a segment from the signal. The length of this segment (the window length) is determined by the MinimumPitch parameter, which stands for the lowest fundamental frequency that you want to detect. The window should be just long enough to contain three periods (for pitch detection) or six periods (for HNR measurements) of MinimumPitch. E.g. if MinimumPitch is 75 Hz, the window length is 40 ms for pitch detection and 80 ms for HNR measurements.

   (b) Subtract the local average.

   (c) The first candidate is the unvoiced candidate, which is always present. The strength of this candidate is computed with two soft threshold parameters. E.g., if VoicingThreshold is 0.4 and SilenceThreshold is 0.05, this frame bears a good chance of being analyzed as voiceless (in step 4) if there are no autocorrelation peaks above approximately 0.4 or if the local absolute peak value is less than approximately 0.05 times the global absolute peak value, which was computed in step 2.

   (d) Multiply by the window function ($a(t) = (x(t_{mid} - \frac{1}{2}T + t) - \mu_x)w(t)$).

   (e) Append half a window length of zeroes (because we need autocorrelation values up to half a window length for interpolation).

   (f) Append zeroes until the number of samples is a power of two.

   (g) Perform a Fast Fourier Transform.

   (h) Square the samples in the frequency domain.

   (i) Perform a Fast Fourier Transform. This gives a sampled version of $r_a(\tau)$.

(j) Divide by the autocorrelation of the window, which was computed once with steps 3e through 3i. This gives a sampled version of $r_x(\tau)$.

(k) Find the places and heights of the maxima of the continuous version of $r_x(\tau)$

4. For every frame n, p_n is a number between 1 and the number of candidates for that frame. The values $\{p_n|1 \leq n \leq number$ of frames$\}$ define a path, and every possible path is associated with a cost. The globally best path is the path with the lowest cost.

Statistics of pitch such as, minimum pitch value, maximum pitch value, mean pitch, median pitch, standard deviation of pitch, baseline pitch (calculated as the mean pitch - 1.43 standard deviation of pitch), an alternative baseline of pitch (calculated as the estimation of the pitch value below which the 7.62% of all pitch values are expected to lie) and the range of pitch (calculated as $PitchRange = maximumPitch - minimumPitch$), are calculated. Also F0 points are calculated as the number of glottal pulses within a voiced interval. Moreover, statistics of intensity, such as minimum intensity, maximum intensity, mean intensity and standard deviation of intensity, are calculated. Finally, energy and audio file duration (in sec) are also calculated.

4.1.2.3 Emotion Classification Features

For emotion recognition we use features extracted by users' speech. The audio features are the same with those presented in Table 4.1.

The emotions can be imprinted onto a two dimensional space, where one dimension is valence and the other is arousal [39, 20, 59, 45]. Valence, as used in psychology, especially when discussing about emotions, means the intrinsic attractiveness (positive valence) or aversiveness (negative valence) of an event, object or situation [55]. However, the term is also used to characterize and categorize specific emotions. On the other hand, arousal is a physiological and psychological state of being reactive to stimuli. In emotion theory, arousal indicates the experience of excitation. In the current work, the four emotions are separated into two large categories: the one, namely positive/negative, indicates the valence and the

second, namely active/passive, indicates the arousal. These two categories are going to be examined separately.

4.2 Modeling and Feature Normalization

The models that we use for the optimal factor and the emotion classification problems are two linear classifiers (naive Bayes and 3-NNR), a neural network and a support vector machine (SVM). The main idea is to use two simple classifiers and compare them with two more sophisticated.

4.2.1 naive Bayes

The naive Bayes is a simple probabilistic classifier based on applying the Bayes' theorem with strong (or else "naive") assumptions. The classifier assumes that the presence/absence of a particular feature of a class is unrelated to the presence/absence of any other feature. Even if these features depend on each other, the naive Bayes classifier considers all of these properties to independently contribute to the probability that a sample belongs to a certain class. The naive Bayes classifier combines naive Bayes probability model with a decision rule. One common rule is to pick the hypothesis that is most probable; this is known as the maximum a posteriori or MAP decision rule. The corresponding classifier is the function *classify* defined as follows:

$$classify(f_1, ..., f_n) = \arg\max_c p(C = c) \prod_{i=1}^{n} p(F_i = f_i | C = c)$$

4.2.2 3-NNR

The 3-nearest neighbors algorithm (3-NNR) is a method for classifying objects based on the three closest training samples in the feature space. In the current work the 3-NNR algorithm is using the Euclidean distance metric. The Euclidean distance between points p and q is:

$$d(p, q) = d(q, p) = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}$$

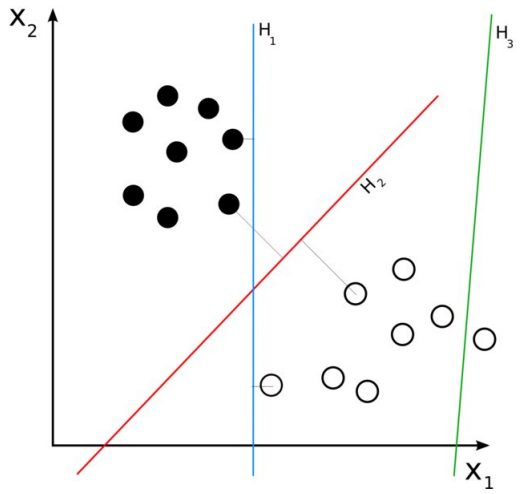


Figure 4.5: H3 (green) doesn't separate the two classes. H1 (blue) does, with a small margin and H2 (red) with the maximum margin

4.2.3 Neural Network

The neural network we use in the current work, is a multilayer perceptron classifier (MLP). The MLP is a feedforward artificial neural network model that maps sets of input data onto a set of appropriate output. The MLP consists of three layers (an input and an output layer with one hidden layer) of nonlinearly-activating nodes. MLP utilizes a supervised learning technique called backpropagation for training the network. MLP is a modification of the standard linear perceptron and can distinguish data that is not linearly separable.

4.2.4 Support Vector Machine

A support vector machine (SVM) is a supervised learning method that analyzes data and recognizes patterns, used for classification analysis. In this case the SVM uses a linear hyperplane for classification (Figure 4.5).

All the four classifiers mentioned above, are used as the WEKA [48] data mining software provides them [48]. We used the WEKA data mining software as the main scope of this work is not centered on the implementation of classification

methods, but we are mainly interested in parameters prediction using well known models.

4.2.5 Feature Selection

In search of better features, we also conducted feature selection. The feature selection method we used, is the wrapper method.

The wrapper method uses a searching algorithm to search through the space of possible features and evaluate each subset by running a model on the subset. The searching algorithm we use is the best first, that is provided by the WEKA data mining software. Best-first search is a search algorithm which explores a graph by expanding the most promising node chosen according to a specified rule.

Best-first search estimates the promise of node n by a heuristic evaluation function $f(n)$ which, in general, may depend on the description of n, the description of the goal, the information gathered by the search up to that point, and most important, on any extra knowledge about the problem domain.

Although, best first is a greedy search algorithm, is fast and the features selected by this method are quite promising compared to other search algorithms (e.g. exhaustive search).

4.2.6 Normalization

Additional to the feature selection we also performed feature normalization. Preliminary analysis, indicated that the task each child played, seems to be an important factor, which influences the child's behavior during game play.

Normalization is the process of isolating statistical error in repeated measured data. Also by normalization we allow underlying characteristics of the data on different scales to be compared, by bringing them to a common scale. The type of normalization we used is the z-normalization. It is derived by subtracting the population mean from an original observation value and then dividing the difference by the population's standard deviation: $z = \frac{x-\mu}{\sigma}$.

4.3 Experimental Methodology

4.3.1 Subjects

Fifteen preschool children, of both genders and ages 4-6 participated, in this survey by playing different versions of the application. All the children are native Greek speakers. Six of them were girls and nine boys. The users distribution relatively to age and gender is shown in the Table 4.2.

	AGE			
	Four	Five	Six	Total
Female	1	2	3	6
Male	2	4	3	9
Total	3	6	6	15

Table 4.2: Users distribution relatively to age and gender.

4.3.2 Experimental Procedure

The experimental procedure took place in a noisy preschool environment, where each child played different versions of the application. Before the experimental process starts we let the child to familiarize with the application, by playing a demo session using both mouse and voice. After finishing the demo session, each child was asked to play three different versions (or else sessions in the rest of the thesis) of the game and then choose the most entertaining according to their opinion. At each session only the value of one factor was modified (taking level values 0, 1, or 2), while the other two factors remained constant to level value 1. Level value 1 is chosen as it represents an intermediate state, where the factor's level contribution is not too obvious to the user. Meaning that level values 0 and 2 are those which affect the user's interest, negatively or positively.

During game play children believed that they were interacting with a fully automated system, meaning that they had no knowledge about the existence of the wizard. At each session children played at least once all the tasks, that are appropriate for their age. At the end of the whole process, each child had played

a total of nine different application set-ups. The order that each factor and each factor level was presented to the child was random.

4.3.3 Emotion

We have five different emotions: angry, happy, neutral, boring and sad. Analysis, indicated that neutral emotion is not important, since it does not indicate the user's emotional state. An emotion can be positive (e.g. happy), negative (e.g. sad, agree and boring), active (e.g. happy and agree) or passive (e.g. sad and boring), depending on what processes it activates in the human body.

Three post graduate students were asked to label the collected audio data. To be more specific, the three labelers, independently, had to assign each recorded answer into one of the five emotions according to their opinion. In our experiments we use separately the data that all the three labelers agreed on, and the data that at least two labelers agreed on. The number of utterances that resulted by the previous procedure is shown in Table 4.3.

	At-least two agree	Three agree
Positive utterances	233	82
Negative utterances	88	19
Active utterances	255	89
Passive utterances	66	12
Total emotional utterances	321	101

Table 4.3: Number of emotional utterances

Figure 4.6 presents the valence-arousal distribution normalized on age:

We can observe that in Figure 4.6 there is a "V" structure, same as proposed in [42]. The discrete image that the distribution displays is attributed to the fact that we consider emotions as discrete points onto the 2-dimensional space, and not as whole regions.

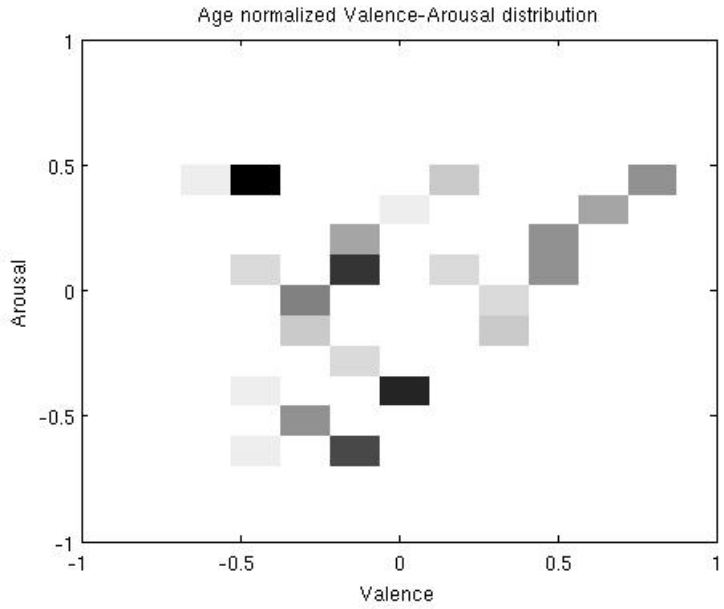


Figure 4.6: Distribution of valence/arousal for age normalized data

4.4 Results

In this section we present the results from the experimental process. By analyzing the results, we attempt to deduce helpful information about the interaction patterns used by children during game play, and their preferences. Additionally the prediction of affect and optimal factor level will help us build reliable user models.

It is important to establish what elements are those, which affect the interaction patterns formation in early ages. To be more specific, we want to present in what extend emotion affects the interaction and the child's perception about what is fun or not.

4.4.1 System Evaluation Results

The evaluation of multimodal dialogue systems is a complex task and different metrics (objective and subjective) are typically used to evaluate different aspects of such systems.

In Table 4.4, a summary of the objective evaluation metrics is shown, as a

function of age and gender. Specifically, for each age and gender, average response, average inactivity and average interaction times, as well as speech/mouse input and task completion percentages are shown. It is clear that speech usage decreases with age, while mouse usage increases with age. Also older children display lower interaction time during gaming. As far as the gender is concerned, results indicate that there are not significant differences during the interaction.

	AGE			GENDER	
	4	5	6	Male	Female
Avg. Response Time(sec)	5.21	4.69	4.20	4.56	4.47
Avg. Inactivity Time(sec)	1.18	1.24	1.32	1.16	1.41
Avg. Interaction Time(sec)	4.02	3.45	2.88	3.40	3.06
Mouse Usage(%)	16.7	19.48	19.11	19.30	18.40
Speech Usage(%)	83.30	80.52	80.89	80.68	81.6
Task Completion(%)	93.62	97.5	98.19	96.15	98.68

Table 4.4: Objective metrics per age and gender

As Table 4.4 indicates there is no significant difference in interaction patterns between the two genders. This yields that gender is not a significant factor, as far as the interaction patterns are concerned in this small sample of children. Age seems to affect the interaction patterns formation. Significant differences are displayed mostly by the four year old children, especially in input modality usage and task completion. To be more specific, younger children prefer mostly speech as input modality and less mouse usage, compared to older ones. This is attributed to the fact that they are not well familiarized with the mouse device usage, in comparison to older children. Also four year olds display low inactivity and high interaction times, meaning that they tend to give more spontaneous and descriptive answers. Another important outcome is that younger children tend to be less successful in task completion, than older ones. This proves that the younger the child is, the harder is to keep their interest constantly in high levels.

These results are also being verified by the correlations in Table 4.5. In Table 4.5, correlations among various objective metrics and age are presented along with their p-values. Correlations, accord important information about the user patterns, relatively to age and gender. We observe that there is high negative

correlation between average response time and age, and high positive correlation between average response time and wrong answers. Moreover, age and mouse skill are highly correlated, as well as gender and inactivity time.

	Corr. Coef.	p-value
Avg. Response Time/Age	-0.29	0
Mouse Skill/Age	0.17	0.06
Avg. Response Time/Mouse Skill	-0.13	0.13
Avg. Response Time/Wrong Answers	0.24	0.01
Inactivity Time/Gender	0.34	0

Table 4.5: Correlations for time, mouse skill and age

Table 4.5 indicates that there is negative correlation between average response time and mouse usage, which is attributed to the fact that kids are not well familiar with the mouse input device. As a result it takes them more time to complete a request using the mouse rather than just speak to the microphone. Moreover, there is a positive correlation between the number of wrong answers and the average response time, meaning that kids when make a mistake they become more cautious.

In Table 4.6, correlations among various objective metrics and speech usage are presented. By this matrix we expect to conclude how speech is used by infants during the game and how speech influences the interaction. The results indicate that the number of wrong answers and task completion are highly correlated with the speech input modality. Also as it is expected there is negative correlation between speech and age.

Since our survey is concentrated mainly on the speech modality, it is important to study how speech as input usage influences the interaction. The correlations in Table 4.6, indicate that there is high negative correlation between speech and age, which is quite rational as the younger the child is, the less familiar is with the mouse input device. There is high positive correlation between speech and task completion, meaning that the existence of speech as input modality, constitutes a strong motivator for child users to complete a task. This is attributed to the fact that, speech is a natural mean of communication, and gives the child the illusion

	Corr. Coef.	p-value
Speech Usage/Age	-0.17	0.06
Speech Usage/Avg. Response Time	0.14	0.13
Speech Usage/Interaction Time	0.07	0.41
Speech Usage/Wrong Answers	0.24	0.01
Speech Usage/Task Completion	0.41	0

Table 4.6: Correlations for speech usage

of interacting with an other human, instead of a machine. Also, there is hight positive correlation between speech and the number of wrong answers. This is due to the fact that, infants are spontaneous when talking and as a result they often make mistakes.

In Figure 4.7, the average response times per task and age, along with the 95% confidence intervals (to indicate the reliability of an estimate) are shown. This figure indicates that average response time, in general, drops with age. The most indicative image is that of the numbers task, where there is significant difference in average response time among the three ages. Also four and six year olds display the highest and lowest variances in all tasks respectively.

In Figure 4.7 the results are quite clear. Four year olds display mainly the highest response time in most tasks, with the most representative being the numbers recognition task. This is attributed to the fact that younger children are less familiar to the numbers, and they need more time to recognize the displayed number. Also variances indicate that the older the child is, the more consistent the time patterns are.

In Figures 4.9, 4.10 and 4.8 times, modality usage and task completion per age, gender and task, along with the 95% confidence intervals, are shown.

Such figures are going to give us a visual representation about how much and in which way interaction varies according to age, gender and task. The task completion, in general, does not vary significantly. Also we observe that mouse skill increases with age, while speech usage decreases with age. As far as the tasks and the modality usage are concerned, the most preferred input modality is that of speech. Specifically, the highest mouse usage is displayed for the farm task,

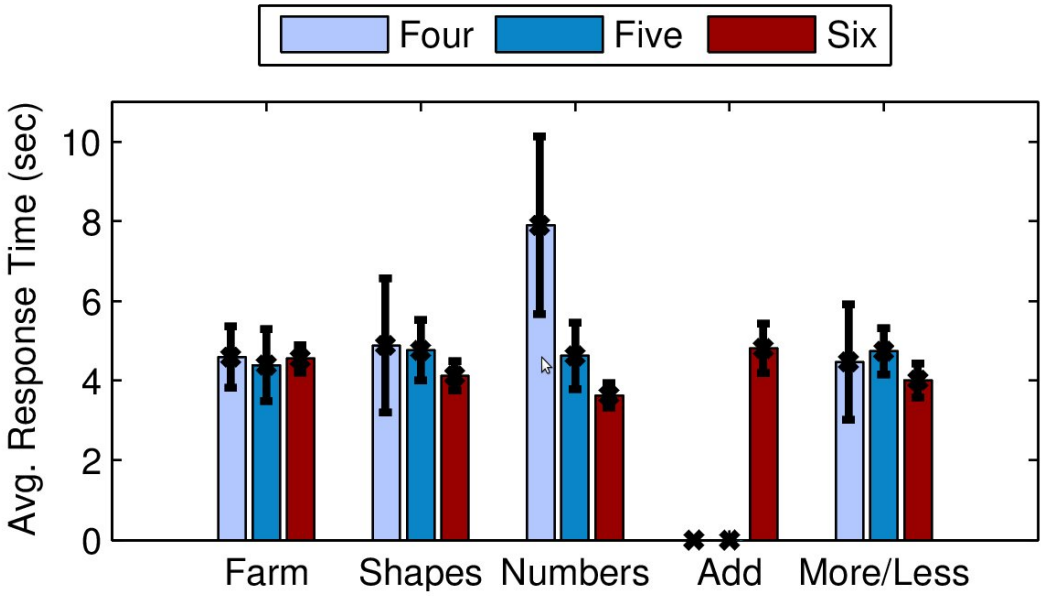


Figure 4.7: Avg. Response time per task and age

as it is more obvious for the child, that they can click on the displayed items. Something else that is noteworthy, is the time that users devote during the game. There is an increase in inactivity time for older children, while interaction time decreases with age. As it is expected there is high inactivity time during the additions task, because of the difficulty that this task displays. Finally, we note that there are not significant differences in interaction times, between the two genders.

To be more specific, Figure 4.10, indicates that interaction variations between the two genders are not significant. Also, Figure 4.9 provides us a visual representation of the changes in interaction patterns with respect to the age. Higher inactivity time in six year olds, indicates that older children are more cautious before giving an answer and tend to give more exact answers, as low interaction time indicates. On the other hand, younger children tend to be more spontaneous and give more descriptive answers. Moreover, in Figure 4.8 we can see how each task contributes to the times. Difficult tasks, such as additions and quantity comparisons (more/less), is presumable to display higher inactivity times.

Also, the 95% confidence intervals over the mean values are not very sound.

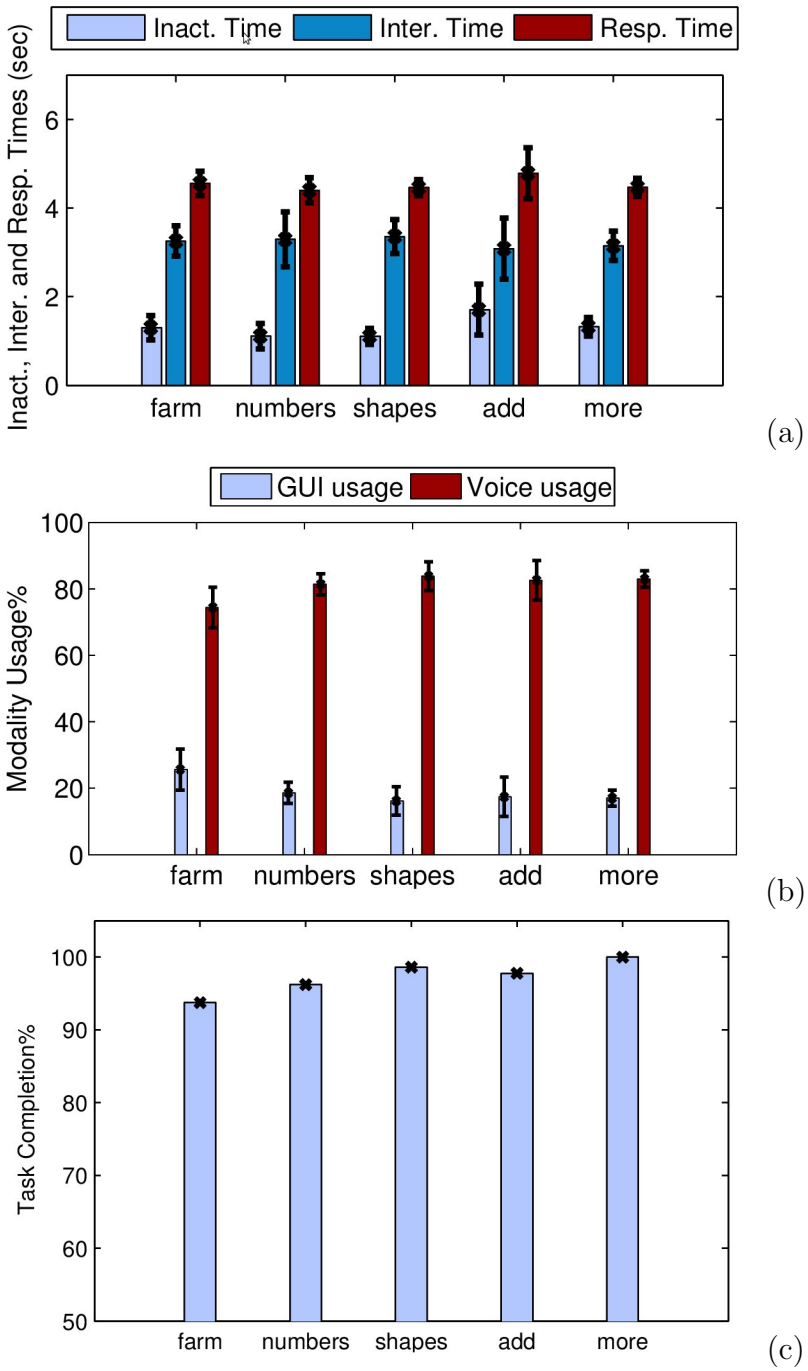


Figure 4.8: Various objective metrics per task (a) Response, Inactivity and Activity times per task, (b) Modality usage per task, (c) Task Completion per task.

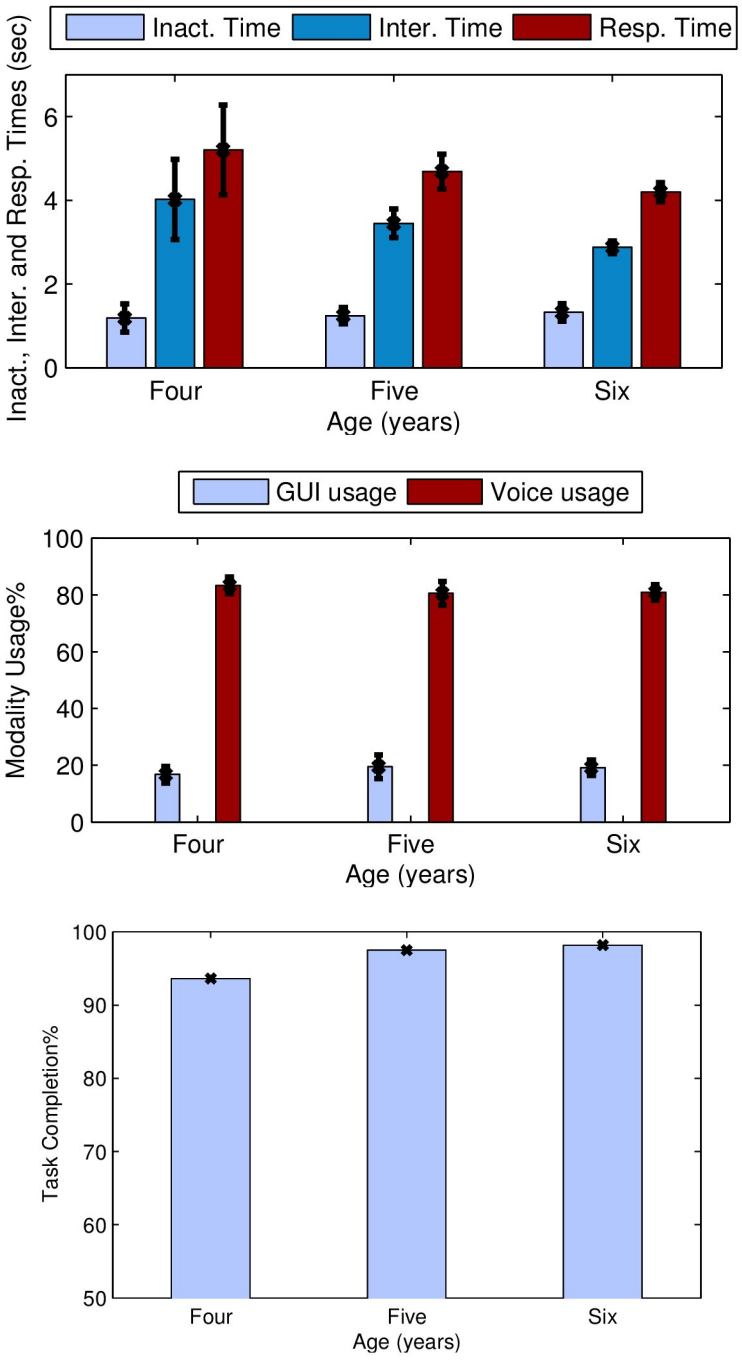


Figure 4.9: Various objective metrics per age (a) Inactivity, Interaction, Response times per age, (b) Modality usage per age, (c) Task Completion per age.

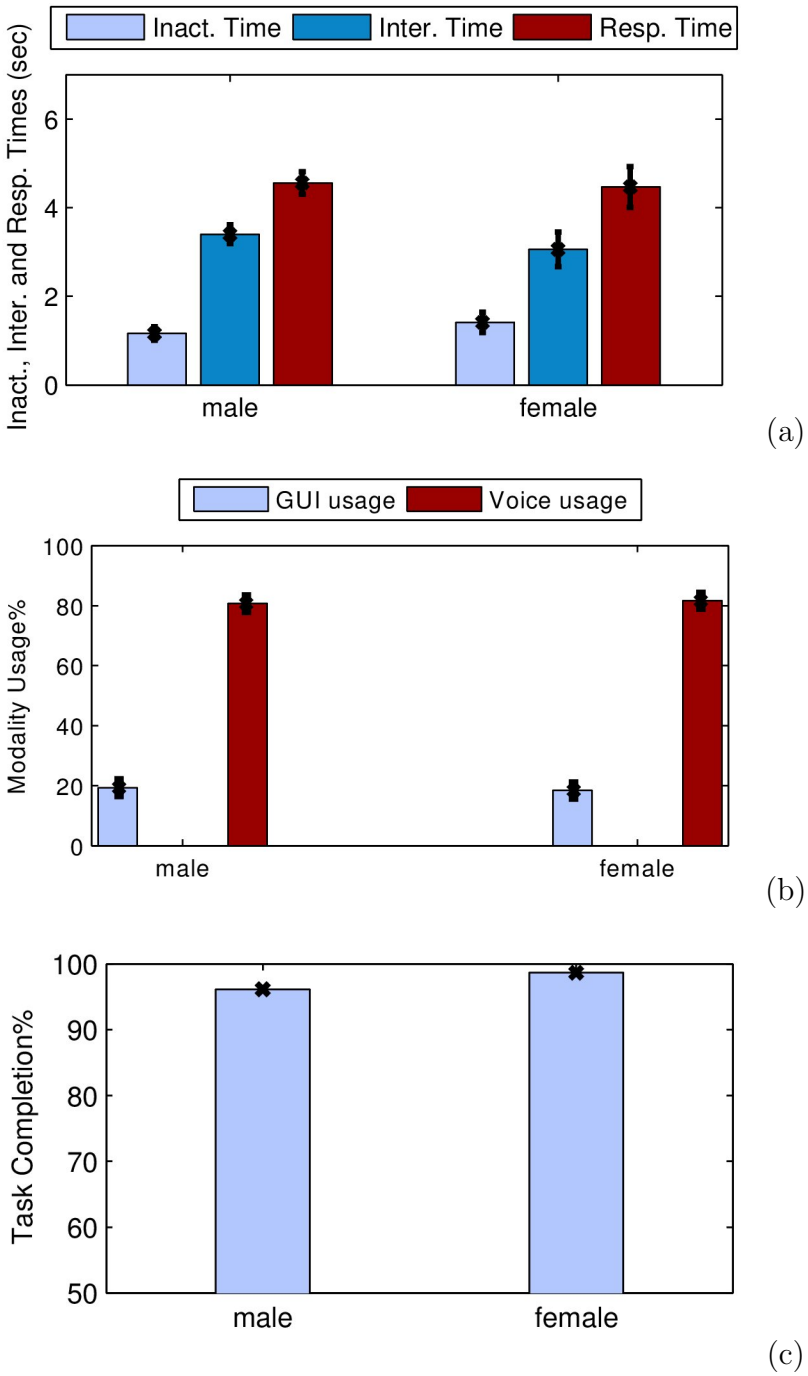


Figure 4.10: Different objective metrics per gender (a) Inactivity, Interaction, Response times per gender, (b) Modality usage per gender, (c) Task Completion per gender.

	Corr. Coef.	p-value
Fantasy/Entertainment	0.30	0
Curiosity/Entertainment	0.22	0.01
Challenge/Entertainment	0.17	0.05

Table 4.7: Correlations between entertainment and the three factors

This is attributed to the fact that we have very few users and each user interacted with the application at least 180 times. The means are calculated over the number of interactions (which is very large), so the user factor is exterminated. This is why we observe such small variances.

In Table 4.7, the correlations between the level of each factor (fantasy, curiosity, challenge) and entertainment (binary variable 1,0) are presented. By studying this table we can conclude how the three Malone factors contribute to users' entertainment. The table indicates that there is high positive correlation between the fantasy factor and the entertainment, as well as curiosity and entertainment. On the other hand challenge is moderately correlated with entertainment.

Table 4.7 represents the correlation between each Malone factor and entertainment, as it was selected by the users. Fantasy is high positively correlated with the entertainment, which indicates that high levels of fantasy make games more interesting. Specifically, in our game system, in high fantasy level the games are presented as part of a story, in which the child plays active role. Also curiosity displays high positive correlation with the entertainment. This is attributed to the fact that sounds and animations, as well as the existence of constructive feedback (rewards), attract users attention and make the game play more interesting. Finally, challenge has moderate correlation with entertainment. Challenge level in a game, may lead to gamers discouragement, as the less obvious a task's goal is, the more possible is the user abandon the task. In our case, mostly six year olds selected as more entertaining the most challenging level. This is attributed to the fact that older children can manage difficult tasks efficiently.

In Table 4.8, the correlation between the three Malone factors and various objective metrics is shown. This table is going to help us understand how fantasy, curiosity and challenge levels affect the interaction. By studying this table,

	Corr. Coef.	p-value
Fantasy/Speech usage	0.14	0.13
Curiosity/Avg. Interaction Time	0.17	0.05
Curiosity/Task completion	0.22	0.01
Challenge/Wrong Answers	0.17	0.06
Challenge/Avg. Inactivity Time	0.16	0.07

Table 4.8: Correlations between the three factors and objective metrics

someone can acquire a first insight about what makes a game entertaining. As it is expected challenge is positively correlated with the number of correct answers and the inactivity time. Also, interaction time and task completion are positively correlated with the curiosity factor.

Also correlations between the three Malone factors and different objective metrics are also examined, in Table 4.8. Curiosity and Task completion are highly correlated, meaning that the existence of attention stimuli motivates children to complete the task. Also, as it is expected there is positive correlation among challenge and the number of wrong answers given, as well as inactivity time. The more difficult a task is the more possible is someone to answer wrongfully and the more cautious become before giving a final answer. Last but not least, fantasy and speech usage are moderately correlated, meaning that fantasy triggers don't affect the input modality selection.

Also, Table 4.9 represents the correlation between various objective metrics and the emotion. This table aims to present how various objective metrics affect and/or are affected by the user's emotional state. The emotion percentages, when at least two labelers were agreed on, are used for the correlation calculation. As it is shown age is negatively correlated with positive valence and arousal dimensions. Also, times are positively correlated with positive emotion dimensions.

Table 4.9 presents the correlation between various objective metrics and emotion. To be more specific, correlations between age and emotion, indicate that as children grow up, they tend to be less enthusiastic. This is attributed to the fact that, older children have more expectations from the game system, in comparison to younger ones, who are easily impressed. Also the correlation between the user's

	Corr. Coef.	p-value
Age/Positive Valence (%)	-0.30	0
Age/Positive Arousal (%)	-0.24	0.01
Gender/Positive Arousal (%)	-0.17	0.07
Avg. Interaction Time/Positive Valence (%)	0.25	0.01
Avg. Interaction Time/Positive Arousal (%)	0.23	0.01
Avg. Response Time/Positive Valence (%)	0.25	0.01
Avg. Response Time/Positive Arousal (%)	0.23	0.01
Total Session Time/Positive Valence (%)	0.26	0
Total Session Time/Positive Arousal (%)	0.23	0.01

Table 4.9: Correlation between Objective metrics and Emotion

gender and positive arousal, indicate that females seem to be less active than the male users are. Moreover, average interaction time is positively correlated with emotion, at both positive arousal and valence dimensions. This indicates that as users are positively engaged with the game, they all the more want to interact with the application. The same outcome is also valid for the average response and total session times, which are also positively correlated with positive and active emotions.

Finally, Table 4.10 presents the correlations among audio features and the affect. Also here the affect results from agreement of two or three labelers. Results indicate that energy and pitch characteristics are highly correlated with emotion in both valence and arousal dimensions.

Results in Table 4.10 indicate that energy and pitch characteristics are highly correlated with the emotion, in both dimensions of valence and arousal. This confirms the statements in [59], that energy and pitch are good emotional indicators. Specifically, emotions that are positively stimulating are characterized by increase in pitch and speech energy.

	Corr. Coef.	p-value
Maximum Energy/Valence	0.50	0
Maximum Energy/Arousal	0.47	0
Standard Deviation of Energy/Valence	0.38	0
Standard Deviation of Energy/Arousal	0.44	0
Energy Skewness/Valence	0.29	0
Energy Skewness/Arousal	0.20	0
Energy Kurtosis/Valence	0.29	0
Energy Kurtosis/Arousal	0.19	0
Maximum F0/Valence	0.41	0
Maximum F0/Arousal	0.25	0
Standard Deviation of Pitch/Valence	0.36	0
Standard Deviation of Pitch/Arousal	0.40	0
Maximum Pitch/Valence	0.37	0
Maximum Pitch/Arousal	0.40	0

Table 4.10: Correlations between Audio features and Affect

	Positive/Negative	Active/Passive
At-least-two agree	72.6%	79.4%
Three agree	81.2%	88.1%

Table 4.11: Emotion classification baseline

4.4.2 Emotion Classification Results

Next we present the results from the emotion classification. Here we attempt to classify the emotions into the arousal and valence dimensions, using the audio feature sets, as we described them previously at 4.1.2.

As we mentioned in 4.1.2.3 the two emotion categories are going to be examined separately. In Figure 4.11 we present the mean valence and mean arousal differences among the factor level values. Meaning that:

$$Mean_d(factor = i) - Mean_d(factor = i + 1)$$

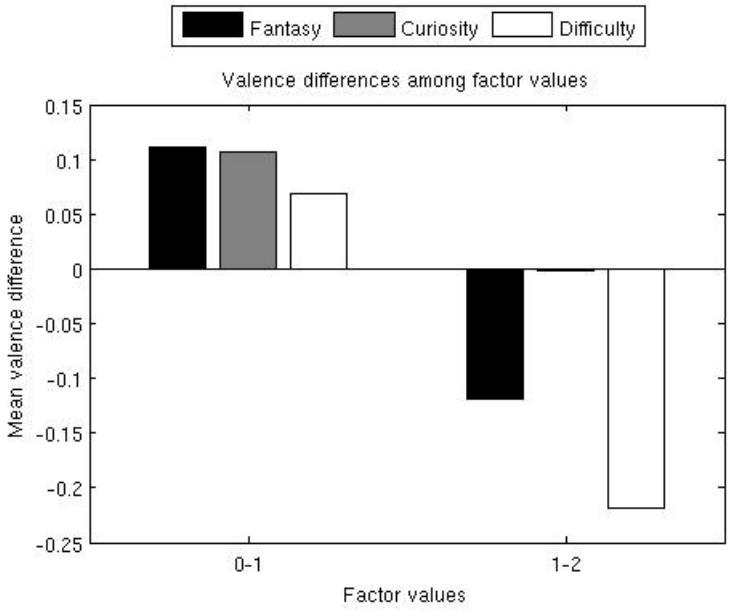
, where $d = \{Valence, Arousal\}$, $factor = \{Fantasy, Curiosity, Challenge\}$ and $i = \{0, 1\}$

As Figure 4.11(a) indicates, valence increases from factor value 1 to factor value 2 for factors Fantasy and Difficulty. Arousal for the factor of Curiosity continuously increases from factor value 1 to factor value 2. While as Figure 4.11 shows arousal for Fantasy and difficulty presents the same trend as valence.

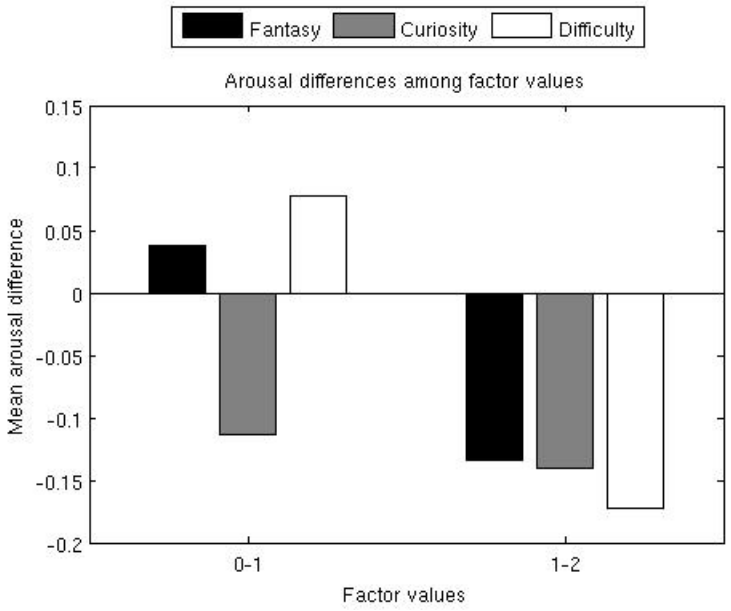
Table 4.11 represents the baseline for the two emotional categories. The baseline is simply calculated as the probability of the most popular class: $\frac{\mathbf{C}}{\mathbf{N}}$, where C = the number of utterances assigned in the most popular emotional class and N is the total number of the utterances.

The features selected as the best for the two audio data sets are shown in Table 4.12.

As we can see in Table 4.12, the wrapper method selected mostly pitch features as the most appropriate for the emotion classification problem, which is quite rational. As it is mentioned in [45], pitch characteristics are very good emotional indicators. Specifically, in arousal dimension, emotions that are stimulating, such as anger and happiness are characterized by increase in fundamental frequency (F0) and speech intensity. While on the other hand, less stimulating emotions,



(a)



(b)

Figure 4.11: (a) Mean valence differences among the factor values, (b) Mean arousal differences among the factor values

Selected Features	
Praat	Standard deviation of pitch Standard deviation of intensity
openSMILE	Maximum F0

Table 4.12: Selected Audio Features

	Praat Toolkit			
	At least two agree		Three agree	
Classifier	All features	Selected features	All features	Selected features
Naive Bayes	72%	**81.6%**	73.3%	**85.2%**
3NNR	**83.5%**	**75.7%**	**86.1%**	**83.2%**
	openSMILE Toolkit			
	At least two agree		Three agree	
Classifier	All features	Selected features	All features	Selected features
Naive Bayes	**77%**	**84.8%**	80.2%	**89.1%**
3NNR	**84.5%**	**81.1%**	**82.2%**	**93.1%**

Table 4.13: Classification accuracy for negative/positive emotion classification problem

such as boredom and sadness, are characterized by decreasing in fundamental frequency [59] and speech intensity.

Emotion classification results before and after feature selection are shown in Tables 4.13 and 4.14. With bold we present classification accuracies that exceed the baseline, as presented in Table 4.11:

At a first glance we can see that even before the feature selection we manage to outreach the baseline, with the 3NNR classifier. Classification accuracies indicate that emotion prediction is quite an easy task, for preschool ages. This yields that emotion can be demonstrated as a strong indicator for adaptation.

As results indicate in Tables 4.13 and 4.14, emotion recognition seems to perform quite well, even if the classifiers that are used are quite simple.

It is interesting that the performance of emotion recognition does not follow the same age-trend as that of speech recognition, which is known to degrade for younger children [19]. This is attributed to the fact that infants are quite expres-

	Praat Toolkit			
	At least two agree		Three agree	
Classifier	All features	Selected features	All features	Selected features
Naive Bayes	76%	**86%**	82.2%	**95.1%**
3NNR	**87.2%**	**81.9%**	**92.1%**	**93.1%**
	openSMILE Toolkit			
	At least two agree		Three agree	
Classifier	All features	Selected features	All features	Selected features
Naive Bayes	**79.5%**	**89.1%**	86.1%	**95.1%**
3NNR	**89.1%**	**86.7%**	**91.1%**	**93.1%**

Table 4.14: Classification accuracy for active/passive emotion classification problem

sive when talking. Since, younger children are more spontaneous it is easier to identify their emotional state. Meaning that compared to adults, young children have less control of their emotions.

To be more specific we managed to reach 93.1% and 95.1% accuracy, respectively, in valence and arousal classification. It is clear that the highest accuracies accrue from the three labelers agreement, which is quite rational as the data are more accurate. We believe that affect can be a strong indicator for adaptation, in speech dialogue systems for preschoolers, as it is easily predictable.

4.4.3 Factor Classification Results

For the optimal factor level prediction, we are going to represent the classification results using the audio feature sets (Praat/openSMILE), the objective feature set, as well as the merging of audio feature sets and the objective one.

As preliminary analysis indicated, the class which represents the zero factor value consists of few samples, since rarely such an application set up was selected as the most entertaining. As a result for each factor we have only two classes, one class which represents factor value one and the other represents factor value two. Since, users played different tasks per each session, in some sessions some children played less tasks than the particular number of tasks. Hence the classification results that are presented below, are calculated through leave one out

Baseline 2 Classes		Number of sessions
Fantasy	58.82%	34
Curiosity	68.42%	38
Challenge	64.71%	34

Table 4.15: Baseline for the Two Classes classification problem

cross validation, meaning that each time one session is used for testing while the rest are used for training.

Initially we calculated a baseline, by simply calculating for each of the three Malone factors the probability of the most popular class: $\frac{\mathbf{C}}{\mathbf{N}}$, where C = the number of entertaining sessions assigned in the most popular class for one specific factor and N = the total number of the sessions assigned as entertaining for one specific factor. The baseline results are shown in Table 4.15:

Table 4.16 presents the classification results before the feature selection. With highlight we represent the results that exceed the baseline. The columns of the table represent the classification accuracy when:

- only the objective metrics are used as features
- only the audio features extracted by the Praat toolkit are used
- the merging of objective metrics and Praat features is used
- only the audio features extracted by the openSMILE toolkit are used
- the merging of objective metrics and openSMILE features is used
- the merging of objective metrics and all the audio features (Praat and openSMILE) is used

The recognition of the optimal factor level is a difficult task, mainly because it is very user dependent. In Table 4.16 the classification accuracies of the three classifiers, before the feature selection, are shown. In general the merging of objective and audio features seems to display better performance, in comparison to the isolated feature sets. The challenge and fantasy factors seem to be easily

Neural Network							
Factor	Object.	Praat	Object. + Praat	openSMILE	Object. + openS-MILE	Audio (Praat+ openS-MILE)	Object. + Au-dio
Fantasy	47.1%	52.9%	55.9%	**67.7%**	**64.7%**	**61.8%**	58.8%
Curiosity	**84.2%**	65.8%	**73.7%**	68.4%	**76.3%**	60.5%	**81.6%**
Challenge	**76.5%**	**82.4%**	**85.3%**	**73.5%**	**85.3%**	**76.5%**	**82.4%**
Naive Bayes							
Factor	Object.	Praat	Object. + Praat	openSMILE	Object. + openS-MILE	Audio (Praat+ openS-MILE)	Object. + Au-dio
Fantasy	**67.7%**	**67.7%**	**70.6%**	**61.8%**	**61.8%**	**70.6%**	**70.6%**
Curiosity	**73.7%**	68.4%	**71.1%**	**73.7%**	68.4%	**73.7%**	68.4%
Challenge	**79.4%**	**82.4%**	**82.4%**	**67.7%**	**73.5%**	**70.6%**	**73.5%**
SVM							
Factor	Object.	Praat	Object. + Praat	openSMILE	Object. + openS-MILE	Audio (Praat+ openS-MILE)	Object. + Au-dio
Fantasy	52.9%	**67.7%**	**64.7%**	**73.5%**	**67.7%**	**73.5%**	**67.7%**
Curiosity	68.4%	57.9%	**86.8%**	**71.1%**	**81.6%**	**79.0%**	**81.6%**
Challenge	**73.5%**	**73.5%**	**85.3%**	64.7%	**88.2%**	**70.6%**	**85.3%**

Table 4.16: Factor Classification results before Feature Selection

	Selected Features		
Factor	Praat Feature Set	openSMILE Feature Set	Object. Feature Set
Fantasy	Max. Pitch Alter. Baseline Pitch Mean Intensity	rms min. Energy Mean F0	Min. Resp. Time
Curiosity	Min. Pitch Max. Pitch Alter. Baseline Pitch Range of Pitch F0 Points Mean Intensity Std of Intensity Energy	min. Voice Prob.	Age # of Correct Answers Voice usage Avg. Inact. Time Avg Inter. Time Avg. Resp. Time Min. Resp. Time
Challenge	Max. Pitch Mean Pitch Median Pitch Std of Pitch Wav Duration F0 Points Max. Intensity	Max. rms Energy Std rms Energy F0 Mean	Age Gender Avg. Resp. Time Min. Resp. Time

Table 4.17: Selected audio features for the Praat and openSMILE data sets

separable, while classification results for curiosity factor indicate that optimal curiosity level is not easily recognizable.

Since, fantasy is highly correlated with entertainment, we expect that high fantasy levels make the game entertaining. This is proved by the classification results, since in most cases fantasy recognition outreaches the baseline. Although optimal level of challenge seems to be more user dependent than the other two factors, classifiers manage to recognize the preferred level of challenge in all cases. Results indicate that fantasy and challenge seem to be good indicators for adaptation.

As we mentioned previously, we conducted feature selection. The selected features are presented in Table 4.17:

We confront each Malone factor, as a different classification problem, independently from the other factors. As a result we have different metrics selected for each factor, as it is shown in Table 4.17.

As we can see the wrapper method selected mostly pitch characteristics as the most appropriate for the optimal factor level prediction. This is attributed to the fact that pitch is related with the excitement [45]. Also age is a good indicator of the level of challenge, since older children can deal with difficult tasks more efficiently than younger ones.

Table 4.18 presents the classification results after the feature selection. For the integrated Audio (Praat+openSMILE) we used as best features the merging of the best Praat and best openSMILE features:

Classification results indicate that in most cases we exceed the baseline significantly (over 10%). The results are quite improved, compared to those before the feature selection, but the accuracies indicate that the data are still noisy. This is attributed to the fact that we have few users, and the amount of data are not sufficient to accurately estimate the model parameters.

In general, the prediction of the preferred factor levels, as well as the emotion, would lead to better design in games. Games able to recognize the user's emotional state and interest level in real time, would boost the interaction efficiency. Since, human-to-human interaction is a dynamic process, we expect that by taking advantage of that, and designing systems adaptable to users' needs, will bring the human-computer interaction to a new level.

We managed to show that there are objective and audio criteria that affect the user's preferences. Also, the emotion and entertainment seem somewhat to be connected. This means that emotion can be an indicator towards adaptation.

Neural Network							
Factor	Object.	Praat	Object. + Praat	openSMILE	Object. + openS-MILE	Audio (Praat+ openS-MILE)	Object. + Audio
Fantasy	**61.8%**	**64.7%**	**61.8%**	**73.5%**	**64.7%**	**67.7%**	55.9%
Curiosity	**84.2%**	**76.3%**	**76.3%**	**71.1%**	**84.2%**	**71.1%**	**76.3%**
Challenge	**70.6%**	**85.3%**	**85.3%**	**79.4%**	**97.1%**	**79.4%**	**91.2%**
Naive Bayes							
Factor	Object.	Praat	Object. + Praat	openSMILE	Object. + openS-MILE	Audio (Praat+ openS-MILE)	Object. + Audio
Fantasy	**67.7%**	**82.4%**	**85.3%**	**70.6%**	**70.6%**	**82.4%**	**79.4%**
Curiosity	**73.7%**	**71.1%**	65.8%	**73.7%**	**71.1%**	**71.1%**	68.4%
Challenge	**79.4%**	**88.2%**	**91.2%**	**79.4%**	**82.4%**	**88.2%**	**88.2%**
SVM							
Factor	Object.	Praat	Object. + Praat	openSMILE	Object. + openS-MILE	Audio (Praat+ openS-MILE)	Object. + Audio
Fantasy	**61.8%**	**73.5%**	**70.6%**	**79.4%**	**79.4%**	**79.4%**	**79.4%**
Curiosity	**71.1%**	65.8%	68.4%	68.4%	**71.1%**	68.4%	**71.1%**
Challenge	**76.5%**	**76.5%**	**91.2%**	**73.5%**	**91.2%**	**76.5%**	**91.2%**

Table 4.18: Factor Classification results after Feature Selection

Chapter 5

User Modeling and Affective Evaluation with Physiological Signals

During face-to-face communication, human beings declare intention by expressing their emotions or mood. It is understood that affect and emotion plays an important role, as it enhances communication among participants. Recently the study of the human communication components, presents great interest in the field of Human Computer Interaction (HCI). To be more specific, the field of affective computing aims at incorporating affective and emotional cues in HCI.

In Chapter 4 we concentrated on the system's evaluation and the "definition" of the three Malone factors, as well as emotion by using objective metrics (e.g. inactive, active and response times, modality usage e.t.c.) and audio features extracted from children's speech. On the contrary, in this chapter we study the attention, meditation and arousal patterns as well as various brainwave signals. This will give us a better insight on the child computer interaction in comparison with adults. We concentrate on the Electroencephalography (EEG) which is a rich information source of affective and cognitive states during interaction. Using physiological signals is challenging but also effective as they provide information that cannot be easily faked by the user.

5.1 User Modeling

User modeling is a subdivision of human computer interaction and describes the process of building up and modifying a user model. The main goal of user modeling is the customization and adaptation of systems to the user's specific needs. The system needs to "say the 'right' thing at the 'right' time in the 'right' way" [31]. Another common purpose is modeling specific kinds of users, including modeling of their skills and declarative knowledge, for use in automatic software-tests [11].

5.1.1 User Model

A user model represents a collection of personal data associated with a specific user. Therefore, it is the basis for any adaptive changes to the system's behavior. Which data is included in the model depends on the purpose of the application. It can include personal information such as users' names and ages, their interests, their skills and knowledge, their goals and plans, their preferences and their dislikes or data about their behavior and their interactions with the system.

There are different design patterns for user models, though often a mixture of them is used [11], [38].

- **Static User Models:** Static user models are the most basic kinds of user models. Once the main data is gathered they are normally not changed again, they are static. Shifts in users' preferences are not registered and no learning algorithms are used to alter the model.

- **Dynamic User Models:** Dynamic user models allow a more up to date representation of users. Changes in their interests, their learning progress or interactions with the system are noticed and influence the user models. The models can thus be updated and take the current needs and goals of the users into account.

- **Stereotype Based User Models:** Stereotype based user models are based on demographic statistics. Based on the gathered information users

are classified into common stereotypes. The system then adapts to this stereotype.

- **Highly Adaptive User Models:** Highly adaptive user models try to represent one particular user and therefore allow a very high adaptivity of the system. In contrast to stereotype based user models they do not rely on demographic statistics but aim to find a specific solution for each user.

5.2 Affective Computing

Affective computing is the study and development of systems and devices that can recognize, interpret, process, and simulate human affects. It is an interdisciplinary field spanning computer sciences, psychology, and cognitive science [40]. There are two areas of affective computing: *Detecting and recognizing emotional information* and *Emotion in machines*.

Detecting emotional information begins with passive sensors which capture data about the user's physical state or behavior without interpreting the input. The data gathered is analogous to the cues humans use to perceive emotions in others. For example, a video camera might capture facial expressions, body posture and gestures, while a microphone might capture speech. Other sensors detect emotional cues by directly measuring physiological data, such as skin temperature and galvanic resistance [54]. Recognizing emotional information requires the extraction of meaningful patterns from the gathered data. This is done using machine learning techniques that process different modalities speech recognition, natural language processing, or facial expression detection, and produce either labels (i.e. 'confused') or coordinates in a valence-arousal space.

Another area within affective computing is the design of computational devices proposed to exhibit either innate emotional capabilities or that are capable of convincingly simulating emotions. A more practical approach, based on current technological capabilities, is the simulation of emotions in conversational agents in order to enrich and facilitate interactivity between human and machine [21]. While human emotions are often associated with surges in hormones and other neuropeptides, emotions in machines might be associated with abstract states

associated with progress (or lack of progress) in autonomous learning systems. In this view, affective emotional states correspond to time-derivatives (perturbations) in the learning curve of an arbitrary learning system.

5.3 The Human Brain

The human brain is the center of the human nervous system. It has the same general structure as the brains of other mammals, but is larger than expected on the basis of body size among other primates. Estimates for the number of neurons (nerve cells) in the human brain range from 80 to 120 billion. Most of the expansion comes from the cerebral cortex, especially the frontal lobes, which are associated with executive functions such as self-control, planning, reasoning, and abstract thought. The portion of the cerebral cortex devoted to vision is also greatly enlarged in human beings, and several cortical areas play specific roles in language, a skill that is unique to humans.

The cerebral hemispheres form the largest part of the human brain and are situated above most other brain structures. The cerebral cortex is nearly symmetrical, with left and right hemispheres that are approximate mirror images of each other. Anatomists conventionally divide each hemisphere into four "lobes", the frontal lobe, parietal lobe, occipital lobe, and temporal lobe (Figure 5.1). This division into lobes does not actually arise from the structure of the cortex itself, though: the lobes are named after the bones of the skull that overlie them, the frontal bone, parietal bone, temporal bone, and occipital bone.

Neuroscientists, along with researchers from allied disciplines, study how the human brain works. Such research has expanded considerably in recent decades. Information about the structure and function of the human brain comes from a variety of experimental methods. By placing electrodes on the scalp it is possible to record the summed electrical activity of the cortex, using a methodology known as electroencephalography (EEG). EEG measures mass changes in synaptic activity from the cerebral cortex and can detect changes in electrical activity over large areas of the brain. In addition to measuring the electric field directly via electrodes placed over the skull, it is possible to measure the magnetic field that the brain generates using a method known as magnetoencephalography (MEG).

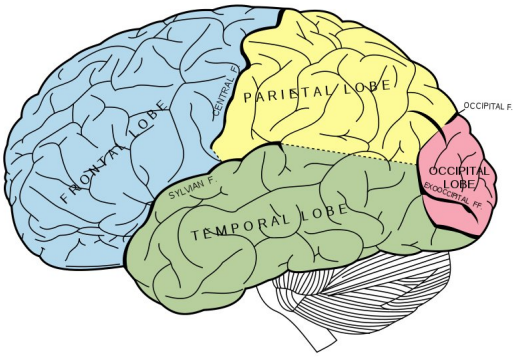


Figure 5.1: The Human Brain

This technique also has good temporal resolution like EEG but with much better spatial resolution. There are several methods for detecting brain activity changes by three-dimensional imaging of local changes in blood flow. The older methods are SPECT and PET, which depend on injection of radioactive tracers into the bloodstream. The newest method, functional magnetic resonance imaging (fMRI), has considerably better spatial resolution and involves no radioactivity [62].

5.4 The NeuroSky Device

The usual process to obtain brainwave measurements is a medical procedure supervised and performed by trained personnel, involving several electrodes positioned around the head and attached with conductive gel. This may cause discomfort for the subject, and consequently will lead to erroneous results.

The NeuroSky MindSet [3] is a BCI, which is a simplified version of the traditional EEG technology. The MindSet monitors electrical potential between the sensing electrode, positioned on the forehead, and the reference electrodes, positioned on the left earlobe (Figure 5.2).

By using an EEG recording device, such as the MindSet, with a single dry electrode at the forehead, the measuring and analysis of brain states becomes less complex and more comfortable for the subject. The user can use the device without the help of trained personnel and wherever she wants. The single point

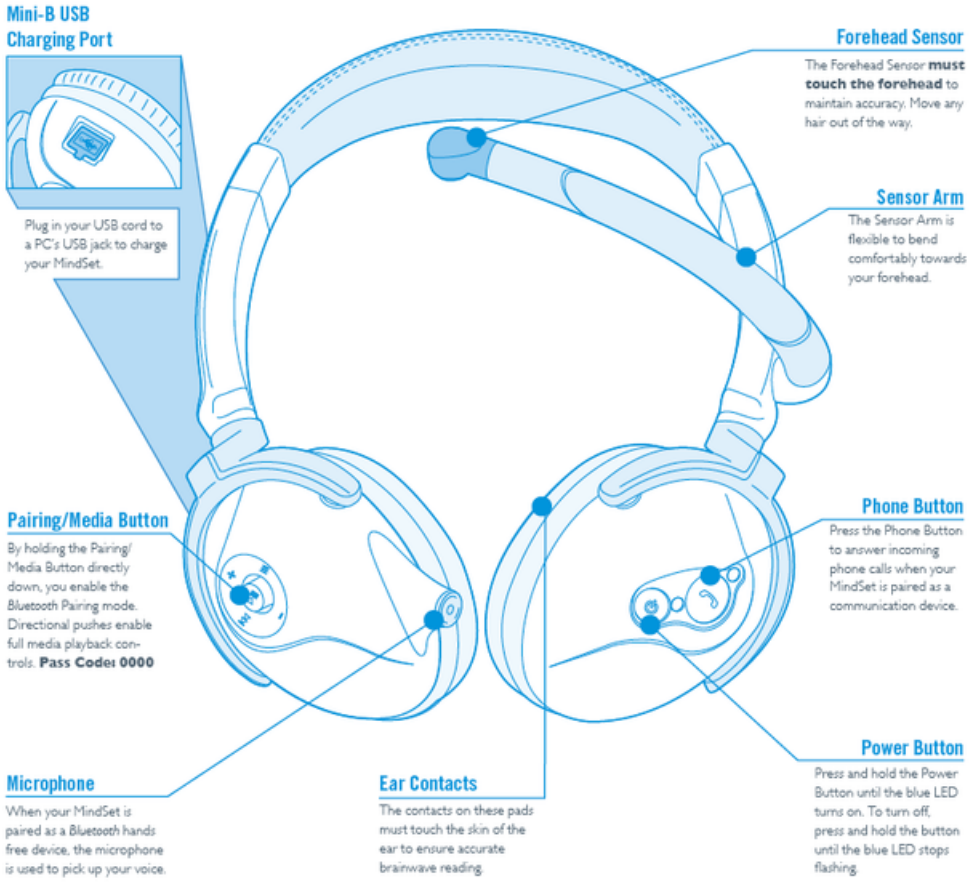


Figure 5.2: NeuroSky MindSet Diagram [3]

electrode means that changes in brainwave activity in different parts of the brain cannot be monitored. However, volume conduction makes it possible to measure electrical potentials at some distance from their source generators. Therefore, the single point electrode is able to monitor a substantial part of the entire brain's activity. The sensing electrode is positioned on the forehead. There is no hair between the electrode and the scalp, which makes for a stronger, steadier signal. Also, the cognitive signals linked to higher states of consciousness originate from the frontal cortex, which lies directly below the forehead.[1]

Inside every NeuroSky product, including the MindSet, is the ThinkGear chip. This chip enables the device to interface with the wearer's brainwaves by amplifying the raw brainwave signal and removing the ambient noise and artifacts. The noise is filtered out of the raw EEG before the calculation of the eSense values, described below. This noise elimination is performed by a proprietary algorithm and no information pertaining to this internal process has been released. When the ThinkGear chip detects too much noise to be filtered out in a satisfactory way, the same eSense meter values are repeated. Therefore, all meter values that are consecutive and equal need to be marked as noise and removed before data analysis begins. ThinkGear communicates with the MindSet through the ThinkGear Socket Protocol.[2]

5.4.1 MindSet Data Types

5.4.1.1 eSense Meters

By using NeuroSky's proprietary eSense algorithm, the two eSense meters are calculated, with an interpretation update rate of 1 Hz [2]. They each describe a mental state and draw their names from them:

[1] *Details about NeuroSky's MindSet can be found at Appendix A* and for full details about the data types, including meanings and ranges, please download the free Mindset Development Tools (MDT) and refer to the MindSet Communications Protocol (download from http://store.neurosky.com/products/developer-tools-2-1).

[2] *Details about the ThinkGear Connector can be found at http://developer.neurosky.com/docs/doku.php?id=thinkgear_connector_tgc*

1. **Attention (or Engagement hereafter)** indicates the intensity of a user's level of mental "focus" or "attention", such as which occurs during intense concentration and directed (but stable) mental activity. Distractions, wandering thoughts, lack of focus, or anxiety may lower the engagement meter levels.

2. **Meditation** indicates the level of a user's mental "calmness" or "relaxation". Note that meditation is a measure of a person's mental levels, not physical levels, so simply relaxing all the muscles of the body may not immediately result in a heightened Meditation level. However, for most people in most normal circumstances, relaxing the body often helps the mind to relax as well.

Meditation is related to reduced activity by the active mental processes in the brain, and it has long been an observed effect that closing one's eyes turns off the mental activities which process images from the eyes, so closing the eyes is often an effective method for increasing the meditation meter level. Distractions, wandering thoughts, anxiety, agitation, and sensory stimuli may lower the Meditation meter levels.

The eSense values range between 1 and 100. On this scale, a value between 40 to 60 at any given moment in time is considered "neutral". A value between 60 to 80 is considered "slightly elevated" and a value between 80 to 100 is considered "elevated", meaning they are strongly indicative of heightened levels of that eSense. Similarly, on the other end of the scale, a value between 20 to 40 indicates "reduced" levels of the eSense, while a value between 1 to 20 indicates "strongly lowered" levels of the eSense. These levels may indicate states of distraction, agitation, or abnormality, according to the opposite of each eSense.

5.4.1.2 Brainwaves Band Powers

Besides, eSense values NeuroSky also measures brainwave (eeg) signal values, such as alpha, beta, gamma (split in low and high frequencies), as well as delta and theta with a sampling rate of 512 Hz. The different characteristics of brainwave types, in terms of frequency ranges (bands) are shown in Table 5.1 and Figure 5.3

Brainwave Type	Frequency Range
Delta (δ)	0.1-3Hz
Theta (θ)	4-7Hz
Alpha (α)	8-9Hz (Low) 10-12Hz (High)
Beta (β)	12-17Hz (Low) 18-30Hz (High)
Gamma (γ)	30-40Hz (Low) 41-100Hz (High)

Table 5.1: Different characteristics of brainwave types [6]

To be more specific:

- **Delta (δ) Wave** A delta wave is a high amplitude brain wave with a frequency of oscillation between 0.1-3 hertz. Delta waves, like other brain waves, are recorded with an electroencephalogram (EEG) and are usually associated with the deepest stages of sleep (3 and 4 NREM), also known as slow-wave sleep (SWS), and aid in characterizing the depth of sleep.

- **Theta (θ) Wave** A theta wave refers to frequency components in the 4-7 Hz range, regardless of their source. Theta waves are usually associated with working memory and cognitive load. Working memory can be constructed as an outcome of the ability to control attention and sustain its focus on a particular active mental representation. In other words, this notion is nearly synonymous with what we commonly understand as the ability to effortfully concentrate on task performance.

- **Alpha (α) Wave** Alpha waves are neural oscillations in the frequency range of 8-12 Hz arising from synchronous and coherent (in phase/constructive) electrical activity of thalamic pacemaker cells in humans. They are also called Berger's wave in memory of the founder of EEG. The α-band tends to be attenuated in high load tasks. α-waves are typical for an alert, but relaxed mental state.

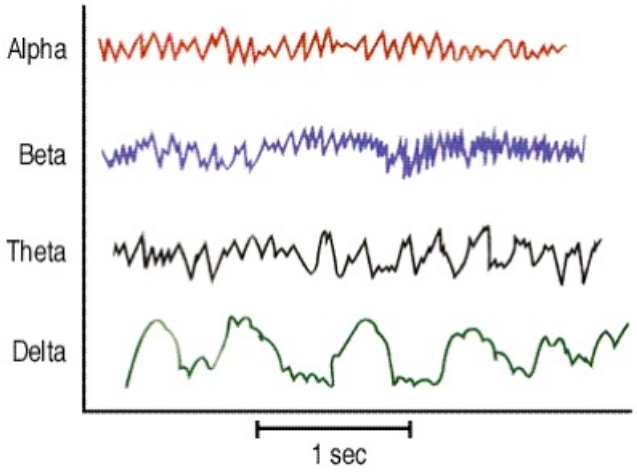


Figure 5.3: Brainwaves Graph

- **Beta (β) Wave** Beta wave, or beta rhythm, is the term used to designate the frequency range of human brain activity between 12 and 30 Hz. The β-band is expected to be increased in power during high load tasks. β-waves is related to an active state of mind, most prominent in the frontal cortex and over other areas during intense focused mental activity.

- **Gamma (γ) Wave** A gamma wave is a pattern of neural oscillation in humans with a frequency between 30 to 100 Hz, though 40 Hz is prototypical. According to a popular theory, gamma waves may be implicated in creating the unity of conscious perception (the binding problem).

EEG typically involves analyzing amplitudes (powers) of activity in certain frequency ranges (bands). ThinkGear reports the relative power of each EEG band, typically at 1 second intervals. The ASIC-EEG-POWER-INT values are indications of relative amplitudes of the individual EEG bands. Typically, power spectrum band powers would be reported in units such as Volts-squared per Hz ($\frac{V^2}{Hz}$), but since our values have undergone a number of complicated transforms and rescale operations from the original voltage measurements, there is no longer a simple linear correlation to units of Volts. Hence, we do not try to label them with any conventional unit. You can think of them as ASIC-EEG-POWER units, if you must.

The reason we say they are only meaningful compared to each other and to themselves is primarily due to the fact they have their own units as described above. It would not necessarily be meaningful nor correct to directly compare them to, say, values output by another EEG system. In their currently output form, they are useful as an indication of whether each particular band is increasing or decreasing over time, and how strong each band is relative to the other bands [1].

5.4.1.3 Poor Signal Quality

This value gives an indication of how poor the signal is, such as whether the signal is contaminated with noise, or is clearly not even connected to a person's head [5]. It is an integer value that is generally in the range of 0 to 200, with 0 indicating a good signal and 200 indicating an off-head state.

5.4.2 Arousal

Arousal is a physiological and psychological state of being awake or reactive to stimuli, and often is referred to be associated with β and α waves [22]. It is known that β brainwaves indicate an alerted state of mind, while α brainwaves indicate a more relaxed state. So the arousal is calculated by the $\frac{\beta}{\alpha}$ ratio. So $\frac{\beta}{\alpha} > 1$, indicates an active state, while $\frac{\beta}{\alpha} < 1$, indicates a passive state. A neutral state is indicated by $\frac{\beta}{\alpha} = 1$.

5.5 Experimental Methodology

5.5.1 Subjects

Fourteen adults and six preschool children participated in this survey. Adults of both genders participated (5 females and 9 males) and the age ranged from 20 to 59 years. On the other hand only female, 6 years old children participated.

5.5.2 Experimental Procedure

The experimental procedure took place in a sound attenuated office room, where the subjects sat on a comfortable chair. During the whole process subjects should wear the NeuroSky MindSet, so that we could collect EEG and eSense data during interaction. Before the experiment begins there was a test "period", where the subject should find the most comfortable position in front of the laptop's screen and we tried to best fit the NeuroSky device on the subject's head. The procedure was a slightly different for the two user groups.

5.5.2.1 Children

During the experiment, each child was asked to play one session of the game application. During that session we modified only the challenge value from 1 to 2, while the other two factors (fantasy and curiosity) remained constant. Also, after finishing the session, we asked the child play twice more the game (all the three factors where at level 1), once by using only her voice and once by using only the mouse input device.

5.5.2.2 Adults

Each adult was asked to play only one application setup, that where all the three factors are taken the value 1. They also played the game by using both input modalities.

5.6 Results

In each game-task, the user had to answer five simple questions in order to successfully complete it. We concentrate on engagement, meditation and arousal values, in order to observe certain behavioral patterns.

For a better representation, results are displayed through interaction turns. An interaction turn is split into four interaction types, as follows:

- While user waits for the system's prompt: **WP**
- The system's prompt (this is one value): **P**

- While user is thinking the answer (user is inactive): **WA**
- When the user speaks out the answer (user is active): **A**

When the user waits to exit the game is called as **WE**. The WE interaction type is unique for each game and so it can be omitted since the user just waits to exit the game without doing anything. Each game consists of five WP turns, five P turns, as well as five WA and A turns. Average interaction turn patterns are calculated as:

$$AvgTurn = \frac{\sum_{i=1}^{N} turn_i}{N}$$

, where $N = \#Turns$ and $turn_i \equiv [WP, P, WA, A]$

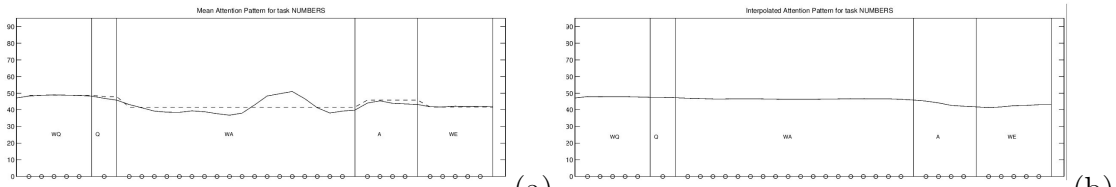
5.6.1 Adults

To our knowledge the five tasks are very simple for adults, so we do not expect to see extreme variations in users' behavior.

5.6.1.1 Engagement, Meditation and Arousal per task

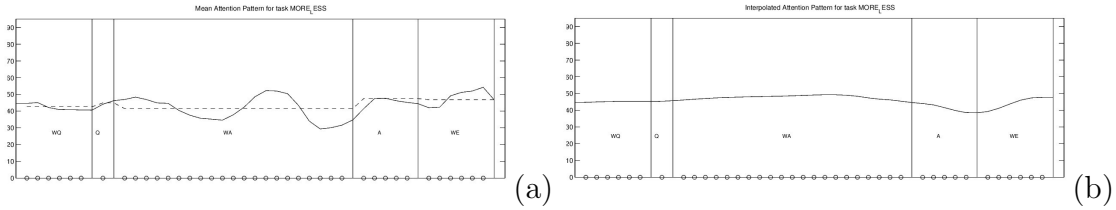
Tables 5.2, 5.3, 5.4 display the mean and the standard deviation of engagement, meditation and arousal signals, per task averaged on all users. While Figures 5.4, 5.5 and 5.6 show the mean turn and the interpolated engagement, meditation and arousal signals per task.

We chose to draw both the average and interpolated signals because of the time variability users display during interaction. Each user displays different inactivity (WA) and activity times (A), meaning that some users have larger inactivity times or give more detailed answers (they are more active) compared to others. So as the average signal is time dependent, the interpolated signal does not depend on time. The interpolation is used to estimate missing data, within the range of a discrete set of known data points. The interpolation was computed by the spline function.



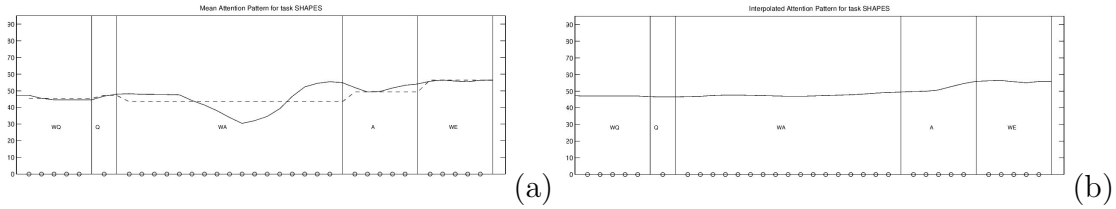
(a) (b)

(a) Avg Engagement Pattern for Numbers (b) Interpolated Engagement Pattern for Numbers



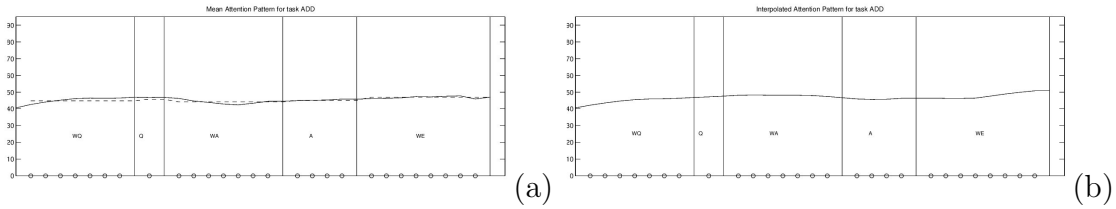
(a) (b)

(a) Avg Engagement Pattern for More/Less (b) Interpolated Engagement Pattern for More/Less



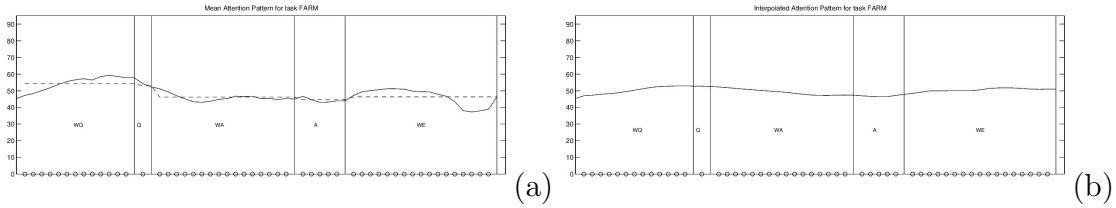
(a) (b)

(a) Avg Engagement Pattern for Shapes (b) Interpolated Engagement Pattern for Shapes



(a) (b)

(a) Avg Engagement Pattern for Additions (b) Interpolated Engagement Pattern for Additions



(a) (b)

(a) Avg Engagement Pattern for Farm (b) Interpolated Engagement Pattern for Farm

Figure 5.4: Average and interpolated engagement per task averaged on all users

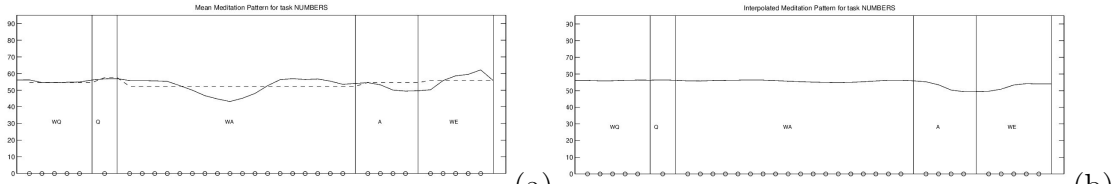
		Engagement					
		Interaction Types					
		WP	P	WA	A	WE	Avg. turn
NUMBERS	MEAN	48.74	47.94	41.62	45.85	41.69	43.72
	STD	1.15	19.49	9.42	1.89	1.56	8.09
MORE/LESS	MEAN	42.58	44.91	41.50	47.49	46.83	42.71
	STD	6.21	18.91	11.61	7.34	12.29	10.08
SHAPES	MEAN	45.45	47.25	43.41	49.26	56.39	44.96
	STD	3.85	18.46	10.13	2.42	2.44	8.31
ADD	MEAN	44.79	45.74	44.14	45.10	47.01	44.67
	STD	2.39	18.00	3.26	1.86	2.69	2.50
FARM	MEAN	54.11	53.03	46.15	44.63	46.33	49.08
	STD	6.19	13.35	3.87	6.01	7.97	6.44
	Mean of Means	47.13	47.77	43.36	46.47	47.65	45.03

Table 5.2: Mean and std engagement values for each task, per interaction turn

We observe a decrease in engagement during the WA fashion, a pattern that is reversed in the case of meditation. During WA the whole processing is occurred, so we would expect an increase of engagement instead of a decrease. In fact, in trivial tasks there is no need for sustained focus, so the brain works on a focused but relaxed mode.

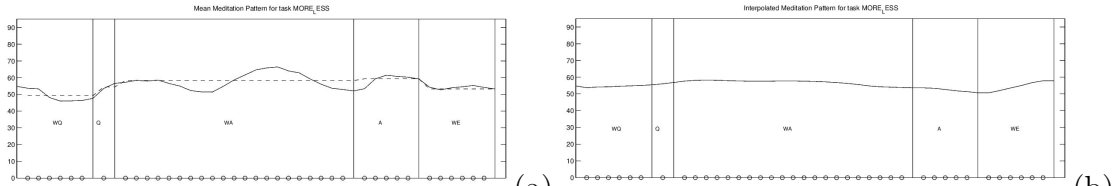
Even if there are some differences among interaction turn types, they are insignificant, as one-way ANOVA results in a $p \approx 0.15$. This is also verified by Figure 5.4, where in the first column we present the average engagement pattern and in the second column we present the interpolated engagement over the interaction turns. In the average engagement pattern with dashed line we draw each interaction type's mean engagement value. We observe that the engagement displays a flat like pattern and there are no significant differences among tasks (one-way ANOVA gives $p \approx 0.18$).

Farm displays the largest engagement and meditation in WP and P. In farm there is only audio output; audio and visual stimuli activate different neural processes, and it is known that the presence or absence of a modality affects the



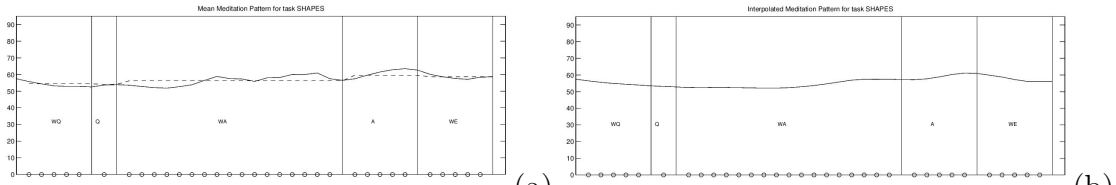
(a) (b)

(a) Avg Meditation Pattern for NUMBERS (b) Interpolated Meditation Pattern for NUMBERS



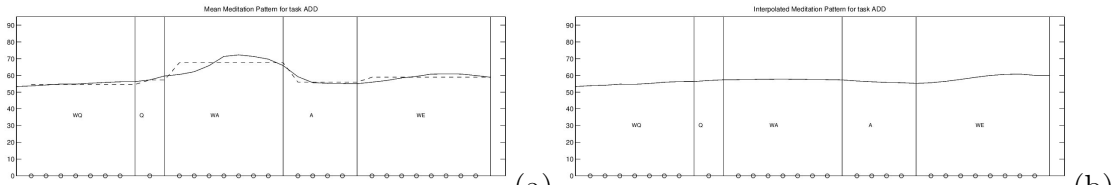
(a) (b)

(a) Avg Meditation Pattern for MORE/LESS (b) Interpolated Meditation Pattern for MORE/LESS



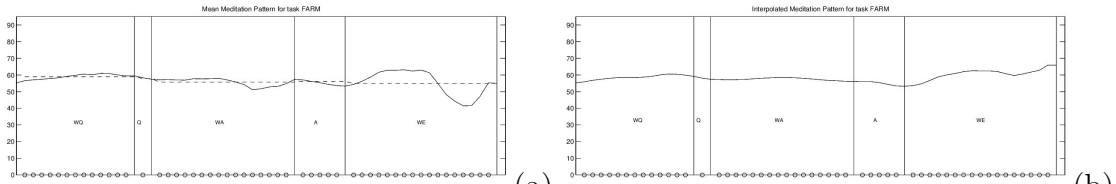
(a) (b)

(a) Avg Meditation Pattern for SHAPES (b) Interpolated Meditation Pattern for SHAPES



(a) (b)

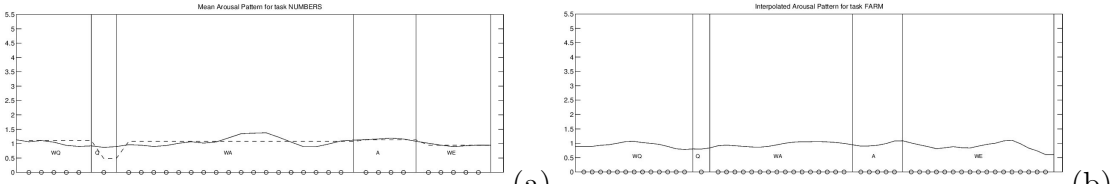
(a) Avg Meditation Pattern for ADDITIONS (b) Interpolated Meditation Pattern for ADDITIONS



(a) (b)

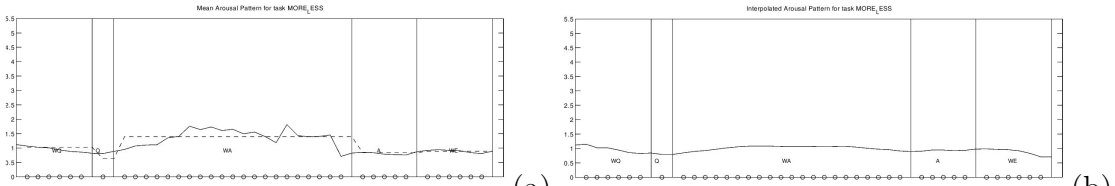
(a) Avg Meditation Pattern for FARM (b) Interpolated Meditation Pattern for FARM

Figure 5.5: Average and interpolated meditation per task averaged on all users



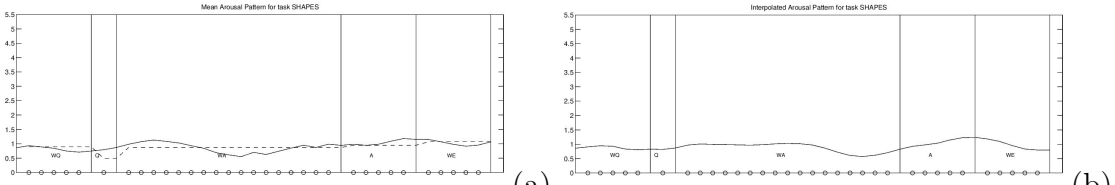
(a) (b)

(a) Avg Arousal Pattern for NUMBERS (b) Interpolated Arousal Pattern for NUMBERS



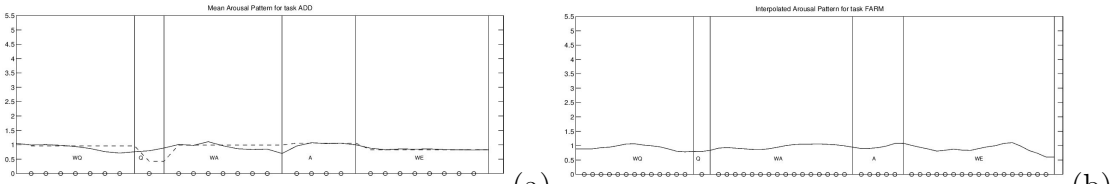
(a) (b)

(a) Avg Arousal Pattern for MORE/LESS (b) Interpolated Arousal Pattern for MORE/LESS



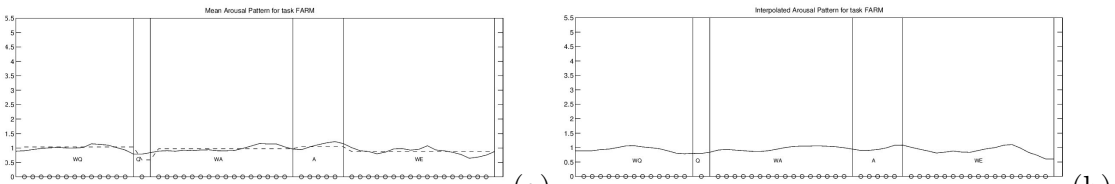
(a) (b)

(a) Avg Arousal Pattern for SHAPES (b) Interpolated Arousal Pattern for SHAPES



(a) (b)

(a) Avg Arousal Pattern for ADDITIONS (b) Interpolated Arousal Pattern for ADDITIONS



(a) (b)

(a) Avg Arousal Pattern for FARM (b) Interpolated Arousal Pattern for FARM

Figure 5.6: Average and interpolated arousal per task averaged on all users

		Meditation					
		Interaction Types					
		WP	P	WA	A	WE	Avg. turn
NUMBERS	MEAN	54.59	57.50	52.31	54.63	55.75	53.23
	STD	2.49	18.13	7.08	1.72	12.19	5.89
MORE/LESS	MEAN	49.23	54.41	58.25	59.29	53.26	56.65
	STD	10.03	17.24	6.65	12.17	3.81	8.67
SHAPES	MEAN	54.38	54.12	56.20	59.60	58.63	56.41
	STD	3.19	14.89	6.51	4.34	3.88	5.68
ADD	MEAN	54.55	57.26	67.56	55.99	58.86	59.79
	STD	1.13	14.85	10.43	1.66	3.00	8.65
FARM	MEAN	58.90	57.82	55.65	56.14	54.91	56.99
	STD	2.50	16.15	4.54	1.54	12.05	3.74
	Mean of Means	54.33	56.22	57.99	57.13	56.28	56.61

Table 5.3: Mean and std meditation values for each game, per interaction turn

stimuli processing. When a modality is absent, the human brain tries to generate information of that modality and the user must be more focused.

Also, Figure 5.4 indicates a temporal difference in response time ($WA + A$ time) among tasks. Given that both the auditory and visual information from a single event will inform the observer about that event, it is not surprising to observe such variations in response time. To be more specific in numbers, shapes, more/less and farm tasks the response time is larger that in additions task. Research has shown that observers are faster when responding to redundant bimodal stimuli (e.g. visual and audio) than they are to either of the component unimodal stimuli alone [79]. Indeed, additions task is the only one that provides audiovisual information during the prompt.

In 5.6 we can observe that the highest values are displayed mostly in the WQ, WA and A interaction turns. Indeed, when someone knows that is going to be exposed to a stimulus, they feel more aroused in order to be "prepared". After the stimulus appears (in our case the question is posed) the arousal is slightly dropped while they are thinking the answer, and during the answer the arousal

		Arousal					
		Interaction Turns					
		WQ	Q	WA	A	WE	Avg. turn
NUMBERS	MEAN	1.11	0.48	1.08	1.13	0.95	1.07
	STD	0.08	0.40	0.42	0.18	0.18	0.36
MORE/LESS	MEAN	1.03	0.63	1.40	0.84	0.89	1.22
	STD	0.12	0.67	1.00	0.16	0.20	0.83
SHAPES	MEAN	0.89	0.49	0.87	0.95	1.07	0.87
	STD	0.08	0.42	0.32	0.13	0.37	0.26
ADD	MEAN	0.96	0.43	0.99	1.06	0.83	0.97
	STD	0.10	0.60	0.50	0.34	0.07	0.35
FARM	MEAN	1.03	0.59	0.98	1.04	0.88	1.00
	STD	0.20	0.59	0.17	0.29	0.29	0.21
	Mean of Means	1.00	0.52	1.06	1.00	0.92	1.03

Table 5.4: Mean and std arousal values for each game, per interaction turn

is slightly raised again as they get ready for the next for the next question.

Nevertheless, a "flat" image appears also in meditation and arousal as Figures 5.5, 5.6 indicate. Also in meditation and arousal there are no significant differences among tasks ($p \approx 0.57$ in meditation and $p \approx 0.90$ in arousal).

5.6.1.2 Correlations among eSense and arousal

Here we are going to present the most important correlations among eSense and arousal signals over interaction turn types. Since interaction is split into turns ($\approx$ 5 turns per interaction) and each turn is split into four types-parts (WP,P,WA,A), each signal consists of $(\#OfUsers)*(\#OfTasksTheUsePlayed)*(\#OfTurnsPerTask)*(\#OfInteractionTypes) \approx 1400data$. We want to examine how engagement, meditation and arousal depend on a certain interaction type e.g. what happens to engagement when the user is active, or in which way engagement is related to meditation etc. The most important findings are presented in Table 5.5.

As it is expected, arousal is positively correlated with engagement and negatively correlated with meditation, since engagement is referred to an intense

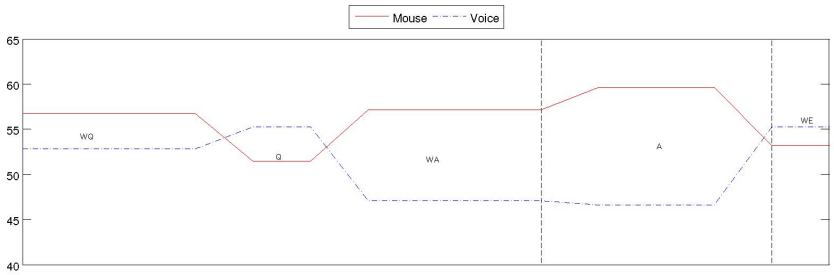
	cor. coef.	p-value
Arousal / Engagement	0.16	0
Arousal / Meditation	-0.15	0
Engagement / Meditation	0.23	0

Table 5.5: Correlations between eSense and arousal

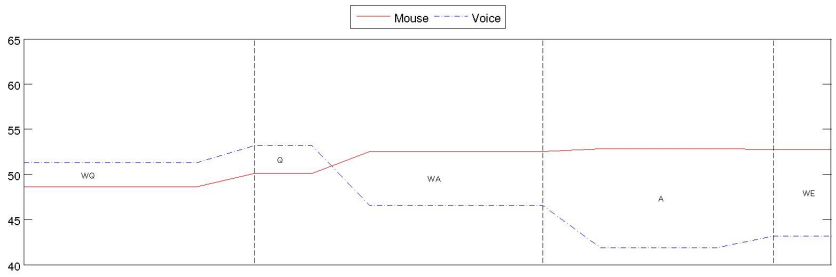
concentration state. The unexpected here, is the positive correlation between the engagement and the meditation. Someone would expect that engagement would be negatively correlated with meditation. But meditation is a measure of a person's mental levels, not physical levels and often it is mentioned to be a "relaxed focus" mental state. Considering meditation as a relaxed but focused state, it is justified that relationship between engagement and meditation.

5.6.1.3 Engagement, Meditation and Arousal patterns through modality usage

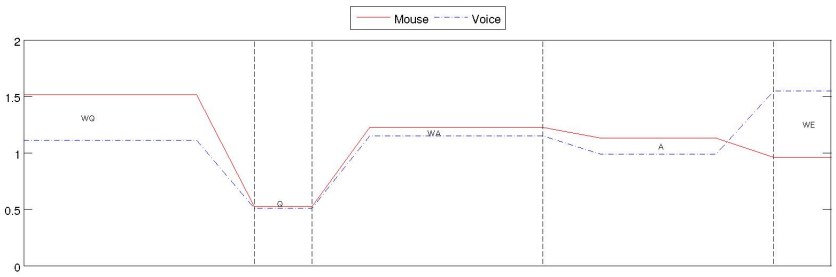
In this subsection we are going to discuss how engagement, meditation and arousal are affected by the input modality. As we mentioned in subsection 1 there is a bias through voice input, since four out to five games support only voice as input modality. In order to examine how eSense values and arousal are affected by input modalities, we asked users play again the farm game, once using only their voice and then using only the mouse input device. The findings are shown in Figure 5.7.



Mouse and voice engagement patterns by mean values



Mouse and voice meditation patterns by mean values



Mouse and voice arousal patterns by mean values

Figure 5.7: Modality eSense mean and std values per interaction turn averaged on all users

We can observe that in WA and A interaction turn types engagement level increases when mouse is used, compared with engagement in the case of voice input. One-way ANOVA in the case of engagement suggest that there are significant differences between mouse and voice ($p \approx 0.07$). Although adults are familiar with the mouse device, firstly they have to detect the target and then click on it, while in the case of voice input they only have to speak out the name of the target. The target detection in the case of mouse, requires extra information processing, which elevates the attention level. Meditation pattern is similar to attention, but there are no significant differences among modalities ($p \approx 0.34$). Arousal in both input modalities displays insignificant differences ($p \approx 0.55$).

5.6.2 Children

5.6.2.1 Engagement, Meditation and Arousal per task

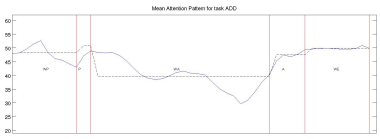
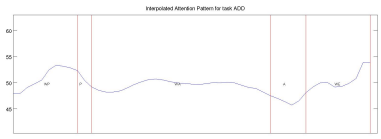
Tables 5.6, 5.7, 5.8 display the mean and standard deviation engagement, meditation and arousal values per task, averaged on all users. The μ and σ are each task's overall mean and standard deviation respectively.

Even if there are some differences among interaction turn types, they are insignificant (one-way ANOVA gives $p \approx 0.44$ for difficulty 1 and $p \approx 0.8$ for difficulty 2). This is also verified by Figure 5.9, where in the first column we present the average engagement pattern and in the second column we present the interpolated engagement over the interaction turns. In the average engagement pattern with dashed line we draw each interaction type's mean engagement value. In Figure 5.9 we observe that the engagement displays a flat like pattern and there are no significant differences among tasks (one-way ANOVA gives $p \approx 0.46$ for difficulty 1 and $p \approx 0.37$ for difficulty 2). This is attributed to the fact that, tasks are easy enough for six year old children, consequently there is no need for sustained focus and users display near neutral engagement levels. Numbers task displays the largest engagement in WA turn. Figure 5.9 indicates also temporal difference in response time ($WA + A$ time) among tasks. Farm displays the less response time, compared to other tasks. Perhaps because as a task is very easy.

The same image appears also in meditation and arousal as Figures 5.10, 5.12 indicate. Also in meditation and arousal there are no significant differences among

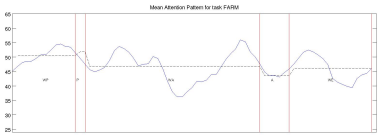
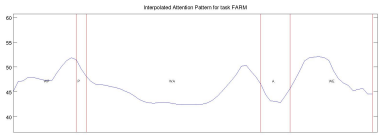
			Engagement					
			Interaction Parts					
			WP	P	WA	A	WE	Avg.Turn
NUMBERS	Diff=1	MEAN	43.67	43.87	47.06	46.28	42.50	46.20
		STD	1.25	21.03	16.61	1.97	3.14	13.13
	Diff=2	MEAN	48.32	50.77	80.84	38.20	48.30	65.70
		STD	2.04	26.71	21.81	18.24	11.93	25.68
MORE/LESS	Diff=1	MEAN	41.65	41.77	45.27	47.60	52.19	45.14
		STD	2.44	12.69	11.13	4.79	1.57	10.45
	Diff=2	MEAN	41.63	42.20	38.66	55.47	35.93	42.15
		STD	2.25	20.68	6.71	16.19	6.96	10.61
SHAPES	Diff=1	MEAN	48.44	51.30	48.63	47.20	45.07	48.55
		STD	3.81	20.96	5.84	1.49	1.72	5.35
	Diff=2	MEAN	45.88	39.43	47.55	41.10	38.77	46.42
		STD	2.57	19.21	8.68	10.16	2.02	8.46
ADD	Diff=1	MEAN	48.23	50.80	39.71	47.45	49.79	42.69
		STD	9.20	19.06	6.98	1.75	3.19	8.06
	Diff=2	MEAN	42.21	45.30	35.34	41.78	45.04	36.65
		STD	1.49	19.16	8.11	3.19	5.69	7.85
FARM	Diff=1	MEAN	50.57	51.87	46.72	43.53	45.98	47.39
		STD	3.85	14.13	7.78	1.14	6.35	6.88
	Diff=2	MEAN	49.26	46.77	46.03	40.68	45.47	46.22
		STD	3.51	18.89	9.57	5.99	4.84	7.81
	Diff=1	Mean of Means	46.51	47.92	45.48	46.41	47.11	46.69
	Diff=2	Mean of Means	45.46	44.89	49.68	43.45	42.70	45.24

Table 5.6: Mean and std engagement values for each task, per interaction turn

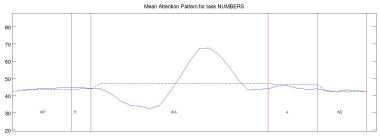
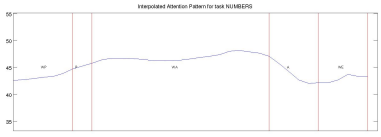
(a) (b)

(a) Avg Engagement Pattern for Additions (b) Interpolated Engagement Pattern for Additions

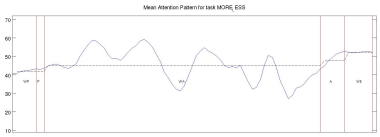
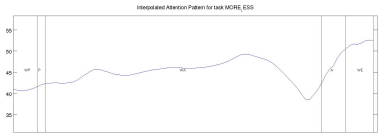
(a) (b)

(a) Avg Engagement Pattern for Farm (b) Interpolated Engagement Pattern for Farm

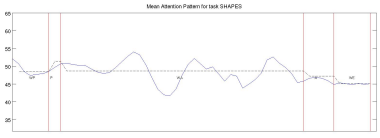
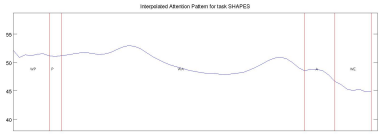
(a) (b)

(a) Avg Engagement Pattern for Numbers (b) Interpolated Engagement Pattern for Numbers

(a) (b)

(a) Avg Engagement Pattern for More/Less (b) Interpolated Engagement Pattern for More/Less

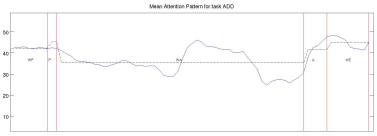



(a) (b)

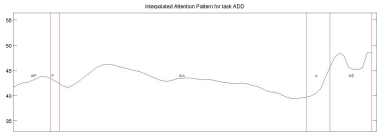
(a) Avg Engagement Pattern for Shapes (b) Interpolated Engagement Pattern for Shapes

Figure 5.8: Average and interpolated engagement per task averaged on all users for difficulty level 1

tasks ($p \approx 0.60$ for difficulty 1 and $p \approx 0.46$ for difficulty 2 in meditation and $p \approx 0.97$ for difficulty 1 and $p \approx 0.53$ in arousal).

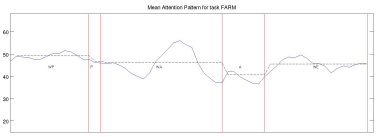


(a)

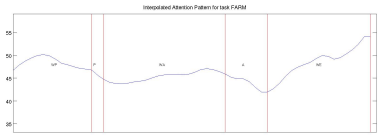


(b)

(a) Avg Engagement Pattern for Additions (b) Interpolated Engagement Pattern for Additions

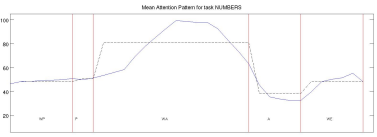


(a)

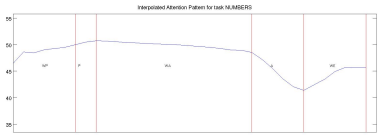


(b)

(a) Avg Engagement Pattern for Farm (b) Interpolated Engagement Pattern for Farm

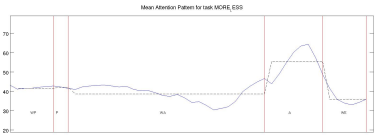


(a)

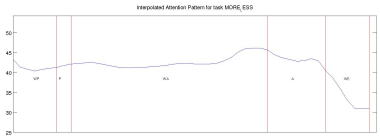


(b)

(a) Avg Engagement Pattern for Numbers (b) Interpolated Engagement Pattern for Numbers

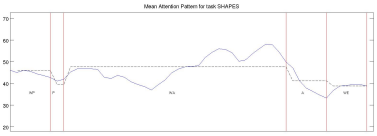


(a)

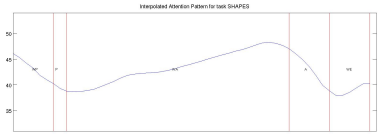


(b)

(a) Avg Engagement Pattern for More/Less (b) Interpolated Engagement Pattern for More/Less



(a)



(b)

(a) Avg Engagement Pattern for Shapes (b) Interpolated Engagement Pattern for Shapes

Figure 5.9: Average and interpolated engagement per task averaged on all users for difficulty level 2

			Meditation					
			Interaction Types					
			WP	P	WA	A	WE	Avg.Turn
NUMBERS	Diff=1	MEAN	50.96	52.87	47.43	58.54	56.83	49.93
		STD	2.43	20.83	14.47	4.68	3.49	12.18
	Diff=2	MEAN	39.89	46.83	49.80	40.80	45.90	46.11
		STD	1.69	18.81	7.16	9.37	2.18	7.91
MORE/LESS	Diff=1	MEAN	50.55	47.87	53.80	59.45	61.61	53.88
		STD	2.22	18.68	11.22	12.07	7.62	10.91
	Diff=2	MEAN	45.75	44.77	46.36	56.39	46.23	48.04
		STD	2.01	13.96	6.69	14.95	5.35	9.00
SHAPES	Diff=1	MEAN	52.29	52.23	52.88	56.23	50.07	53.08
		STD	7.69	19.38	9.15	5.45	7.70	8.60
	Diff=2	MEAN	47.41	45.57	53.27	48.76	47.03	51.89
		STD	0.85	19.92	9.32	5.26	9.33	8.53
ADD	Diff=1	MEAN	52.16	60.93	49.73	54.94	55.73	51.12
		STD	6.31	15.79	11.31	2.55	3.76	9.79
	Diff=2	MEAN	45.81	47.60	46.92	45.52	46.15	46.73
		STD	3.07	15.06	8.10	0.95	1.62	7.37
FARM	Diff=1	MEAN	61.81	57.77	52.54	55.46	52.08	55.01
		STD	12.64	15.71	5.99	3.44	4.21	8.55
	Diff=2	MEAN	50.63	50.67	49.29	43.76	54.55	48.88
		STD	6.22	14.31	2.42	5.45	5.86	4.88
	Diff=1	Mean of Means	53.55	54.33	51.28	56.92	55.26	54.27
	Diff=2	Mean of Means	45.90	47.09	49.13	47.05	47.97	47.73

Table 5.7: Mean and std meditation values for each game, per interaction turn

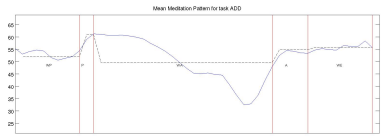
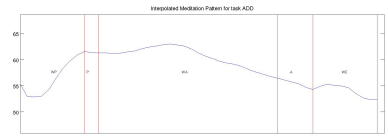
5.6.3 Correlations among eSense and arousal

As it is expected (Table 5.9), arousal is positively correlated with engagement since engagement is referred to an intense concentration state. The unexpected also here, is the positive correlation between engagement and meditation. But as we mentioned previously meditation is "relaxed focus" mental state. Finally, when wrong answers increase, the engagement also increase. This is attributed to the fact that when someone realizes a mistake, they concentrate more to correct it.

As tasks are quite easy for six year old children, we would expect no significant

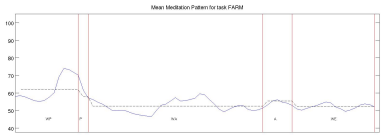
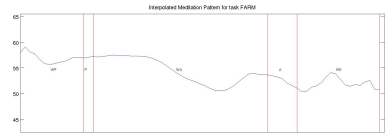
			Arousal					
			Interaction Turns					
			WQ	Q	WA	A	WE	Avg.Turn
NUMBERS	Diff=1	MEAN	0.79	0.37	0.85	0.67	0.84	0.80
		STD	0.08	0.40	0.56	0.15	0.59	0.46
	Diff=2	MEAN	1.00	0.34	0.59	0.77	0.77	0.69
		STD	0.39	0.32	0.19	0.18	0.33	0.29
MORE/LESS	Diff=1	MEAN	0.80	0.44	0.79	0.68	0.80	0.78
		STD	0.10	0.39	0.43	0.16	0.39	0.41
	Diff=2	MEAN	0.76	0.56	0.95	0.89	0.68	0.91
		STD	0.05	0.47	0.31	0.41	0.14	0.31
SHAPES	Diff=1	MEAN	0.95	0.42	0.86	0.79	1.08	0.85
		STD	0.17	0.37	0.36	0.09	0.43	0.33
	Diff=2	MEAN	0.86	0.37	0.71	0.83	0.68	0.73
		STD	0.09	0.25	0.45	0.31	0.11	0.40
ADD	Diff=1	MEAN	0.86	0.39	0.93	0.79	0.81	0.89
		STD	0.19	0.29	0.29	0.15	0.27	0.27
	Diff=2	MEAN	0.87	0.61	1.08	0.99	0.96	1.04
		STD	0.12	0.54	0.55	0.19	0.33	0.51
FARM	Diff=1	MEAN	0.78	0.48	0.96	0.72	1.01	0.89
		STD	0.17	0.33	0.45	0.18	0.59	0.39
	Diff=2	MEAN	0.80	0.43	0.67	0.73	0.72	0.72
		STD	0.23	0.37	0.24	0.13	0.29	0.23
	Diff=1	Mean of Means	0.84	0.42	0.88	0.73	0.91	0.76
	Diff=2	Mean of Means	0.86	0.46	0.80	0.84	0.76	0.74

Table 5.8: Mean and std arousal values for each game, per interaction turn

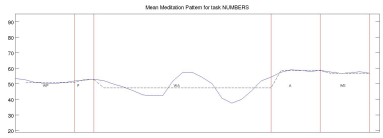
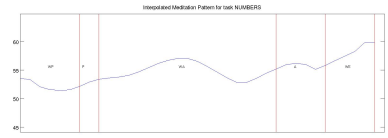
(a) (b)

(a) Avg Meditation Pattern for Additions (b) Interpolated Meditation Pattern for Additions

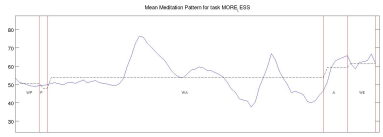
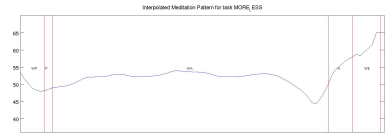
(a) (b)

(a) Avg Meditation Pattern for Farm (b) Interpolated Meditation Pattern for Farm

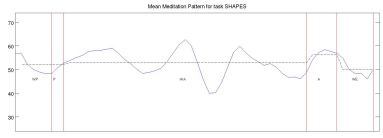
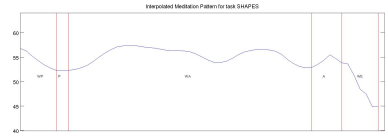
(a) (b)

(a) Avg Meditation Pattern for Numbers (b) Interpolated Meditation Pattern for Numbers

(a) (b)

(a) Avg Meditation Pattern for More/Less (b) Interpolated Meditation Pattern for More/Less

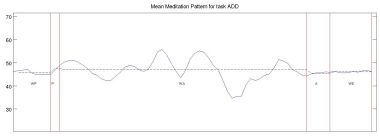



(a) (b)

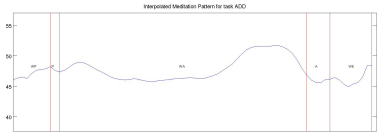
(a) Avg Meditation Pattern for Shapes (b) Interpolated Meditation Pattern for Shapes

Figure 5.10: Average and interpolated Meditation per task averaged on all users for difficulty level 1

differences between the two difficulty levels, since difficulty level 2 is not much challenging. Indeed, between difficulty level 1 and 2 there are no significant differences. One-way ANOVA indicated $0.08 < p < 0.8$.

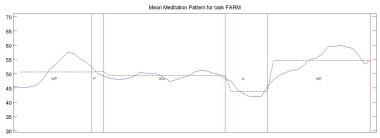


(a)

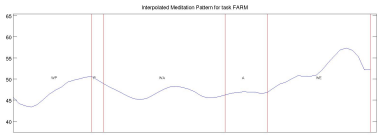


(b)

(a) Avg Meditation Pattern for Additions (b) Interpolated Meditation Pattern for Additions

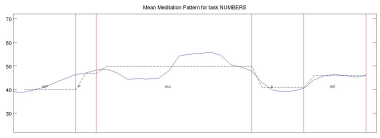


(a)

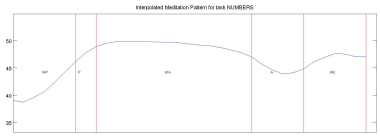


(b)

(a) Avg Meditation Pattern for Farm (b) Interpolated Meditation Pattern for Farm

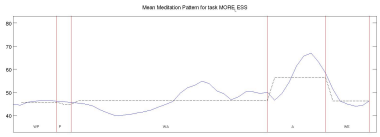


(a)

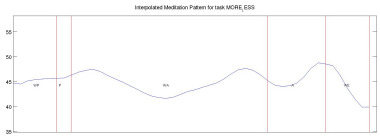


(b)

(a) Avg Meditation Pattern for Numbers (b) Interpolated Meditation Pattern for Numbers

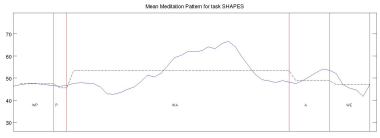


(a)

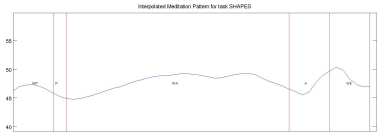


(b)

(a) Avg Meditation Pattern for More/Less (b) Interpolated Meditation Pattern for More/Less



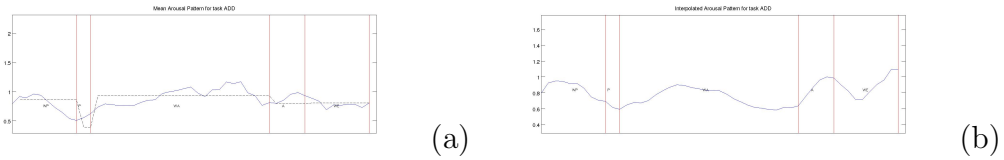
(a)



(b)

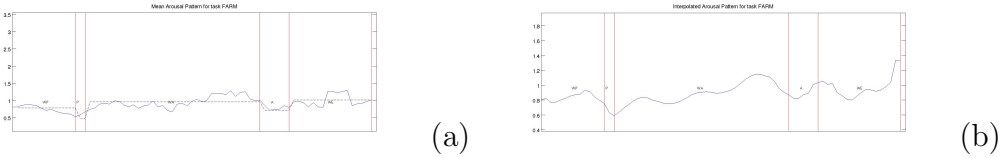
(a) Avg Meditation Pattern for Shapes (b) Interpolated Meditation Pattern for Shapes

Figure 5.11: Average and interpolated meditation per task averaged on all users for difficulty level 2



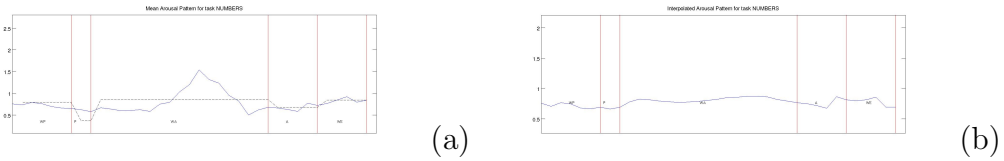
(a) (b)

(a) Avg Arousal Pattern for Additions (b) Interpolated Arousal Pattern for Additions



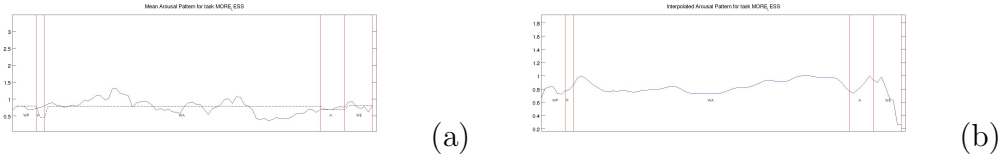
(a) (b)

(a) Avg Arousal Pattern for Farm (b) Interpolated Arousal Pattern for Farm



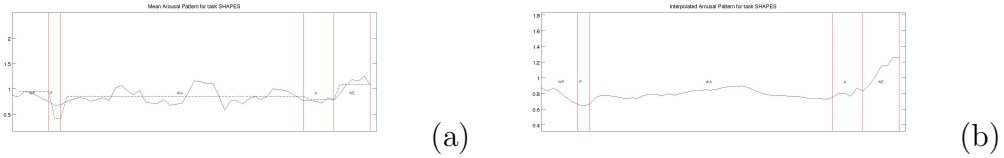
(a) (b)

(a) Avg Arousal Pattern for Numbers (b) Interpolated Arousal Pattern for Numbers



(a) (b)

(a) Avg Arousal Pattern for More/Less (b) Interpolated Arousal Pattern for More/Less



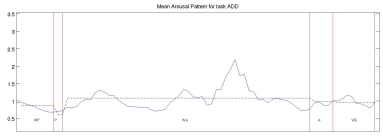
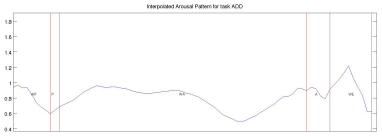
(a) (b)

(a) Avg Arousal Pattern for Shapes (b) Interpolated Arousal Pattern for Shapes

Figure 5.12: Average and interpolated Arousal per task averaged on all users for difficulty level 1

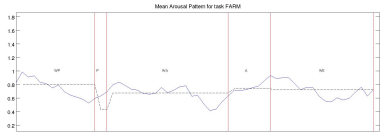
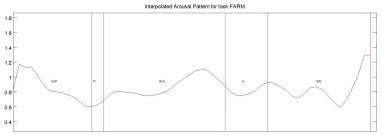
5.6.4 Engagement, Meditation and Arousal patterns through modality usage

In this subsection we are going to discuss how engagement, meditation and arousal are affected by the input modality. As we mentioned in subsection 1 there is a bias through voice input, since four out to five games support only voice as input

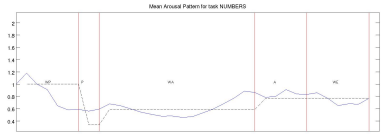
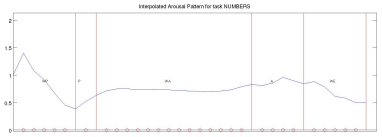
(a) (b)

(a) Avg Arousal Pattern for Additions (b) Interpolated Arousal Pattern for Additions

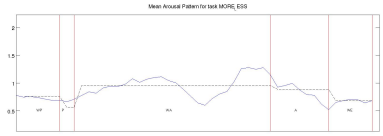
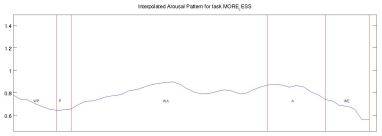
(a) (b)

(a) Avg Arousal Pattern for Farm (b) Interpolated Arousal Pattern for Farm

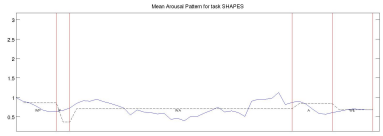
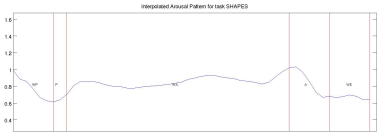
(a) (b)

(a) Avg Arousal Pattern for Numbers (b) Interpolated Arousal Pattern for Numbers

(a) (b)

(a) Avg Arousal Pattern for More/Less (b) Interpolated Arousal Pattern for More/Less

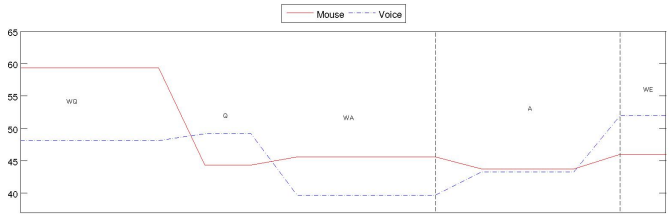
(a) (b)

(a) Avg Arousal Pattern for Shapes (b) Interpolated Arousal Pattern for Shapes

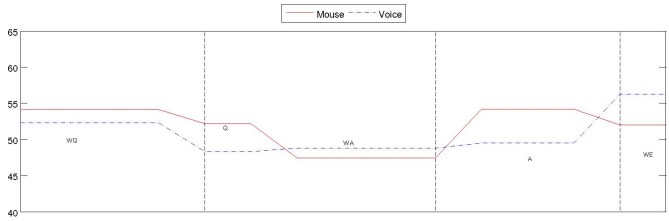
Figure 5.13: Average and interpolated Arousal per task averaged on all users for difficulty level 2

modality. In order to examine how eSense values and arousal are affected by input modalities, we asked users play again the farm game, once using only their voice and then using only the mouse input device. The findings are shown in Figure 5.14.

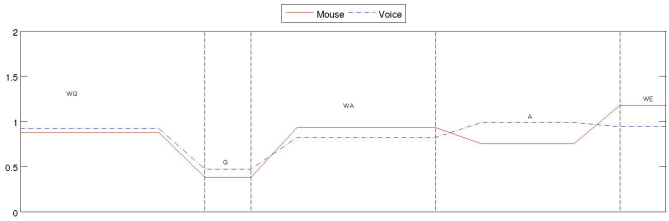
We can observe that in WA and A interaction turn types engagement level increases when mouse is used, compared with engagement in the case of voice



Mouse and voice engagement patterns by mean values



Mouse and voice meditation patterns by mean values



Mouse and voice arousal patterns by mean values

Figure 5.14: Modality eSense mean and std values per interaction turn

	Interaction turn split in smaller parts	
	cor. coef.	p-value
Activity time / Arousal	-0.27	0.001
Correct-Wrong Ans. / Engagement (WA)	0.15	0.06
Arousal / Engagement	0.28	0.001
Engagement / Meditation	0.26	0
	Interaction turns	
	cor. coef.	p-value
Arousal / Engagement	0.28	0.001
Engagement / Meditation	0.25	0.002

Table 5.9: Correlations between eSense and arousal

input. One-way ANOVA in the case of engagement suggest that there are no significant differences between mouse and voice ($p \approx 0.49$). Although children are familiar with the mouse device, firstly they have to detect the target and then click on it, while in the case of voice input they only have to speak out the name of the target. The target detection in the case of mouse, requires extra information processing, which elevates the engagement level. Meditation pattern is similar to engagement, but there are no significant differences among modalities ($p \approx 0.27$). Arousal in both input modalities displays insignificant differences ($p \approx 0.71$).

5.6.5 Compare children and adults

In this section we are going to present results, in order to compare adults' and children's behavior during game-play.

Children are slightly engaged compared to adult, while adults seem to be more meditated compared to children. Adults when they have to complete a task which is trivial they work on a relaxed mode. On the other hand in young ages, even for easy tasks, more processing is needed and hence more attention for fulfilment.

Although there are some differences, those are insignificant and both children and adults display quit similar patterns for the three affective metrics. That indicates that the completion of a task, which has a low difficulty level, follows

	Engagement					
	WP	P	WA	A	WE	Mean Turn
Children	46.51	47.92	45.48	46.41	47.11	46.69
Adults	47.13	47.77	43.36	46.47	47.65	45.03
	Meditation					
	WP	P	WA	A	WE	Mean Turn
Children	53.55	54.33	51.28	56.92	55.26	54.27
Adults	54.33	56.22	57.99	57.13	56.28	56.28
	Arousal					
	WP	P	WA	A	WE	Mean Turn
Children	0.84	0.42	0.88	0.73	0.91	0.76
Adults	1.00	0.52	1.06	1.00	0.92	1.03

Table 5.10: Mean eSense turn split in interactions

a certain process which is independent of the age. Also as far as arousal is concerned, both adults and children are passive and more specifically near the neutral point. Games are not challenging enough for both adults and children so this justifies the neutral arousal values.

As far as the modality is concerned, also adults and children follow the same patterns. Both adults and children display low arousal levels (near neutral point) for both modalities. The fact that children are not as much familiar with the mouse device as adults, can justify higher levels of arousal children display when use the mouse. But the small differences between children and adults are justified by the fact that children nowadays are familiar with computers, and the use of peripherals such as the mouse.

Both adults and children are in general more engaged during the WA turn when they use the mouse, since the location of the target needs more processing. Than pattern is reversed for the meditation.

Engagement						
Modality	User	WP	P	WA	A	WE
Mouse	Adults	56.74	51.49	57.17	59.62	53.19
	Children	59.33	44.36	45.60	43.72	45.96
Voice	Adults	52.82	55.27	47.10	46.60	55.24
	Children	48.14	49.20	39.66	43.29	51.99
Meditation						
Modality	User	WP	P	WA	A	WE
Mouse	Adults	48.63	50.11	52.51	52.84	57.78
	Children	45.15	52.20	47.41	54.16	52.02
Voice	Adults	51.27	53.18	46.57	41.86	43.19
	Children	52.31	48.30	48.82	49.53	56.25
Arousal						
Modality	User	WP	P	WA	A	WE
Mouse	Adults	1.51	0.53	1.23	1.13	0.96
	Children	0.88	0.38	0.93	0.75	1.18
Voice	Adults	1.11	0.51	1.15	0.99	1.55
	Children	0.92	0.47	0.82	0.99	0.94

Chapter 6

Discussion, Conclusions and Future Work

This work constitutes an extension in previous research [72, 73], the main contribution is to determine how preschoolers experience the entertainment during the game play and in what extend emotion affects the interaction patterns. Various objective and audio features were tested by using simple classification models, in order to predict the desirable factor levels and the users' emotional state during gaming. Also the calculation of correlations between various metrics and the affect, we expect to give us a better insight on the role emotions play during interaction. We believe that correlations give us important information about how two quantities vary relatively to one another. As a result we can understand in what extend and how one quantity affects or is affected by another one. Since technology has began used for educational purposes [8], e.g. educational computer games, it is crucial to build computer applications able to "understand" users' needs and keep the child's interest in high levels. On that basis we believe that this work, will bring forward the first principles in order to understand how interaction patterns are formed and which are the elements that affect such a formation in preschool ages.

6.1 Importance of results and implications

In this work we have investigated how preschool children interact with multimodal dialogue systems. For this purpose, an on-line multimodal platform has been designed, implemented and used as a fresh start point to develop web-based speech-enabled applications for children. Five preschool games were implemented based on popular preschool activities and evaluated by fifteen children of ages 4-6.

Various levels of fantasy, curiosity and challenge are implemented in the games, in order to investigate how Malone's factors affect engagement and enjoyability. The evaluation experiments showed that fantasy and curiosity are positively correlated with children's entertainment, while the level of challenge seems to depend on each child's individual preferences.

Preliminary experiments also showed that interaction patterns and acoustic features are indicators of (subjectively) optimal levels of fantasy, curiosity and challenge. Moreover, we showed that emotion classification is possible in such ages. The classification results were better compared to those in adults and older children cases. This is due to the fact that younger children are more spontaneous and more expressive in their interaction with the computer. Nevertheless more experiments with more subjects and different system setups are needed in order to better understand how to design adaptive multimodal dialogue systems for preschool children that maximize engagement and enjoyability.

Also, physiological metrics such as EEGs helped us collect more accurate data during game play, as an attempt to create better and more reliable features. The data analysis indicated that adults as well as children display quite similar patterns in modality usage. Adults tend to be more meditative in comparison with children during tasks.

6.2 Future work

Further improvement of the platform could be done by supporting more modalities in the future. For example, the touch modality or gesture recognition might proved to be very interesting for the kids. During the experiments we notice that some kids, especially the younger ones, had the tendency to touch the screen in

order to give their answer to the game. That shows that touch could be a very effective modality in multimodal dialogue systems for preschoolers.

Moreover, there are research directions to be taken in terms of the adaptation of the three Malone factors. Since challenge seems to be the most easily recognizable factor, we can implement the challenge adaptation, so that user would stay engaged during game play. By improving the emotion classifier, we can detect when the child is negatively excited and drop the challenge level. On the other hand when boredom is detected, the challenge level should be raised.

Finally, more subjects should participate in an experiment in order to collect more EEG data. Perhaps another device with more electrodes would give us the chance to examine in depth the patterns displayed during interaction.

Appendix A

NeuroSky MindSet Characteristics

A.1 Main Benefit

The MindSet is NeuroSky's flagship headset. Known for its ease-of-use by over 200 universities worldwide, the EEG signal is measured through a comfortable dry electrode. Data collected by this system is output in the form of raw-brainwaves, EEG power spectrums (alpha waves, beta waves, e.t.c) , attention (eSense measured), meditation (eSense measured), and the quality of the data being collected.

A.2 Overview

The MindSet was the first mass market EEG headset when it debuted in 2009. The device consists of earphone, a microphone, and a sensor arm. The headset's reference and ground electrodes are on the ear phones and the EEG electrode is on the end of the sensor arm, sitting on the forehead above the eye. Not only can it collect research-grade EEG data, but it features a set of high-quality Bluetooth enabled wireless headphones complete with microphone. The MindSet is versatile, compact, and portable. It is an excellent choice for gaming, research, or public installations.

A.3 Product Contents

1. MindSet headset with A2DP audio earphone and microphone
2. G-Link Bluetooth transmitter
3. AC plug (3.0x1500mm AC100-240 50/60HZ DC5V 400MA)
4. MindSet Quick Start Guide (also downloadable at: http://developer.neurosky.com/docs/doku.php?id=mindset)
5. MindSet Instruction Manual (also downloadable at http://developer.neurosky.com/docs/doku.php?id=mindset)
6. MindSet Applications Disk: includes Visualizer 2 and NeuroBoy
7. TGAM module featuring a TGAT ASIC
8. Lithium-ion rechargeable battery (non removable), connects to a USB cable port
9. Battery run-time: 7 hours
10. Static Headset ID (headsets have a unique MAC address for pairing purposes)

A.4 Measures

- Raw-Brainwaves
- EEG power spectrums (Alpha, Beta, e.t.c)
- eSence meter for Attention and Meditation
- Quality of data being collected (can be used to detect poor contact and whether the device is off the head)

A.5 Physical

- 190 gr weight
- Length 21 cm x width 17.5 cm x height 5 cm

A.6 Power/Battery

- USB Jack Rating: DC5V 1A
- USB Jack Insertion Force: 2 Kgf max.
- USB Jack Life: 5000 cycles (INTERNAL ONLY. NOT FOR CONSUMER)
- Rate Power: 30 mV
- Max Power: 50 mV
- Battery Run Time: 7 hours
- Battery Capacity: 550 mAh
- Battery Charging Time: 4 hours

A.7 Signal and EEG

- Maximum signal input range: 1 mV pk-pk
- Hardware filter range: 3 Hz to 100 Hz
- MS001: includes 60 Hz environmental AC noise filtering
- MS002: includes 50 Hz environmental AC noise filtering
- Amplification gain: 2000x
- ADC resolution: 12 bits (-2048 to 2047)
- Sampling rate: 512 Hz

- eSense interpretation update rate: 1 Hz
- 0% byte loss (i.e packet loss)
- 1 Hz eSense interpretation rate
- UART Baudrate: 57600 Baud
- SPP through put: 9600 Baud
- S/N ration: ¿70 dB
- Class 2 Bluetooth Radio Range: 10 m 6 dBm RF max power
- 250 kbit/s RF data rate

A.8 Microphone and Headphones

- Microphone Dimension: 6x207 mm
- Microphone Sensitivity: -50 ± 3 dB
- Microphone Direction: Omni-directional
- Microphone Impedance: F = 1 kHz Max 2.2 KW
- Speaker type: plastic shell with internal magnet
- Speaker Impedance: 32 W $\pm$ 15%
- Speaker Frequency Response: 20 Hz - 20 KHz
- Speaker Sensitivity (S.P.L.) at 1 KHz: 93 ± 3 dB
- Audio output power 18 W/24 ohms
- Volume control
- Rate Power: 30 MW/Speaker
- Max Power: 50 MW/Speaker

- Sensitivity: -58 ± 3dB
- Speaker Dimension: 40 mm

A.9 Bluetooth Dongle

- Bluetooth Standard: V2.0+EDR (backward compatible with V1.1/V1.2)
- Profiles support: Serial Port, Dial Up Networking, Fax, Headset, Printer, PIM Item Transfer, PIM Synchronization, File Transfer, LAN Access, Generic Object Exchange, Object Push, HCRP, HID, A2DP, AVRCP
- Frequency Range: 2.402-2.480 GHz / 79 Channel FHSS
- Transfer rates (Max): 3 Mbps(Max)
- Power Class: Class 2 (10M distance)
- Receiving Sensitivity: -80dBm at 0.1% BER
- Antenna: Integrated Antenna
- Current Consumption: Standby mode: 25mA / Transmission mode: 80mA
- Interface: USB A type
- LED indicator: Power on
- OS Support: Windows 2000 / XP / Vista / 7
- Dimension: 18.6 x 14.2 x 4.5 mm (L x W x H)
- Weight: Approx. 2.4 grams

A.10 Compatible/Recommended Bluetooth Receivers

- Windows: Mavin Bluetooth Dongle
- Mac OSX: use the built-in Bluetooth receivers

[4]

Appendix A

Additional Results

A.1 Engagement, Meditation and Arousal per user

A.1.1 Adults

Table A.1 presents the mean and std as well as the normalized engagement values for each user averaged on all games each user played. By observation we can group users in three main categories, relatively to the interaction patterns.

- Users 1, 6, 8 and 12 constitute one group, where users in the WQ interaction turn display high attention, which significantly falls in the Q interaction turn and then raises gradually in the next interaction turns (WA and A).
- Users 2, 3, 4, 5, 11, 13 and 14 constitute another group (the larger one), where attention displays small alternations among the interaction turns. WQ and WA interaction turns display higher attention levels than Q and A. Differences among the interaction turns are insignificant.
- Last but not least, users 9 and 10 constitute the third group, where there is a small drop in attention through the transition from WQ to Q interaction turn, and then there is a gradual increase in the following turns.

First and third groups are quite alike, except the fact that in the first group engagement in the WP interaction type is much higher than attention in the following turns, while in the third group attention differences among the various turns are small. The

findings in Table A.1 suggest that there are significant differences among users ($p \approx 0$). Consequently interaction is much user dependent.

A.1.2 Children

		Interaction Turns				
		WP	P	WA	A	WE
USER 1	MEAN	70.86	40.72	47.31	66.50	62.93
	STD	10.09	22.84	6.89	9.66	5.03
	Mean-norm.	9.33	-8.50	-14.23	4.97	1.39
	Z-norm.	0.84	-0.77	-1.29	0.45	0.13
USER 2	MEAN	40.39	31.13	35.06	32.13	39.15
	STD	6.06	17.43	9.44	7.91	7.91
	Mean-norm.	3.34	1.49	-1.98	-4.91	2.11
	Z-norm.	0.42	0.19	-0.25	-0.62	0.27
USER 3	MEAN	40.65	36.99	37.72	35.87	49.98
	STD	9.64	20.69	11.72	1.55	4.71
	Mean-norm.	-0.39	4.15	-3.31	-5.18	8.93
	Z-norm.	-0.04	0.40	-0.32	-0.50	0.86
USER 4	MEAN	46.95	34.85	49.63	38.57	48.64
	STD	11.23	19.50	12.12	8.86	9.30
	Mean-norm.	0.22	-2.53	2.90	-8.16	1.92
	Z-norm.	0.02	-0.24	0.28	-0.77	0.18
USER 5	MEAN	47.33	43.89	51.84	36.50	46.00
	STD	9.26	24.54	11.97	3.03	19.52
	Mean-norm.	0.08	6.09	4.60	-10.74	-1.24
	Z-norm.	0.01	0.47	0.35	-0.82	-0.10
USER 6	MEAN	40.91	24.88	32.14	37.33	32.78
	STD	6.16	14.01	14.22	8.39	8.48
	Mean-norm.	6.25	-2.85	-2.52	2.68	-1.87
	Z-norm.	0.59	-0.27	-0.24	0.25	-0.18
USER 7	MEAN	62.76	51.20	47.88	57.20	53.76
	STD	5.50	28.71	14.20	7.73	6.14
	Mean-normalized	6.69	6.35	-8.19	1.13	-2.32
	Z-normalized	0.77	0.73	-0.94	0.13	-0.27
USER 8	MEAN	60.45	40.56	42.88	44.65	44.28
	STD	12.95	22.72	17.36	3.27	10.11
	Mean-norm.	11.29	1.24	-6.27	-4.50	-4.87
	Z-norm.	0.86	0.09	-0.48	-0.34	-0.37

Table continues on the next page

USER 9	MEAN	44.70	35.46	47.33	53.60	57.18
	STD	6.68	19.83	5.97	18.73	9.27
	Mean-norm.	-5.72	-4.87	-3.09	3.18	6.76
	Z-norm.	-0.54	-0.46	-0.29	0.30	0.64
USER 10	MEAN	64.44	52.61	54.05	62.30	56.23
	STD	5.44	29.42	13.98	2.83	12.44
	Mean-norm.	6.13	5.96	-4.26	3.99	-2.08
	Z-norm.	0.58	0.57	-0.41	0.38	-0.20
USER 11	MEAN	42.21	34.52	45.46	41.20	44.29
	STD	8.35	19.35	15.70	8.67	6.17
	Mean-norm.	-1.43	-0.39	1.83	-2.44	0.66
	Z-norm.	-0.13	-0.04	0.17	-0.23	0.06
USER 12	MEAN	64.43	35.62	51.38	51.27	55.77
	STD	15.69	20.05	13.28	2.66	17.70
	Mean-norm.	8.89	-6.59	-4.15	-4.27	0.23
	Z-norm.	0.60	-0.44	-0.28	-0.29	0.02
USER 13	MEAN	55.48	49.97	58.34	51.30	60.52
	STD	9.67	27.98	7.53	12.87	6.72
	Mean-norm.	-1.45	4.42	1.40	-5.63	3.59
	Z-norm.	-0.17	0.53	0.17	-0.67	0.43
USER 14	MEAN	50.75	22.29	41.97	36.20	45.34
	STD	17.91	14.81	7.79	1.70	5.99
	Mean-norm.	6.77	-8.49	-2.01	-7.78	1.36
	Z-norm.	0.59	-0.74	-0.17	-0.67	0.12
	Mean of Means	52.31	38.19	45.93	46.04	49.78

Table A.1: Mean, mean-normalized, Z-normalized and std engagement values for each user, per interaction turn (*Cont. from the previous page*)

		Interaction Turns				
		WQ	Q	WA	A	WE
USER 1	MEAN	62.92	47.85	61.28	66.42	73.14
	STD	3.04	26.91	4.13	2.38	16.49
	Mean-norm.	-2.84	-4.76	-4.47	0.66	7.39
	Z-norm.	-0.27	-0.45	-0.42	0.06	0.69
USER 2	MEAN	59.16	49.14	52.92	59.00	55.77
	STD	11.74	27.56	14.89	3.52	14.83
	Mean-norm.	3.54	4.64	-2.70	3.38	0.15
	Z-norm.	0.27	0.36	-0.21	0.26	0.01
USER 3	MEAN	64.61	43.86	60.54	56.27	64.42
	STD	11.22	24.53	5.80	8.76	6.82
	Mean-norm.	3.60	-4.95	-0.48	-4.75	3.41
	Z-norm.	0.42	-0.58	-0.06	-0.56	0.40
USER 4	MEAN	50.91	38.75	50.79	40.02	52.51
	STD	9.72	21.67	7.54	3.30	9.64
	Mean-norm.	1.50	-0.78	1.38	-9.40	3.10
	Z-norm.	0.17	-0.09	0.16	-1.07	0.35
USER 5	MEAN	53.83	51.81	55.18	52.92	59.72
	STD	13.40	29.00	4.78	10.26	8.86
	Mean-norm.	-1.93	7.21	-0.57	-2.84	3.97
	Z-norm.	-0.22	0.82	-0.06	-0.32	0.45
USER 6	MEAN	54.71	36.04	53.15	52.40	53.69
	STD	10.96	20.18	8.98	9.89	11.57
	Mean-norm.	1.83	-6.26	0.28	-0.48	0.81
	Z-norm.	0.19	-0.63	0.03	-0.05	0.08
USER 7	MEAN	49.66	42.60	62.62	54.56	48.97
	STD	6.59	23.85	7.05	4.18	5.61
	Mean-normalized	-2.19	1.12	10.77	2.71	-2.88
	Z-normalized	-0.30	0.15	1.48	0.37	-0.40

Table continues on the next page

USER 8	MEAN	62.69	55.43	48.50	60.53	62.65
	STD	7.59	30.99	13.33	7.37	7.76
	Mean-norm.	2.77	7.49	-11.42	0.61	2.74
	Z-norm.	0.31	0.84	-1.28	0.07	0.31
USER 9	MEAN	58.58	39.33	55.83	49.93	62.48
	STD	8.79	22.06	6.50	11.43	13.43
	Mean-norm.	0.97	-6.75	-1.78	-7.67	4.87
	Z-norm.	0.09	-0.61	-0.16	-0.70	0.44
USER 10	MEAN	58.07	45.93	47.37	62.80	53.21
	STD	10.25	25.72	6.47	12.81	12.29
	Mean-norm.	4.24	2.86	-6.47	8.97	-0.62
	Z-norm.	0.40	0.27	-0.61	0.85	-0.06
USER 11	MEAN	53.89	42.87	53.55	59.35	63.59
	STD	12.18	24.01	4.97	9.97	14.71
	Mean-norm.	-2.85	-2.52	-3.19	2.61	6.85
	Z-norm.	-0.27	-0.23	-0.30	0.24	0.64
USER 12	MEAN	67.36	53.70	67.72	59.98	70.08
	STD	9.26	30.81	8.09	7.15	8.31
	Mean-norm.	0.33	2.76	0.70	-7.04	3.06
	Z-norm.	0.04	0.33	0.08	-0.85	0.37
USER 13	MEAN	57.49	44.62	53.53	50.98	55.03
	STD	11.80	25.02	10.42	5.13	8.20
	Mean-norm.	2.99	1.02	-0.98	-3.52	0.53
	Z-norm.	0.33	0.11	-0.11	-0.38	0.06
USER 14	MEAN	54.09	32.07	45.07	40.80	54.73
	STD	12.09	21.84	7.10	5.37	13.53
	Mean-norm.	4.60	-2.57	-4.41	-8.69	5.24
	Z-norm.	0.41	-0.23	-0.40	-0.78	0.47
	Mean of Means	57.71	44.57	54.86	54.71	59.29

Table A.2: Mean and std Meditation values for each user, per interaction turn

		Interaction Turns				
		WQ	Q	WA	A	WE
USER 1	MEAN	1.30	0.35	1.37	1.29	0.99
	STD	0.55	0.20	0.69	0.49	0.65
	Mean-norm.	0.20	-0.53	0.27	0.19	-0.11
	Z-norm.	0.36	-0.96	0.48	0.35	-0.21
USER 2	MEAN	0.89	0.67	1.26	0.87	0.92
	STD	0.31	0.38	0.87	0.52	0.55
	Mean-norm.	-0.11	-0.13	0.26	-0.13	-0.08
	Z-norm.	-0.17	-0.22	0.43	-0.21	-0.14
USER 3	MEAN	1.78	0.47	1.33	2.49	1.32
	STD	1.83	0.26	0.92	1.96	0.53
	Mean-norm.	0.24	-0.76	-0.20	0.95	-0.21
	Z-norm.	0.20	-0.63	-0.17	0.79	-0.18
USER 4	MEAN	0.69	0.40	0.86	0.64	0.75
	STD	0.41	0.23	0.61	0.36	0.46
	Mean-norm.	-0.07	-0.20	0.11	-0.12	-0.01
	Z-norm.	-0.14	-0.43	0.22	-0.25	-0.02
USER 5	MEAN	1.09	0.40	1.02	0.86	0.92
	STD	0.48	0.24	0.50	0.48	0.42
	Mean-norm.	0.11	-0.38	0.05	-0.11	-0.05
	Z-norm.	0.26	-0.87	0.12	-0.25	-0.12
USER 6	MEAN	0.68	0.35	0.68	0.72	0.78
	STD	0.30	0.20	0.31	0.47	0.21
	Mean-norm.	-0.01	-0.21	-0.02	0.03	0.08
	Z-norm.	-0.04	-0.76	-0.05	0.10	0.30
USER 7	MEAN	1.03	0.50	0.73	0.76	0.93
	STD	0.53	0.28	0.14	0.16	0.46
	Mean-normalized	0.14	-0.21	-0.17	-0.13	0.04
	Z-normalized	0.36	-0.53	-0.41	-0.33	0.11

Table continues on the next page

USER 8	MEAN	1.30	0.42	1.25	0.79	1.26
	STD	0.94	0.25	0.41	0.39	0.75
	Mean-norm.	0.15	-0.50	0.10	-0.36	0.11
	Z-norm.	0.21	-0.72	0.14	-0.52	0.15
USER 9	MEAN	0.52	0.25	0.64	0.83	0.60
	STD	0.28	0.15	0.20	0.14	0.47
	Mean-norm.	-0.08	-0.23	0.04	0.23	0.001
	Z-norm.	-0.25	-0.70	0.12	0.69	0.004
USER 10	MEAN	1.00	0.31	0.69	0.85	0.69
	STD	0.42	0.17	0.38	0.24	0.28
	Mean-norm.	0.24	-0.30	-0.07	0.09	-0.07
	Z-norm.	0.68	-0.87	-0.21	0.26	-0.20
USER 11	MEAN	1.02	0.38	1.09	0.85	1.01
	STD	0.43	0.21	0.45	0.24	0.82
	Mean-norm.	0.01	-0.43	0.09	-0.15	0.002
	Z-norm.	0.02	-0.83	0.17	-0.30	0.004
USER 12	MEAN	1.03	0.39	1.17	0.96	0.92
	STD	0.77	0.22	0.70	0.34	0.54
	Mean-norm.	0.02	-0.38	0.16	-0.05	-0.09
	Z-norm.	0.04	-0.63	0.27	-0.08	-0.14
USER 13	MEAN	0.84	0.38	0.83	1.17	1.08
	STD	0.32	0.22	0.52	0.25	0.40
	Mean-norm.	-0.11	-0.38	-0.12	0.22	0.13
	Z-norm.	-0.26	-0.90	-0.29	0.54	0.30
USER 14	MEAN	0.77	0.39	1.05	0.88	1.02
	STD	0.23	0.23	0.48	0.30	0.41
	Mean-norm.	-0.15	-0.26	0.12	-0.05	0.09
	Z-norm.	-0.41	-0.69	0.32	-0.12	0.24
	Mean of Means	1.00	0.40	1.00	1.00	0.94

Table A.3: Mean and std Arousal values for each user, per interaction turn

			Interaction Turns					
			WP	P	WA	A	WE	Overall
USER 1	Diff=1	MEAN	35.52	38.96	33.30	43.80	49.81	38.80
		STD	9.91	22.03	9.38	4.13	5.55	10.32
		Mean-norm.	-3.28	7.92	-5.50	5.00	11.01	
		Z-norm.	-0.32	0.77	-0.53	0.48	1.07	
	Diff=2	MEAN	50.84	42.23	49.89	51.19	42.09	47.43
		STD	7.04	23.61	7.89	10.20	5.88	8.05
		Mean-norm.	3.41	4.29	2.46	3.76	-5.34	
		Z-norm.	0.42	0.53	0.31	0.47	-0.66	
USER 2	Diff=1	MEAN	54.85	40.73	41.95	54.38	50.85	45.08
		STD	9.63	22.79	14.06	1.19	8.80	13.31
		Mean-norm.	9.77	4.67	-3.12	9.31	5.77	
		Z-norm.	0.73	0.35	-0.23	0.70	0.43	
	Diff=2	MEAN	43.25	21.37	38.15	40.64	43.09	39.16
		STD	6.24	12.21	9.51	5.84	8.14	8.97
		Mean-norm.	4.09	-9.95	-1.01	1.48	3.93	
		Z-norm.	0.46	-1.11	-0.11	0.17	0.44	
USER 3	Diff=1	MEAN	62.13	41.73	47.54	46.28	47.07	49.55
		STD	13.88	23.34	9.02	3.90	10.84	10.79
		Mean-norm.	12.58	2.09	-2.02	-3.28	-2.48	
		Z-norm.	1.17	0.19	-0.19	-0.30	-0.23	
	Diff=2	MEAN	59.11	56.18	66.87	62.67	59.38	62.05
		STD	12.38	31.45	15.99	3.79	10.52	13.10
		Mean-norm.	-2.94	6.54	4.82	0.62	-2.67	
		Z-norm.	-0.22	0.50	0.37	0.05	-0.20	
USER 4	Diff=1	MEAN	49.94	34.07	46.05	47.30	47.23	47.10
		STD	8.52	19.08	8.76	1.25	3.82	7.29
		Mean-norm.	2.84	-3.61	-1.05	0.20	0.13	
		Z-norm.	0.39	-0.49	-0.14	0.03	0.02	
	Diff=2	MEAN	55.55	34.31	53.42	41.71	47.29	51.36
		STD	16.59	19.19	9.31	5.00	8.84	10.98
		Mean-norm.	4.19	-6.78	2.06	-9.65	-4.06	
		Z-norm.	0.38	-0.62	0.19	-0.88	-0.37	

USER 5	Diff=1	MEAN	56.33	39.75	39.32	35.52	46.44	44.63
		STD	14.63	22.33	9.85	5.93	5.63	11.73
		Mean-norm.	11.70	4.05	-5.31	-9.11	1.81	
		Z-norm.	1.00	0.34	-0.45	-0.78	0.15	
	Diff=2	MEAN	29.02	24.04	21.53	26.63	29.93	25.33
		STD	9.18	13.47	11.27	2.67	7.17	9.58
		Mean-norm.	3.68	3.78	-3.80	1.30	4.59	
		Z-norm.	0.38	0.39	-0.40	0.14	0.48	
USER 6	Diff=1	MEAN	38.39	36.21	51.45	45.33	45.22	48.17
		STD	6.52	20.39	8.09	10.98	6.59	8.67
		Mean-norm.	-9.78	-2.32	3.29	-2.84	-2.95	
		Z-norm.	-1.13	-0.27	0.38	-0.33	-0.34	
	Diff=2	MEAN	40.00	34.95	43.05	32.20	47.36	42.19
		STD	6.74	19.56	12.10	2.31	7.19	10.92
		Mean-norm.	-2.19	1.20	0.85	-9.99	5.17	
		Z-norm.	-0.20	0.11	0.08	-0.92	0.47	
	Diff=1	Mean of Means	49.53	38.56	43.27	45.44	47.77	44.91
	Diff=2	Mean of Means	46.30	35.51	45.49	42.51	44.86	42.93

Table A.4: Mean, mean-normalized, Z-normalized and std engagement values for each user, per interaction turn

			Interaction Turns					
			WP	P	WA	A	WE	Overall
USER 1	Diff=1	MEAN	45.39	49.45	47.68	58.03	64.17	51.79
		STD	11.16	27.65	10.59	2.04	5.24	11.41
		Mean-norm.	-6.40	8.02	-4.11	6.24	12.37	
		Z-norm.	-0.56	0.70	-0.36	0.55	1.08	
	Diff=2	MEAN	57.90	43.38	44.12	52.68	50.70	49.20
		STD	5.24	24.31	15.12	12.95	9.52	12.52
		Mean-norm.	8.70	4.01	-5.08	3.48	1.49	
		Z-norm.	0.69	0.32	-0.41	0.28	0.11	
USER 2	Diff=1	MEAN	42.56	32.98	53.71	57.68	51.53	51.81
		STD	10.81	18.44	18.35	2.97	3.88	16.25
		Mean-norm.	-9.25	-8.46	1.91	5.88	-0.27	
		Z-norm.	-0.57	-0.52	0.12	0.36	-0.02	
	Diff=2	MEAN	39.72	26.13	47.16	39.75	36.99	42.68
		STD	3.44	14.62	10.75	1.09	5.03	9.70
		Mean-norm.	-2.96	-8.01	4.49	-2.93	-5.69	
		Z-norm.	-0.31	-0.83	0.46	-0.30	-0.59	
USER 3	Diff=1	MEAN	47.89	37.82	47.92	56.53	41.39	46.76
		STD	7.39	21.15	7.13	3.93	8.71	8.07
		Mean-norm.	1.13	0.41	1.16	9.77	-5.37	
		Z-norm.	0.14	0.05	0.14	1.21	-0.67	
	Diff=2	MEAN	41.27	42.43	49.90	38.13	44.82	45.61
		STD	9.46	24.29	7.06	2.60	13.26	9.20
		Mean-norm.	-4.34	5.94	4.29	-7.48	-0.79	
		Z-norm.	-0.47	0.65	0.47	-0.81	-0.09	
USER 4	Diff=1	MEAN	63.97	37.97	49.85	52.47	48.52	52.65
		STD	16.89	21.24	8.59	2.34	4.82	11.41
		Mean-norm.	11.32	-4.15	-2.80	-0.18	-4.13	
		Z-norm.	0.99	-0.36	-0.26	-0.02	-0.36	
	Diff=2	MEAN	48.79	38.89	60.52	45.40	52.87	55.12
		STD	10.21	21.75	12.95	2.60	6.18	11.93
		Mean-norm.	-6.33	-5.20	5.40	-9.72	-2.25	
		Z-norm.	-0.53	-0.44	0.45	-0.81	-0.19	

USER 5	Diff=1	MEAN	76.94	53.76	70.54	69.12	59.46	68.81
		STD	12.72	30.18	6.38	2.94	7.54	10.06
		Mean-norm.	8.12	-1.29	1.73	0.30	-9.35	
		Z-norm.	0.81	-0.13	0.17	0.03	-0.93	
	Diff=2	MEAN	44.80	39.74	44.74	53.97	52.71	47.46
		STD	15.60	22.22	10.67	5.36	10.36	11.30
		Mean-norm.	-2.66	1.77	-2.72	6.51	5.26	
		Z-norm.	-0.24	0.16	-0.24	0.58	0.47	
USER 6	Diff=1	MEAN	53.83	47.45	55.05	62.26	58.23	55.99
		STD	7.22	26.59	9.60	13.42	8.63	9.22
		Mean-norm.	-2.16	2.66	-0.94	6.27	2.24	
		Z-norm.	-0.23	0.29	-0.10	0.68	0.24	
	Diff=2	MEAN	50.72	37.87	45.80	54.22	62.48	49.24
		STD	6.77	21.21	10.85	9.92	11.72	11.58
		Mean-norm.	1.48	-1.52	-3.44	4.99	13.24	
		Z-norm.	0.13	-0.13	-0.30	0.43	1.14	
	Diff=1	Mean of Means	55.10	43.24	54.13	59.35	55.55	53.47
	Diff=2	Mean of Means	47.20	38.07	48.71	47.36	50.10	46.29

Table A.5: Mean, mean-normalized, Z-normalized and std meditation values for each user, per interaction turn

			Interaction Turns					
			WP	P	WA	A	WE	Overall
USER 1	Diff=1	MEAN	0.92	0.37	1.02	0.78	0.80	0.90
		STD	0.43	0.21	0.42	0.55	0.47	0.42
		Mean-norm.	-0.02	-0.35	0.12	-0.11	-0.10	
		Z-norm.	0.06	-0.83	0.29	-0.27	-0.24	
	Diff=2	MEAN	0.91	0.38	1.02	0.70	0.71	0.85
		STD	0.43	0.21	0.59	0.30	0.33	0.47
		Mean-norm.	0.06	-0.30	0.17	-0.15	-0.14	
		Z-norm.	0.12	-0.65	0.36	-0.32	-0.31	
USER 2	Diff=1	MEAN	0.92	0.29	0.60	0.88	0.70	0.65
		STD	0.33	0.17	0.40	0.41	0.32	0.65
		Mean-norm.	0.27	-0.22	-0.05	0.23	0.05	
		Z-norm.	0.71	-0.58	-0.13	0.59	0.14	
	Diff=2	MEAN	0.76	0.24	0.56	0.85	0.82	0.64
		STD	0.27	0.14	0.36	0.54	0.50	0.38
		Mean-norm.	0.11	-0.27	-0.08	0.21	0.18	
		Z-norm.	0.30	-0.72	-0.21	0.56	0.46	
USER 3	Diff=1	MEAN	0.91	0.25	0.75	0.69	1.11	0.84
		STD	0.20	0.14	0.35	0.40	0.74	0.46
		Mean-norm.	0.07	-0.43	-0.10	-0.16	0.26	
		Z-norm.	0.15	-0.93	-0.21	-0.34	0.58	
	Diff=2	MEAN	0.76	0.51	0.70	1.14	0.93	0.80
		STD	0.26	0.29	0.24	0.15	0.34	0.28
		Mean-norm.	-0.03	-0.13	-0.10	0.34	0.13	
		Z-norm.	-0.12	-0.47	-0.36	1.24	0.47	
USER 4	Diff=1	MEAN	0.61	0.36	0.80	0.66	0.81	0.74
		STD	0.42	0.21	0.41	0.21	0.50	0.40
		Mean-norm.	-0.13	-0.23	0.06	-0.08	0.08	
		Z-norm.	-0.32	-0.58	0.15	-0.19	0.20	
	Diff=2	MEAN	0.77	0.24	0.70	0.69	0.58	0.67
		STD	0.34	0.14	0.51	0.06	0.14	0.41
		Mean-norm.	0.10	-0.29	0.03	0.02	-0.09	
		Z-norm.	0.25	-0.72	0.08	0.05	-0.21	

USER 5	Diff=1	MEAN	0.87	0.34	0.76	0.66	0.97	0.82
		STD	0.58	0.20	0.27	0.06	0.41	0.38
		Mean-norm.	0.05	-0.31	-0.05	-0.15	0.15	
		Z-norm.	0.14	-0.83	-0.14	-0.41	0.40	
	Diff=2	MEAN	0.87	0.33	0.77	0.53	0.73	0.75
		STD	0.71	0.21	0.43	0.08	0.44	0.46
		Mean-norm.	0.12	-0.27	0.02	-0.22	-0.02	
		Z-norm.	0.26	-0.58	0.05	-0.48	-0.04	
USER 6	Diff=1	MEAN	0.80	0.41	1.06	0.88	0.99	0.99
		STD	0.52	0.23	0.43	0.45	0.60	0.47
		Mean-norm.	-0.19	-0.38	0.07	-0.11	-0.002	
		Z-norm.	-0.40	-0.82	0.16	-0.23	-0.004	
	Diff=2	MEAN	0.74	0.41	1.29	0.91	0.95	1.11
		STD	0.34	0.23	0.93	0.30	0.62	0.81
		Mean-norm.	-0.37	-0.48	0.17	-0.21	-0.16	
		Z-norm.	-0.46	-0.60	0.21	-0.26	-0.20	
	Diff=1	Mean of Means	0.80	0.34	0.83	0.76	0.90	0.73
	Diff=2	Mean of Means	0.80	0.35	0.84	0.80	0.79	0.72

Table A.6: Mean, mean-normalized, Z-normalized and std arousal values for each user, per interaction turn

References

[1] EEG Band Power values: Units and Meaning. http://support.neurosky.com/kb/development-2/eeg-band-power-values-units-and-meaning. 73

[2] eSense (tm) Meters. http://developer.neurosky.com/docs/doku.php?id=esenses_tm. 69

[3] NeuroSky: Brain Wave Sensors for Every Body. http://www.neurosky.com. 67, 68

[4] NeuroSky Store. http://store.neurosky.com/products/mindset. 108

[5] Poor Signal Quality. http://developer.neurosky.com/docs/doku.php?id=signal_quality. 73

[6] What are the different EEG Band Frequencies? http://support.neurosky.com/kb/technology/eeg-band-frequencies. 71

[7] A. Bruckman and A. Bandlow. Human Computer Interaction for kids. 2003. 19

[8] A. Druin and K. Inkpen. When are Personal Technologies for Children? *Personal and Ubiquitous Computing*, **5**:191–194, 2001. 8, 12, 99

[9] A. Druin, B. Bederson, A. Boltman, A. Miura, D. Knotts-Callahan and M. Platt. Children as Our Technology Design Partners. *The Design of Children's Technology: How We Design, What We Design and Why*, 1998. 8

[10] A. Jameson. Adaptive Interfaces and Agents. 2008. 12

[11] A. Johnson and N. Taatgen. User Modeling: Handbook of human factors in Web design. pages 424–439, 2005. 64

[12] A. Potamianos, H. K. Kuo, C. H. Lee, A. Pargellis, A. Saad and Q. Zhou. Design Principles and Tools for Multimodal Dialog Systems. In *Proc. of ESCA Workshop on Interactive Dialogue in Multi-Modal Systems (IDS-99)*, 1999. 6

[13] B. Appoloni and R. J. Howlett. Knowledge-Based Intelligent Information and Engineering Systems. In *Proc. 11th international Conf. XVII Italian Workshop on Nueral Networks*, 2003. 15

[14] B. Schuller, G. Rigoll and M. Lang. Speech Emotion Recognition Combinig Acoustic Features and Linguistic information in a hybrid vector Machine - Belief Network Architecture. 2004. 15

[15] B. Xiao, C. Girand and S. Oviatt. Multimodal Integration Patterns in Children. In *Proc. of ICSLP'02*, pages 629–632, 2002. 8

[16] C. Busso, Z. Deng , S. Yildirim, M. Bulut, C. M. Lee, A. Kazemzadeh, S. Lee, U. Neumann , S. Narayanan. Analysis of Emotion Recognition using Facial Expressions, Speech and Multimodal Information. In *ICMI*, 2004. 15, 16

[17] C. Klimmt. Dimensions and Determinants of Enjoyment of Playing Digital Games. In *In Proc. Level Up: Digital Games Research Conf.*, pages pp 246–257, 2003. 15

[18] C. M. Lee, S. S. Narayanan and R. Pieraccini. Combining acoustic and Language information for Emotion Recognition. In *ICSLP*, 2002. 16

[19] C. M. Lee, S. S. Narayanan, R. Pieraccini. Classifying Emotions in human-machine spoken dialogs. 15, 56

[20] C. Osgood, G. Suci and P. Tannenbaum. The Measurement of Meaning. 1957. 13, 36

[21] D. Heise. Enculturing agents with expressive role behavior. *Agent Culture: Human-Agent Interaction in a Multicultural World*, pages 127–142. 65

[22] D. O. Bos. EEG-based Emotion Recognition. Available: `http://emi.uwi.utwente.nl/verslagen/capita-selecta/CS-oudeBos-Danny.pdf`, 2006. 16, 73

[23] D. Salber and J. Coutaz. Applying the Wizard of Oz technique to the study of multimodal systems. In *Proc. of EWCHI'93*, 1993. 7

[24] D. Salher and J. Coutaz. A Wizard of Oz Platform for the study of Multimodal Systems. 7

[25] D. Ververidis and C. Kotropoulos. Emotional speech recognition: Resources, features, and methods. 2006. 15

[26] E. V. Orekhova, T. A. Stroganova, I. N. Posikera and M. Elam. EEG theta rhythm in infants and preschool children. 16

[27] F. Dellaert, T. Polzin and A. Waibel. Recognizing Emotion in Speech. 15

[28] F. Eyben, M. Woellmer and B. Schuller. openSMILE: the Munich open Speech and Music Interpretation by Large Space Extraction toolkit. 28

[29] F. Nasoz, K. Alvarez, C. L. Lisetti and N. Finkelstein. Emotion Recognition from Physiological Signals for Presence Technologies. **vol. 6**, 2003. 15

[30] G. Chanel, K. Ansari-Asl and T. Pun. Valence-arousal evaluation using physiological signals in an emotion recall paradigm. 2007. 15

[31] G. Fischer. User Modeling in Human-Computer Interaction. *User Modeling and User-Adapted Interaction*, **11**:65–68, 2001. 64

[32] G. N. Yannakakis and J. Hallam. Evolving opponents for interesting interactive games. pages pp 499–508, 2004. 12

[33] G. N. Yannakakis and J. Hallam. Entertainment Modeling in Physical Play through Physiology beyond Heart-Rate. In *In Proc. of the International Conference on Affective Computing and intelligent Interaction*, pages pp 256–267, 2007. 12

[34] G. N. Yannakakis and J. Togelius. Experience-Driven Procedural Content Generation. 2011. 12

[35] G. N. Yannakakis, J. Hallam and H. H. Lund. Capturing entertainment through heart-rate dynamics in the play-ware playground. In *In Proc. of the 5th International Conference on Entertainment Computing*, 2006. 12, 15

[36] Joakim Gustafson, Linda Bell, Johan Boye, Anders Lindström, and Mats Wirén. The nice fairy-tale game system. In Michael Strube and Candy Sidner, editors, *Proc. of the 5th SIGdial Workshop on Discourse and Dialogue*, pages 23–26. Association for Computational Linguistics, 2004. 8, 21

[37] J. Cichosz and K. Slot. Emotion recognition in speech signal using emotion-extracting binary decision trees. 15

[38] J. Hothi and W. Hall. An Evaluation of Adapted Hypermedia Techniques using Static User Modeling. In *Proc. of the 2nd Workshop on Adaptive Hypertext and Hypermedia*, 1998. 64

[39] J. Russel and A. Mehrabian. Evidence for a three-factor theory of emotions. **vol. 11**(no. 3):pp 273–294, 1977. 13, 36

[40] J. Tao and T. Tieniu. Affective Computing: A Review. *Affective Computing and Intelligent Interaction*, pages 981–995, 2005. 65

[41] K. Takahashi. Remarks on Emotion Recognition from Bio-porential signals. In *Proc. 2nd International Conference on Autonomous Robots and Agents*, 2004. 15

[42] Kristina Schaaff. *Challenges on Emotion Induction with the International Affective Picture System*. Master's thesis, University of Karlsruhe, 2008. 41

[43] L. F. Barrett. Discrete Emotions or Dimensions? The Role of Valence Focus and Arousal Focus. 1998. 13

[44] L. M. Reeves, J. Lai, J. A. Larson, S. Oviatt, T. S. Balajil, S. Buisine, P. Collings, P. Cohen, B. Kraal, J. C. Martin, M. McTear, T. V. Raman, K. M. Stanney, H. Su and Q. Y. Wang. Guidelines for Multimodal User Interface Design. *Communications of the ACM*, **47(1)**:57–59, 2004. 6

[45] M. Bradley. Emotional memory: A dimensional analysis. 1994. 13, 36, 54, 61

[46] M. Csikszentmihlyi. Flow: The Psychology of Optimal Experiense. 1990. 10

[47] M. Grimm,K. Kroschel, E. Mower, S. Narayanan. Primitives-based evaluation and estimation of emotions in speech. 2006. 15

[48] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann and I. H. Witten. The WEKA Data Mining Software: An Update. *ACM SIGKDD Explorations*, **11(1)**, 2009. 38

[49] M. Mikhail, K. El-Ayat, R. El kaliouby, J. Coan and J. J. B. Allen. Emotion Detection using Noisy EEG Data. In *Proc. Augmented Human Conference*, 2010. 15

[50] M. Shami and W. Verhelst. An evaluation of the robustness of existing supervised machine learning approaches to the classification of emotions in speech. 2007. 15

[51] M.F.McTear. Spoken Dialogue Technology: Enabling the Conversational User Interface. *ACM Computing surveys*, **34(1)**, 2002. 6

[52] M.Johnston, S.Bangalore, G.Vasireddy, A.Stent, P.Ehlen, M.Walker, S.Whittaker and P.Mallor. MATCH: An Architecture for Multimodal Dialogue Systems. In *Proc. of the 40th Annual Meeting on Association for Computational Linguistis*, 2002. 8

[53] M.Perakakis and A.Potamianos. A Study in Efficiency and Modality Usage in Multimodal Form Filling Systems. *Audio, Speech, and Language Processing, IEEE Transactions*, **16**, 2008. 21

[54] N. Garay, C. Idoia, J. M. Lopez and F. Inmaculada. Assistive Technology and Affective Mediation. *Human Technology*, **2(1)**:55–83, 2006. 65

[55] N. H. Frijda. The Emotions. page p 207, 1986. 36

[56] N. Lazzaro. Why We Play Games: Four Keys to More Emotion Without Story, 2004. 11, 15

[57] P. Boersma and D. Weenink. Praat: doing phonetics by computer (Version 5.2.21). 28

[58] P. Ekman and W. V. Friesen. Facial Action Coding System: Investigator's Guide. 1978. 15

[59] P. J. Lang. The Network Model of Emotion: Motivational Connections. 1995. 13, 36, 52, 56

[60] P. M. Niedenthal and S. Kitayama. The Heart's Eye-Emotional Influences in Perception and Attention. 1994. 13

[61] P.Cohen, M.Johnston, D. McGee, S.Oviatt, J.Pittman, I.Smith, L.Chen and J.Clow. QuickSet: Multimodal Interaction for Distributed Applications. In *Proc. of the fifth International Multimedia Conference (Multimedia '97)*, pages 31–40, 1997. 7

[62] R. Horlings. *Emotion recognition using brain activity.* Master's thesis, TU Delft, 2008. 67

[63] R. W. Picard, E. Vyzas and J. Healey. Toward Machine Emotional Intelligence: Analysis of Affective Physiological State. **vol. 23**(no. 10), 2001. 15

[64] Richard A. Bolt. "Put-that-there": Voice and Gesture at the Graphics Interface. *Computer Graphics*, 1980. 8

[65] S. Lee, A. Potamianos and S. Narayanan. Acoustics of childrenb••s speech: Developmental changes of temporal and spectral parameters. **vol. 10**. 19, 28

[66] S. Narayanan, A. Potamianos and H. Wang. Multimodal Systems For Children: Building a Prototype. In *Proc. of EuroSpeech*, 1999. 8

[67] S. Narayanan and A. Potamianos. Creating Conversational Interfaces for Children. *IEEE Transactions on Speech and Audio Processing*, **10(2)**, 2002. 8, 19, 21

[68] S. Oviatt. Designing Robust Multimodal Systems for Universal Access. In *Proc. of WUAUC'01*, 2001. 7

[69] S. Oviatt and P. Cohen. Multimodal Interfaces that Process What Comes Naturally. **43(3)**, 2000. 5, 7

[70] S. Yildirim, C. M. Lee, S. Lee, A. Potamianos, S. Narayanan. Detecting Politeness and Frustration State of a Child in a Conversational Computer Game. 16

[71] T. Harmony, T. Fernandez, J. Silva, J. Bernal, L. Diaz-Comas, A. Reyes, E. Marosi, M. Rodriguez and M. Rodriguez. EEG delta activity: an indicator of attention to internal processing during performances of mental tasks. 1996. 12

[72] T. Kannetis, A. Potamianos and G.N. Yannakakis. Fantasy, Curiosity and Challenge as Adaptation Indicators in multimodal Dialogue Systems for Preschoolers. In *Proc. Workshop on Child, Computer and Interaction (WOCCI b••09)*, 2009. xv, 17, 19, 20, 21, 22, 24, 99

[73] T. Kannetis and A. Potamianos. Towards adapting fantasy,curiosity and challenge in multimodal dialogue systems for preschoolers. In *Proc. International Conf. on Multimodal Interaction (ICMI b••09)*, 2009. 17, 21, 24, 99

[74] T. L. Nwe, S. W. Foo, L. C. De Silva. Speech emotion recogition using hidden Markov models. 2003. 15

[75] T. O. Zander, C. Kothe, S. Jatzev and M. Gaertner. Enhancing Human-Computer Interaction with Input from Active and Passive Brain-Computer Interfaces. *Brain-Computer Interfaces*, pages 181–199, 2010. 9

[76] T. Partala, V. Surakka and T. Vanhala. Real-time estimation of emotional experiencies from facial expressions. **vol. 18**(no. 2):pp 208–226, 2006. 15

[77] T. W. Malone. Heuristics for designing enjoyable user interfaces: Lessons from computer games. In *Proc. of the Conf. on Human Factors in Computing Systems (CHI '82)*, 1982. 2, 10

[78] T.W. Malone. What make things fun to learn? Heuristics for Designing Instructional Computer Games. In *Proc. of the 3rd ACM SIGSMALL symp. and the first SIGPC symp. on Small systems (SIGS-MALL b••80)*, 1980. 2, 10, 23, 24, 25

[79] W.A. Teder-Salejarvi, J.J. McDonald, F. Di Russo, S.A. Hillyard. An analysis of audio-visual crossmodal integration by means of event-related potential (ERP) recordings. **14**:106–114, 2002. 80

[80] Y. Du , W. Bi , T. Wang , Y. Zhang , H. Ai. Distributing Expressional Faces in 2-D Emotional Space. In *CIVR*, 2007. 15

[81] Y. Li and Y. Zhao. Recognizing emotions in speech using short-term and long-term features. 15