Multimodal Dialogue Systems For Preschoolers

by

Theofanis Kannetis

Submitted in partial fulfillment of the
requirements for the degree of
M.Sc in Electronic and Computer Engineering
at the
TECHNICAL UNIVERSITY OF CRETE

CHANIA, GREECE
July 2009

“Anybody who has been seriously engaged is scientific work of any kind realizes that over the entrance to the gates of the temple of science are written the words: ’Ye must have faith.’ It is a quality which the scientist cannot dispense with.”

– Max Plank.

Abstract

Multimodal dialogue systems are becoming increasingly part of our everyday life, mainly because of their naturalness, robustness and efficiency. One interesting and relevant field of research is multimodal dialogue systems for children. Computers have become a popular learning and playing tool for young children, and multimodal approach seems promising for the creation of interfaces for children. The design and implementation of such interfaces however, can be tricky and challenging, mainly because the way that children interact with such systems is very different than the adults.

In this work, we investigate how preschool children interact with multimodal dialogue systems, and identify factors that are good indicators towards adapting multimodal dialogue systems for preschoolers. For this purpose, an on-line multimodal platform has been designed, implemented and used as a starting point to develop web-based speech-enabled applications for children. In order to investigate how fantasy, curiosity and challenge contribute to the user experience, varying levels of fantasy and curiosity elements, as well as, variable difficulty levels were implemented in five task oriented games, suitable for preschoolers.

Nine preschool children were asked to play these games in three sessions; in each session only one of the fantasy, curiosity or challenge factor was varied. Both objective and subjective criteria were used to evaluate the factors and applications. Results show that fantasy and curiosity are correlated with children's entertainment (user satisfaction), while the optimum level of difficulty seems to depend on each child's individual preferences and capabilities. Preliminary experiments also showed that interaction patterns and acoustic features are indicators of (subjectively) optimal levels of fantasy, curiosity and difficulty. Moreover the collected data was used in order to build emotion classifiers as a step toward adapting multimodal dialogue systems for preschoolers, with promising results.

Acknowledgements

I would like to thank my supervisor Associate Professor Alexandros Potamianos for his valuable guidance. Also I wish to thank Professor Georgios N. Yannakakis (IT University of Copenhagen) and Adjunct Assistant Professor A. Mouzaki (University of Crete), for the valuable discussions we had, helped me to become familiar with research fields that are related with my work.

I would also like to thank Professor Vasilis Digalakis and Associate Professor Katerina Mania for participating in the examining committee.

Next I would like to thank Manolis Perakakis (Ph.D) who was always there for me, giving me guidelines for several aspects of my work. Furthermore, i would like to thank Spyros Meliopoulos for implementing the "Farm" task, and Giorgos Evgeneiadis for helping out in the evaluation. I could not forget to send a big "thank you" to the preschool teachers from 1st Preschool of Kounoupidiana for their help during the data collection and evaluation, and all the kids who participate in the evaluation.

Also I want to acknowledge all my wonderful friends and colleagues, some of whom spent hours with me in the laboratory, for their company and guidance during all these years.

Finally, i would like to thank my family, my brother Andreas, my sister Christothea and especially my parents Michael and Arsinoe.

Contents

List of Figures

List of Tables

Dedicated to my family…

Chapter 1

Introduction

1.1 Introduction

In the past few years multimodal systems are becoming increasingly part of our everyday life, e.g., mobile communication devices. Multimodal systems combine multiple input and output modalities, such as, keyboard, pen, speech, touch/multi-touch, in order to increase the naturalness, robustness and efficiency of human-computer interaction. One interesting and relevant field of research in this area is multimodal dialogue systems for children.

Multimodal interaction systems for children have become increasingly popular during the past few years, mainly because of the difficulties that younger children have, using the conventional input modalities like mouse and keyboard. As a result their interaction with interactive systems such as games and educational software, is not as efficient and pleasant compare to adults. For that reason different input modalities such as gestures and speech can help in order to increase usability in computer-children interaction.

Although children are good recipients of technology, there are several pitfalls and challenges in the design and the implementation of multimodal dialogue systems for children. This is because children are quite different than adults, both in terms of their interaction with such systems, and in human factors standpoint. Recent studies have shown that children have a tendency to interact with multimodal systems using more than one modality simultaneously, more often than adults. Moreover, children's enjoy more the mix of exploration and exploitation, especially at younger ages. In terms of human factors, it is shown that there is acoustic and linguistic variability in children speech, and this is a significant problem in speech recognition technology.

However, the challenges does not prevent the creation of multimodal dialogue systems for children. In the recent years lot of research has be done in the design and implementation of multimodal dialogue systems for children. A number of prototype systems with advanced spoken dialogue interfaces, multimodal interaction capabilities and/or embodied conversational characters have been implemented. However, almost all of these systems have focused in the age group 6-15. And that is because older children can be more easily controlled in experimental conditions than younger children.

To investigate how preschool children interact with multimodal dialogue systems, we have designed and implemented an on-line (web-based) multimodal platform, in order to be able to quickly prototype, deploy and evaluate multimodal dialogue systems for preschoolers. Using this platform we have designed five such games that use speech and mouse as input modalities.

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

Based on these prior studies, one of the main goals of this work, is to identify some of the parameters that we can adapt in such systems so that we can increase user satisfaction and entertainment. For that reason several entertainment factors have been implemented in our games and evaluate in order to identify if they are good indicators towards adapting such systems.

The main contribution of this work is to provide a multimodal platform that can be used in the future, not only for studying children-computer interaction, but for any web-based multimodal application. Furthermore there are only few works studying the integration patterns at preschool ages and the child computer interaction with such systems. This work provides more information that could help in further studies towards developing better applications for children. Finally by identifying which of the factors that make a good game, that exist in the literature, are good indicators towards adapting multimodal dialogue systems for children, the goal for better multimodal interfaces for children is even closer. That is because in this ages entertainment and learning are intertwined activities and having a usable and entertain interface is crucial in the learning procedure.

The remainder of the thesis is organized as follows. Chapter 2 discuss about multimodal dialogue systems and the challenges in the design of multimodal dialogue systems for childern. The existing work towards adapting multimodal dialogue systems for preschoolers is discussed in Chapter 3. The architecture and the implementation of the multimodal platform is described in Chapter 4. Then the functionality and user interface of the five implemented multimodal games are presented in Chapter 5. The development of fantasy and curiosity triggers, as well as, the different levels of game difficulty are described in the same chapter. In Chapter 6, the evaluation method is describe while in Chapter 7 the objective and subjective evaluation results for nine subjects are presented and discussed. Finally conclusions and future directions are presented in Chapter 8.

Chapter 2

Multimodal Dialogue Systems

2.1 Introduction

During the past three decades, computers from simple “calculation machines” have become indispensable components of our daily life with multiple uses. As a result of the growing computer usage, more people interact daily with computers or machines that contain computers. From the early years of the computers an important issue was human-computer interaction. Human was trying to find new ways to interact with computers in order to make the interaction with the machines more efficient and easy. As the computers were evolved, the increase of their computational power allowed the implementation of the first graphic user interfaces and formed the human computer interaction as we know it today, with the mouse and keyboard as the main input modalities and GUI as the main output modality.

Nevertheless the quest for new ways of interaction was never abandoned. As a result computational devices with different sizes, computational power and input/output capabilities are appeared, and the future of computing is likely to include modalities such as gestures, speech, haptics, eye-gaze, opening the door to new applications designed specifically for such devices.

But what are the multimodal interfaces? Multimodal interfaces are interfaces where the communication between the system and the user is made through various modalities such as speech, gesture recognition, pen, GUI, etc. These systems have the ability to fuse the information through the various modalities, to decide which is the best modality for communication at each point of the interaction and are able to disambiguate the input from a modality with the use of another one. The various modalities are not concern only the input to the system but also the output as well.

Multimodal applications range from map-based and virtual reality systems for simulation and training, to multi-biometric person identification systems for security purposes, to medical, educational, and web-based transaction systems. They have evolved rapidly during the past decade, with steady progress, mainly because they have many advantages regarding the human computer interaction.

2.2 Advantages of multimodal systems

One particularly advantageous feature of such multimodal systems is their superior error handling. By developing multimodal interfaces the performance stability and robustness of recognition-based systems is improved. From a usability standpoint, multimodal systems offer a flexible interface in which users can exercise intelligence about how to use input modes effectively so that errors are avoided. To reap these error-handling advantages fully, a multimodal systems are designed so that users can achieve their goals with any of the available modalities. In that way, when an error is occur (e.g. from speech recognizer) users can use another modality (keyboard, pen, gesture etc) in order to fix it. Also in a case of input ambiguity (this happens usually with speech modality), multimodal systems are able to disambiguate input with the use of another modality. As a result multimodal systems demonstrate a relatively greater performance compared to unimodal systems. The error suppression achievable with a bimodal system, compared with a unimodal one, can be in excess of 20-40% and can be improve in systems that use more modalities.

Another important reason for developing multimodal interfaces are their potential to greatly expand the accessibility of computing to diverse and non-specialist users, and to promote new forms of computing not previously available. Since there can be large individual differences in people's ability and preference to use different modes of communication, multimodal interfaces will increase the accessibility of computing for users of different ages, skill levels, cognitive styles, sensory and motor impairments, native languages, or even temporary illnesses. This is because a multimodal interface permits users to have the control over how they interact with the system. For example, a visually impaired user may prefer speech input, as may a manually impaired user with a repetitive stress injury or with his arm in a cast. In contrast, a user with a hearing impairment, strong accent, or a cold may prefer pen input. Well before the keyboard is a practiced input device, a young preschooler could use either speech or pen-based drawing to control an educational application.

Multimodal interfaces also provide the adaptability that is needed to accommodate the continuously changing conditions of mobile use. Systems using speech, pen or touch input are suitable for mobile tasks. When these modalities combined users can select among them from moment to moment, depend on environmental conditions [**?**]. For example, the user of an application in a vehicle, may frequently be unable to use pen or keyboard input, although speech is relatively more available. In this respect, a multimodal interface permits the modality choice and switching that is needed during the changing environmental circumstances of actual field and mobile use.

2.3 Multimodal systems architectures

The architecture of multimodal dialogue systems has steadily evolved from monolithic to modular, with well defined communication protocols between modules. A modular architecture is important because it adds flexibility, scalability and robustness to the system. Usually the communication between the modules is delegate by a central controller, although a peer-to-peer model can also be used, for communication between the modules. In practice, a combination of the two ways for communication is used. There is no clear guidelines for the communication protocol between the various modules so both synchronous and asynchronous communications are allowed, although synchronous communication is preferred.

There are also agent-based architectures where the communication between agents or modules can be done via a shared communication space (aka blackboard). Agent-based architectures can handle asynchronous delivery, triggered responses, multi-casting and other concepts from distributed systems. For example, using a multi-agent architecture, speech and gestures can arrive in parallel or asynchronously via individual modality agents. In our study, a combination of peer-to-peer and a central controller is used.

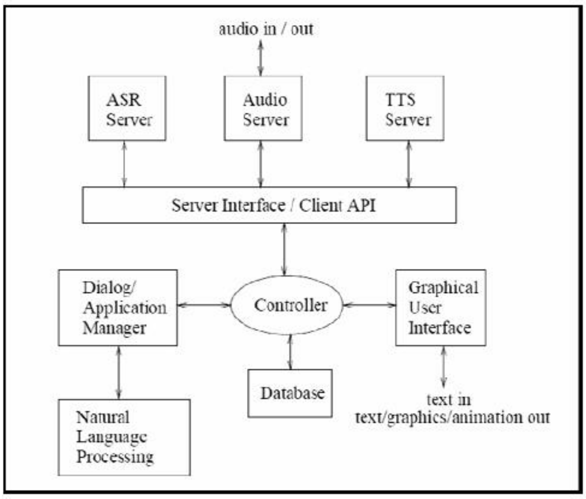

audio in / out
ASR Server
Audio Server
TTS Server
Server Interface / Client API
Dialog/ Application Manager
Controller
Graphical User Interface
Database
Natural Language Processing
text in
text/graphics/animation out

Figure 2.1: A common multimodal system architecture.

In Fig. 2.1 a typical multimodal architecture with emphasis on the speech modality is shown. In muldimodal interfaces the separation of the application and the interface is very important. Application states and communication goals are implemented for multiple modalities, so the separation of the application from the interface can lead to more “clear” implementation. Also the separation of the logic of the application is important because by separating logic, a single application manager can be build for all modalities.

2.4 Design process and principles for multimodal dialogue systems

To exploit the advantages of multimodal systems, the design should conform to some principles. In this way, such systems obtain the maximum functionality. Designing a good application is the first step towards building a successful multimodal system. In the process of building an interactive dialogue system, data collection and analysis are crucial part of multimodal dialogue system development. Data can be collected at various stages of application development and used to improve the system. As a result the development of a multimodal application is an iterative process. Usually “Wizard of Oz” scenarios are used where the unfinished parts of the system are replaced by a software interface that is operated by a human that he/she provide the missing functionality. Through this iterative process, we can lead to a more robust and stable system.

Also the general guidelines that are apply to any interactive application must be follow. First of all, learnability is an important aspect of an interactive interface. The interface should be design in a way that can help users to learn the functionality and familiarize with the system very easy. Moreover flexibility is another important issue. In multimodal systems by flexibility we usually mean the different modalities which user can use in order to interact with the system. More modalities are usually mean more flexible system. Finally robustness is the last general guidline that multimodal systems should be follow. Robustness is the ability of a system, to continue to function despite the existence of faults in its component subsystems or parts. So, a multimodal system should be robust and help users to achieve their goal despite of any faults. For example if an input modality is unavailable for any reason, the system should help user to choose another one.

Beside from the general guidelines like learnability, flexibility and robustness that are apply to any interactive system, there are some design guidelines that are especially important for multimodal dialogue systems as reported below:

- **Consistency:** A multimodal dialogue system should be consistence. Providing consistency between the various input and output modalities is not an easy task especially when speech is involved, so special care should be taken in such cases. One aspect of consistency is symmetric multimodality. By using the same modalities for both input and output, the cognitive load is reduce and the efficiency is improved [33]. By consistency we mean that the same input or output modalities should be used in any state of the system and that the goals should be achieved by using any of the available modalities. For example if a ticket reservation system uses speech and mouse as input, then user should be able to use them at all the necessary steps through the reservation, and should be able to achieve the reservation goal using either speech, mouse or both of them.

- **Efficiency and Synergy:** This principle applies to multimodal interfaces that are use more that one input or output. Synergy of the modalities is very important in multimodal systems. In a synergistic interface the performance should be better than the performance of the individual unimodal interface. Efficiency is one of the advantages of multimodal interfaces and that is because of the synergy between the different input and output modalities.

2.5 Multimodal systems for children

One interesting and relevant field of research in the area of multimodal dialogue systems, is multimodal dialogue systems for children. Computers have become an increasingly popular learning and playing tool, especially with young children, and multimodal approach seems promising in the creation of interfaces for children. That's mainly because the usage of different modalities like speech and gestures can overcoming the limitations of keyboard and mouse as input devices, especially at the early ages, where children can not use them efficiently.

Although children are early adopters of new technologies and interfaces, designing multimodal systems for children is challenging both from the core technology development and the human factors standpoint. Core technology challenges include getting speech recognition technology to work for children users. Interface and human factor challenges have to do with the interaction patterns of children (mix of exploration and exploitation) and the variable capability in using a specific modality (e.g., language, mouse). Overall, variability is one of the greatest challenges when designing multimodal interfaces for children, one size does not fit all.

In the next paragraphs we are going to report these challenges and the relevant work in each field.

2.5.1 Acoustic and linguistic variability

Acoustic and linguistic variability in children speech, is one of the main challenges in the design of multimodal dialogue systems. There are several differences between adult and children's speech. Due to a shorter vocal tract and smaller vocal folds, children have higher fundamental and formant frequencies than those of adults, and greater spectral variability [6, 7, 34]. These differences affect seriously speech recognition performance.

Research has been conducted targeted on adapting adult speech recognition technology towards children, with promising results. In [34] a vocal tract length normalization (VTLN) was applied to an adult recognizer to adapt it to children speech. Also in [35] they normalize the signal, prior feature extraction, using voice transformation techniques. Finally in [1] they used a combination of VTLN and linear parameter transformation to adapt a recognizer. The experiments shows that a rapid improvement of the performance was occurred from ages 7 to 13 years old. At the age of 13 the performance was close to that of adult recognition.

Therefore, using different algorithms, the problem of acoustic and linguistic variabilities can be solved with children of ages 7-13. Although problems are still exist in younger ages.

2.5.2 Integration patterns

In studying integration patterns (voice and gestures) in children, Xiao and colleagues [13] have shown that modality usage was similar between children and adults, although children tend to use both input modes simultaneously rather than sequentially. If children enjoy using simultaneously different input modalities and that make their interaction with the computer more efficient, then we must have it in mind.

Also in interactive interfaces children tend to explore more, than to accomplish only specific tasks. Exploration can intensify children's curiosity and thus hold their span of attention for a longer period of time. Exploration can also help children to develop their problem solving skills, encourage them to learn by trial and error, and improve their knowledge developing skills. But with out the exploitation (task assignment) the game can not used for educational purposes. Thus the combination of exploitation and exploration is crucial in such systems.

2.5.3 HCI for children

One of the main principles we must keep in mind when design interfaces for children, is usability (see the design guidelines in [2]). Although technology is an enabler, usability is the prerequisite for learning and entertainment. Note, however, that for children user requirement vary significantly with age. According to Piaget [23] the cognitive development of a child can be divided into a series of stages with different characteristics. Attention span varies with age, with older children capable of longer periods of attention than younger children [22]. As a result a different mix of sounds, good animation and graphics has to be used for each child to keep him/her engaged. Moreover, most children in the 4-6 age group are at a preliterate level, so the use of text as an output modality must be avoided. Text output is thus substituted or complemented with sounds, graphics and animation.

Recent work in the field of multimodal dialogue systems shows that children enjoy to interact with virtual characters and that the user experience is enhanced if the animated characters possess a specific "personality" and/or social role [6, 29].

Most preschool children cannot use keyboard and mouse efficiently. They can click on specific mouse targets but these targets must be relative large. For these children (especially the 3 and 4 year-olds) speech and touch is the natural modality choice. Although speech is a good choice for the specific target group, the age-depended acoustic and linguistic variability in children's speech [6, 7] makes automatic speech recognition for children more difficult.

2.5.4 Multimodal Applications for children

The increase in usage of interactive technology by children has not gone unnoticed. More than ever before, technology manufacturers and service providers are turning their attention to children as a growing market segment. Even more important, societies are becoming concerned to ensure that appropriate products and services, namely those that can support development and enhance well-being, are being made available for children.

Also the scientific community although higher variability and different interaction patterns create additional challenges, there has done notable efforts in the designing, implementing and testing prototype multimodal systems for children, such as word games for preschoolers [24], aids for reading [25] and pronunciation tutoring [26]. Recently a number of systems with advanced spoken dialogue interfaces, multimodal interaction capabilities and/or embodied conversational characters have been implemented.

In [29] a fairy-tale world was implemented in where users interact with lifelike conversational characters using speech and gestures. What primarily distinguishes the NICE fairy-tale system from other systems was the attempt to move away from strictly task-oriented dialogue and that the dialogue was take place within the context of an interactive computer game.

Furthermore in [43, 44], interactive storytelling systems where the user unfold and affect the course of the story by acting on physical objects on screen, where implemented. In the Agent CHIMP project [15] a prototype consists of two applications (communication agent and spelling bee) was implemented in order to providing design guidelines for building multimodal-input multimedia-output applications for children. The system involves the integration of multiple input and output modalities such as voice, audio, keyboard, mouse, graphics and animations.

Moreover in [27] the initial work towards development of a children's speech recognition system for use within an interactive reading and comprehension training system was presented. The speech technology was used in order to enable students to provide spoken answers to questions about stories, to perform tracking about the words that are pronounced correctly and to provide pronunciation verification capabilities.

Finally in [28], the StoryMat system that engage in story-listening rather than storytelling, is introduce. StoryMat supports and listens to childrens voices in their own storytelling play and offers a child-driven, story-listening space by recording and recalling childrens narrating voices, and the movements they make with their stuffed animals on a colorful story-evoking quilt.

However, almost all of these systems have focused in the age group 6-15. A significant advantage when working with the 6-15 age group is that experimental conditions can be more easily controlled, subjects are collaborating and the subjective evaluation results are easier to interpret. However, for the 4-6 age group speech technology can be very relevant, especially since children are not very adept at using traditional human-computer interfaces, i.e., keyboard and mouse.

Chapter 3

Towards adapting multimodal interfaces for preschoolers

3.1 Introduction

Interactive technologies such as video games have become significant part of children's play culture. Video games present virtual worlds in which children can control and interact with fantasy figures and receive feedback responses tailored to their interactions. At ages 4-6, learning and playing are intertwined activities. Thus the main goal of a successful game for preschooler is to provide fun, excitement and engagement.

In this chapter we are going to investigate some aspects of adapting multimodal dialogue interfaces, in our case some preschool games, in order to increase entertainment. One of the main goals of this work, is to identify some of the parameters that we can adapt in such systems so that we can increase user satisfaction and entertainment. First we discuss what exactly are adaptive interfaces and then we review some of the existing literature which we use in this work, toward identifying some factors that can be used as indicators toward adapting multimodal dialogue systems for preschoolers.

3.2 Adaptable systems

As applications become more complex, it is more hard to build applications and interfaces that satisfy the needs of all users. 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 [37] adaptation can be defined as follow: "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".

Adaptation can play a very important role in the quality of the educational experience, allowing the learning environments to cater to students with different learning styles, different levels of initial knowledge and different expectations and objectives. Educational games are a highly interactive medium, and reactive to the actions of the user. From a technological perspective, this makes them the ideal medium to support an adaptive learning experience, as the game can both monitor the activity of the user within the game and change its own behavior accordingly.

In order to be able to build adaptation models, we have to identify some of the parameters that they provide fun, cause emotions and increase the entertainment for the users.

3.3 Measuring user experience and fun in games

Several theoretical studies have attempted to identify what is "fun" in a game. According to Malone [3] the essential characteristics of a good computer game can be organized into three categories: fantasy, curiosity and challenge. Alternatively, Lazzaro [9] identified 4 relevant categories (hard fun, easy fun, altered states and the people factor) based on Malone's factors and facial expressions/data obtained from actual games. Another well known study is the theory of flow [10], i.e., strong involvement in a task occurs when the skill of an individual meets the challenge of the task. Finally, in the field of entertainment capture, Yannakakis [11] showed that the player-opponent interaction is a major factor in entertainment. In a later work he use physiological measurements in order to measure entertainment. In this section we are going to discuss these studies and try to figure out how we can use these ideas towards adapting multimodal dialogue systems for preschoolers.

3.3.1 Malone's quality characteristics

Malone at his research into games as tools for learning he organized the essential characteristics of good computer games into three categories namely fantasy, curiosity and difficulty [3, 4]. Based on these categories he propose some guidelines that a video game should be follow in order to be fun. More detail analysis of the three categories is reported next.

- **Fantasy:** Fantasy often make computer games more interesting. To make learning motivating and appealing for learners, one way is to present the material to them either in an imaginary context which is familiar to them or in a fantasy context which is emotionally appealing for the learner. As Malone points out, fantasies in games "derive some of their appeal from the emotional needs they help to satisfy in the people who play them". He separates fantasy into intrinsic fantasy and extrinsic fantasy. Intrinsic fantasy depends on the use of a skill in order to achieve some fantasy goal but not vice versa. In contrast in intrinsic fantasy the skill also depends on the fantasy.
- **Curiosity:** Malone identifies curiosity as "the motivation to learn, independent of any goal-seeking or fantasy fulfillment". He divides curiosity into two variants: sensory curiosity, which is about maintaining interest in the senses, and cognitive curiosity, which is more about the semantic content of information. For example, one picks up a National Geographic because the photo on the cover is intriguing this is sensory curiosity. One picks up a newspaper because of a surprising headline this is cognitive curiosity. Audio and visual effects, particularly, in computer games may enhance sensory curiosity. When learners are surprised or intrigued by paradoxes, or incompleteness, it arouses cognitive curiosity. Also Malone is suggesting that randomness is useful in games because it can provoke curiosity.
- **Challenge:** The challenge element of a game is an obvious essential. If it is too easy, the outcome is likely to be certain, making the game boring. If it is too hard the players are quickly demotivated. Challenge is created by having clear, fixed goals that are relevant for the user. Uncertain outcomes provide challenge by offering variable difficulty levels, hidden information, and randomness. Feedback on performance should be frequent, unambiguous, and supportive. Lastly, the activity should promote feelings of competence for the person involved.

3.3.2 Immersion and flow theory

Another well know study is Flow Theory. Csikszentmihalyi [10] studied what makes experiences enjoyable to people. He discovered that central to all the experiences was a psychological state he called flow. An optimal state of enjoyment where people are completely absorbed in the activity. He also found that this experience was similar for everyone, independent of culture, social class, age or gender.

The theory of flow applies in many activities of our everyday lives from education, music and religion to gaming and even in sports. For example the Formula 1 driver Ayrton Senna during qualifying for the 1988 Monaco Grand Prix explained: "I was already on pole, and I just kept going. Suddenly I was nearly two seconds faster than anybody else, including my team mate with the same car. And suddenly I realized that I was no longer driving the car consciously. I was driving it by a kind of instinct, only I was in a different dimension. It was like I was in a tunnel."

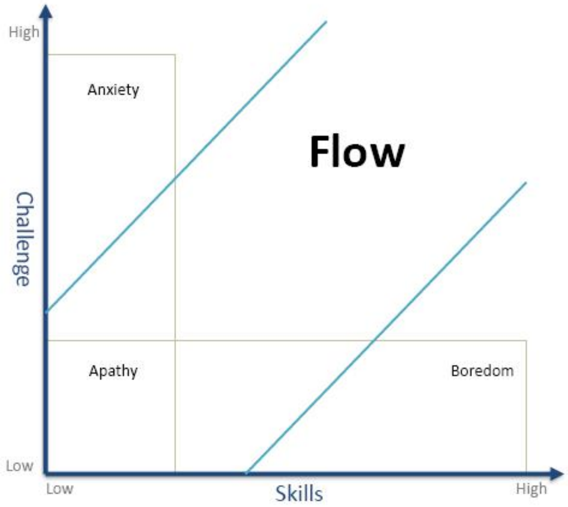


FIGURE 3.1: The two-dimensional model of flow based on Csikszentmihalyi.

According to Csikszentmihalyi the flow experience is characterized by different elements (see [10]). One of them is challenge. Challenge requires skill, and as he says enjoyment arises when the opportunities for action perceived by the individual are equal to his/her capabilities. Thus as shown in Fig. 3.1, flow can be regarded as a state of balance between challenge and skill. And that exactly positive state of mental absorption sounds familiar to a frequent player of computer games.

A similar concept to flow theory is immersion. In the game domain, immersion is mostly used to refer to the degree of involvement or engagement with the game. Based on analysis of children's game-play an immersion model is proposed in [36] where immersion in the game is differentiated in sensory immersion, challenge-based immersion and imaginative immersion, basically the same as Malone's curiosity, challenge and fantasy factors.

3.3.3 Emotion in games and emotion recognition

Emotions play an important role in human to human communication and interaction, allowing people to express themselves beyond the verbal domain. The ability to understand human emotions is desirable for the computer in several applications. It is argued that to truly achieve affective human computer interaction, there is a need for the computer to be able to interact naturally with the user, similar to the way human-human interaction take place, and emotions are part of this.

In applications where computers take on a social role such as an “instructor”, “helper” or even “companion”, it may enhance their functionality to be able to recognize user’s emotional state. For example knowing user’s emotions the computer can become more effective tutor. Also synthetic voice with emotions in the speech would sound more pleasant than a monotonous voice.

In [9] Lazzaro shows that people play games not so much for the game itself as for the experience the game creates(e.g adrenaline rush) or the structure games provide (e.g the company of a friend). She also identified 4 relevant categories that can unlock emotions in games (hard fun, easy fun, altered states and the people factor), based on Malone’s factors and facial expressions data, obtained from actual games. So it seems that emotions are important part of the gaming experience. They play a vital role in decision making, attention, performance, learning and enjoyment. Identifying negative emotions can identify help hot-spots in the interaction.

The vocal aspect of a communication carries various kinds of information. Traditionally as well as in more recent studies, emotions can be recognized using features from prosodic information which includes pitch, duration and energy statistics [38, 39], extracted from user’s speech input. Facial expressions also have significant amount of information, regarding emotions. In [40] Ekman and Friesen developed the Facial Action Coding System (FACS) to code facial expressions where movements on 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, have used these basic expressions in order to recognize emotions.

Also emotion recognition is also possible by combining different modalities [41]. Audio, linguistic, pragmatic and visual information can be combined to obtain a good prediction of the child’s emotional state [30]. More recently there has been interest in emotion recognition and modeling of children’s mood in spoken dialogue and gaming applications [14].

3.3.4 Physiological measurements

In recent studies physiological measurements during gameplay are computed and used in order to build models that can predict entertainment. In [17, 20, 21], the Playware game platform has been used for recording physiological signals of children, like heart rate (HR), blood volume pulse (BVP) and skin conductance (SC), during gameplay. The study shows that several statistical parameters of the recorded signals where correlated with the reported entertainment. In a later study they use the same physiological features in order to evolve neural networks that can predict entertainment.

Physiological signals are also used for emotion recognition [42]. Also in [16] physiological signals where used in order to identify emotions like anxiety, boredom and engagement during a tetris game. Using the collected data they shows that boredom, engagement (flow) and anxiety are good indicators for difficulty adaptation in games.

3.4 Towards adaptable mutimodal systems

Based on these prior works, one of the goals of this thesis is to identify how fantasy, curiosity and challenge affect the entertainment value of multimodal dialogue computer games for preschoolers. In previous work with children (ages 8-10) playing on Playware game platform, Yannakakis et al [12] has shown that fantasy is correlated with entertainment but curiosity and difficulty depends on each child's preferences. However, it is unclear if these results hold for younger children interacting using a multimodal dialogue interface.

Another goal is to investigate how these factors can be adapted to increase the entertainment value of the game, i.e., which are good indicators (or features) of the right fantasy, curiosity and challenge level in a game. For this purpose both interaction patterns and acoustic features of the speech input have been studied.

Our last goal is to identify user's emotions using audio features from the collected data. By knowing user's emotions during the game we can help him/her reach the optimal state of enjoyment, by adapting the game (e.q Malone's quality factors). When the children is negative excited (Anxiety) the difficulty level could be drop down. On the other hand when negative calm (Boredom) is detected the difficulty level could be increase. By adapting the difficulty value and by offering high values of curiosity and fantasy in our games, we are going to achieve high entertainment.

Finally objective metrics from the game such as interaction times and correct/wrong answers could also be used in order to adapt difficulty. Long interaction times usually means that the child is boring or that the difficulty level is too high. The same thing happens when lot of wrong answers are given.

Chapter 4

Platform Design and Implementation

4.1 Introduction

The main advantage of building a web-based platform for multimodal game-development is that it can be used for (remote) data collection and analysis of educational software and games. The collected data can be used to train language and acoustic models for automatic speech recognition (ASR) and for analysis of user interaction patterns to improve or adapt the user interface. Educational software and games are also used extensively by linguists and psychologists, e.g., to diagnose and solve language development problems. In this work we use the platform to study the children-computer interaction for preschool children and to see how we can adapt Malone's quality characteristics in order to improve the user experience.

4.2 Platform Architecture

The system follows a modular architecture, the full functionality of the system being the result of the collaboration between the modules. The architecture of the system is shown in Fig. 4.1.

Since this is a web-based platform it is (by nature) distributed. The *Application Manager* is responsible for the synchronization and cooperation of the modules. It consists of two parts that follow the client/server architecture, i.e., the client/browser side and the server side. The two parts communicate through a two way socket connection. The *Speech Module* is responsible for capturing and streaming the audio, as well as, performing the

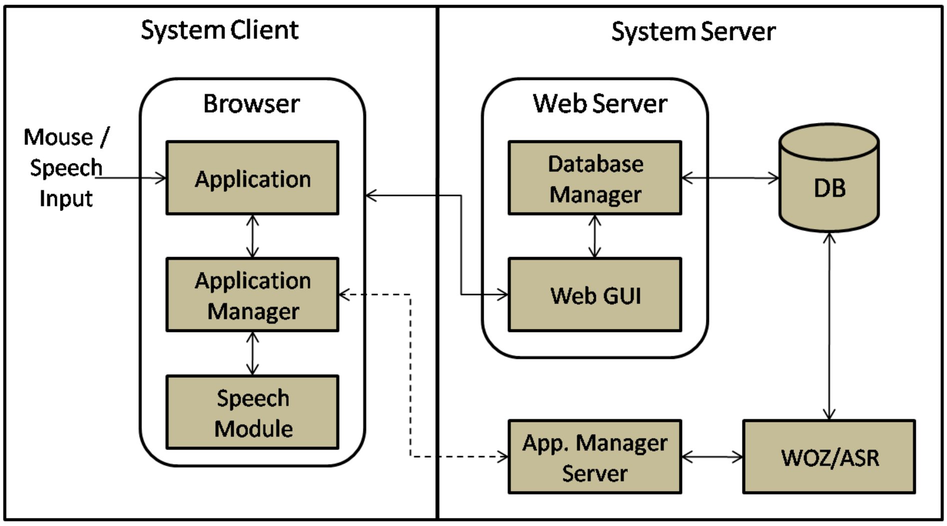


FIGURE 4.1: The modular architecture of the platform.

voice activity detection (VAD) to determine if the user is speaking. When voice activity is detected the Speech Module starts the audio capture. At the same time, the streaming of audio data to the client part of Application Manager begins. Finally, the multimodal *Application* module may contain any interactive application implemented by the system designer. In our case, we have integrated five preschool games into a single application, as explained in Chapter 4. All games were implemented in Flash [49] in order to provide an easy and platform-independent way to manipulate sounds, animations and graphics.

On the server side of the system, the *Application Manager* (server part) is responsible for executing the speech requests that are being received from the client side of the platform. It receives and then streams the audio data to the *ASR/WoZ* module for automatic recognition or manual transcription (by the wizard). The result of the recognition is send back to the client part of the Application Manager. The module also receives and stores the necessary log files for further processing. On the server side of the system, we have also implemented the *Web Interface Module* and the *Database Communication Module*. The web interface was designed using Java Server Pages (JSP) technology. Using the web interface users can register and login to the platform. Functionality such as profile management and preferences configuration (e.g., microphone configuration) is also provided. The Database Communication Module is responsible for all the necessary database queries. Both the Web Interface and Database Communication modules are parts of the Apache Tomcat Web Server.

The *ASR/WOZ* module has been implemented on the server side of the system. In this study, the ASR module has been replaced by a Wizard of Oz (WoZ) module which is operated by a human transcriber. The WoZ module is actually a graphic user interface (GUI) that plays the audio stream received by the Application Manager and allows the wizard to supply the appropriate transcription via a GUI interface. Both the audio and transcription files are stored in the database.

4.3 Modules Implementation

4.3.1 Speech Module

The *Speech Module* is responsible for capturing and streaming the audio, as well as, performing the voice activity detection (VAD) to determine if the user is speaking. The module is implemented as a java class using the Java Sound API. Three main functions have been implemented in the module. *CaptureAudio* function is responsible for capturing and streaming the audio to the *Application Manager*(server part) using the Java Sound API. The captured audio is in PCM 16000Hz 16bit stereo format. *VoiceDetection* function is responsible for Voice Activity detection (VAD). Voice activity detection (also known as speech activity detection or, more simply, speech detection) is a technique used in speech processing wherein the presence or absence of human speech is detected in regions of audio (which may also contain music, noise, or other sound).

VAD is an important enabling 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. We have been implemented a very simple algorithm using the energy of the captured frames by computing the Root Mean Square (RMS) for each frame. When RMS is above a specific threshold (which can be changed through microphone calibration), the *CaptureAudio* function is called. When RMS is below that threshold for a short period of time, the *StopCapture* function is called, which is the function that stops the recording procedure.

4.3.2 Application Manager

The most important part of the system, is *Application Manager* module. As we already said, *Application Manager* is distributed and follows the client/server architecture. The client part of *Application Manager* was implemented as a java applet and it is running on the client part of the system (the browser), while the server part of the module was implemented as a java multi threading server.

When a VAD event occurred, the client part of the module sends a “voice request” message to the server part of *Application Manager* so that the streaming connection could be established. Then the module waits for the appropriate response from the server part of the *Application Manager*. If the response is positive, the *Application Manager* notifies the speech module, so that the recording and the streaming of the audio data could be start.

When the user stops talking and a stop speech VAD event is occurred, the client side of the *Application Manager* notifies the server side of the module, that the streaming is over and then waits for the audio recognition result. After receives the answer from the server part of the module, the result is send to the multimodal application where the appropriate action is taking place, depends on the answer.

The server part of the *Application Manager* is responsible for receiving the audio data and send them to the Speech Recognition Module, which in our case is the Wizard of Oz module. When it has the result of the recognition, the appropriate message is send to the client part of the module. Also the server part of *Application Manager* is responsible for receiving and storing the log files of the application for further process. The server part of the module is implemented as a java multi threading server using the network API of the Java. The communication between the client side and the server site of the module is occurred through a two way TCP/IP socket. The process of a speech request through the system is shown in 4.2.

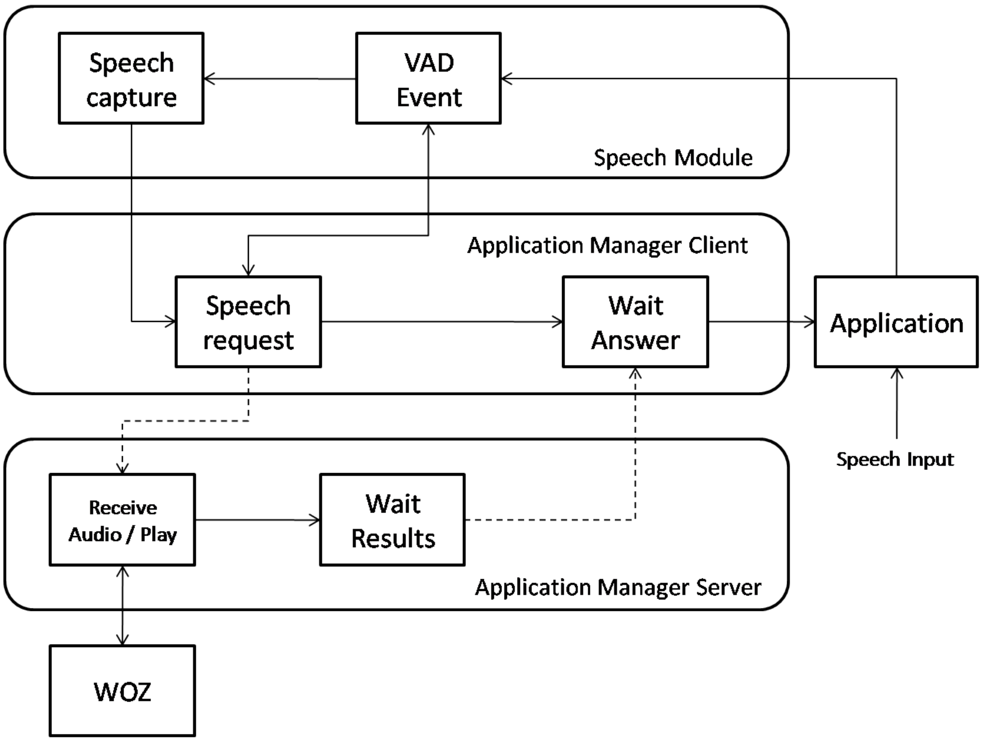


FIGURE 4.2: Service of a voice request process.

In order to achieve the cooperation between the client and the server part of the *Application Manager*, a communication protocol has been implemented. We have built a fairly simple protocol for the communication between the two part of the *Application*

Manager based on simple text messaging. Table 4.1 shows the protocol that we have use for all the communications between the client and the server part of the *Application Manager*.

App. Manager Client	**App. Manager Server**	**Action**
VOICEREQ#	VOICEREADY#	Audio streaming
—	AUDIODONE#	Audio streaming
ANSWEROK#	ANSWER#result	Speech Recognition result
WRITELOG#log	WRITELOGOK#	Saving Log File

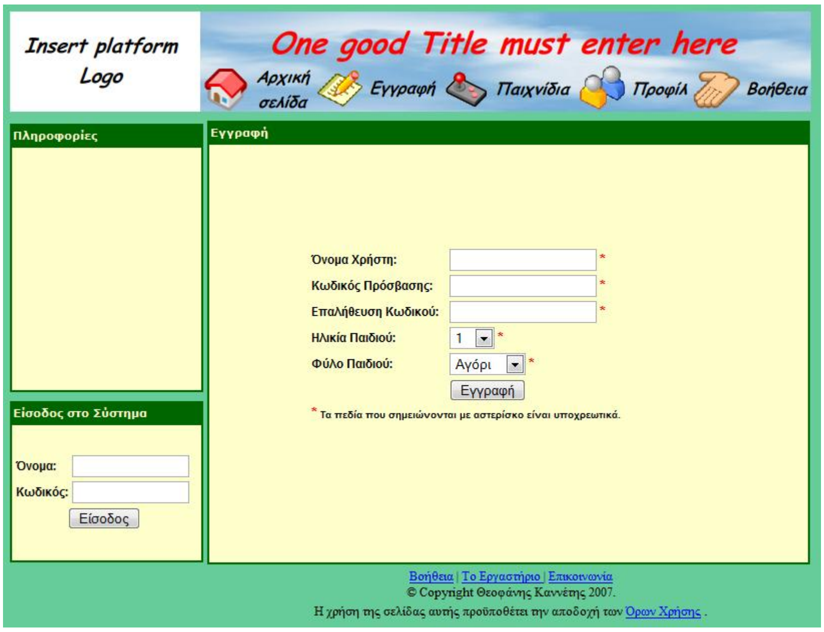
Table 4.1: The communication protocol between client and server parts of the Application Manager.

When a connection for audio streaming is needed, the client part of the *Application Manager* sends the "VOICEREQ#" string to the server side of the module. Then it waits for the "VOICEREADY#" answer in order to start streaming the audio data to the server side of the module. When the streaming is over the server side sends the message "AUDIODONE#" so that client part gets ready to receive the audio recognition result. When the server side of *Application Manager* has the result of audio recognition the "ANSWER#result" (where result is the audio recognition result) message is send back. Then if receives from the client "ANSWEROK" it terminates the connection, else it sends again the result. Finally for "Log writing" request, the client part must send "WRITELOG#log" (where log is the log file string). When the log file is successfully written in the database, the message "WRITELOGOK" is send back.

4.3.3 Web interface and Database

As we decide to build an on-line system, the appropriate web user interface should be implemented. Using the web interface, users could register and log-in to the platform so that they could use it. Also profile management and microphone calibration is provided through the web interface. The web interface has been implemented using the Java Server Pages (JSP) technology.

In order to store information regarding to users a database communication module has been implemented. The module is responsible for all the database transactions including register and log-in. The module is actually a collection of JavaBeans that each one of them is responsible for a specific job. JavaBeans are classes written in Java conforming to a particular convention. They are used to encapsulate many objects into a single object (the bean), so that they can be passed around as a single bean object instead of as multiple individual objects. The JavaBeans classes are called through the web interface so that the communication between the database and the web interface could be established. Both JSP pages and JavaBeans are hosted on Apache Web Server. Finally in Fig 4.3 an example of the implemented beta web interface is shown.



Όνομα Χρήστη: *
Κωδικός Πρόσβασης: *
Επαλήθευση Κωδικού: *
Ηλικία Παιδιού: 1 *
Φύλο Παιδιού: Αγόρι *

* Τα πεδία που σημειώνονται με αστερίσκο είναι υποχρεωτικά.

Όνομα:
Κωδικός:

Βοήθεια | Το Εργαστήριο | Επικοινωνία
© Copyright Θεοφάνης Καννέτης 2007.
Η χρήση της σελίδας αυτής προϋποθέτει την αποδοχή των Ορων Χρήσης .

FIGURE 4.3: Web Interface Example

4.3.4 ASR and Wizard of Oz

The Wizard of Oz (WoZ) technique is an experimental evaluation mechanism. It allows the observation of a user operating an apparently fully functioning system whose missing services are supplemented by a hidden wizard. The user is not aware of the presence of the wizard and is led to believe that the computer system is fully operational. The wizard observes the user through a dedicated computer system connected to the observed system over a network. When the user invokes a function that is not available in the observed system, the wizard simulates the effect of the function. Through the observation of users behavior, designers can identify users needs when accomplishing a particular set of relevant tasks and evaluate the particular interface used to accomplish the tasks.

In our system, the WoZ component was implemented in JAVA and replaced the ASR component which was described above in this subsection. Each game uses the VAD component to notify the presence of speech to the *Application Manager* which streams the audio to the Wizard of Oz interface. It then replays the sound to the wizard who chooses the appropriate answer and sends it back to the game.For the synchronization of the various components (game, VAD, Application Manager, WoZ) to be achieved, the Wizard of Oz application has a text flag (wait - give answer) which notifies the wizard when to give the answer. Also note that there are different tabs in the WoZ GUI, one

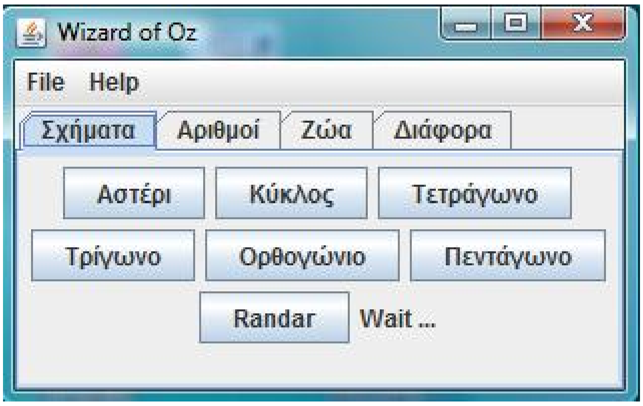


Figure 4.4: The Wizard GUI

for each task (see Fig. 4.4). The tabs are selected automatically by the interface based on the task children choose to play.

4.3.5 Multimodal Application

The last module of the system is the multimodal application module which may contain any interactive application implemented by the system designer. These applications could be implemented in any web-based programming language. The communication between the application and the client part of the *Application Manager* was implemented with Javascript so that the communication between the application and the *Application Manager* be indepented of the implemented language (almost all of the web-based languages support javascrip calls). In our case, we have integrated five preschool games into a single application using Flash [49], as explained in the next chapter of this thesis.

Chapter 5

Application Design and Implementation

5.1 Introduction

As we have already discussed in Chapter 2, there are many challenges regarding the design of a multimodal dialogue system for children. Keeping all these challenges in mind, in this chapter we are going to discuss how we have been implemented five different multimodal games based on popular preschool activities. We are also going to see how we have implement fantasy, curiosity and difficulty factors and how we have combine all the five tasks in a single application in order to study how these three Malone factors affect entertainment.

5.2 Application Functionality

Five different popular preschooler tasks were selected for implementation. The selected tasks were: animal recognition (ages 3-4), shape recognition (ages 4-5), quantity comparison (ages 3-4), number recognition (ages 5-6) and addition (ages 5-6). For each game an embodied agent guides the child through the task. Both mouse and speech are enabled as input modalities. Animation, sounds, graphics, prerecorded prompts and synthesized text-to-speech prompts (where necessary) were used as output. The list of tasks is described next:

- The animal recognition task is taking place in a farm. First the voice of an animal is heard and then the farmer asks the child to select the appropriate animal in order to guide it into the farm. There are (up to) nine different farm animals in the game (see Fig. 5.1). The specific task has been implemented by Spyros Meliopoulos as a part of his diploma [45].

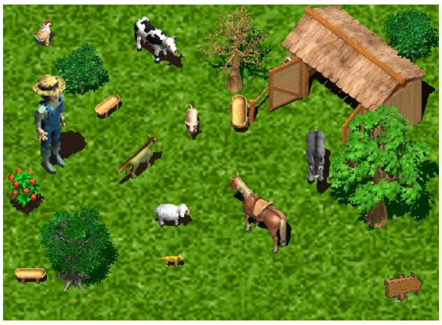


FIGURE 5.1: Animal recognition task.

- The number recognition task takes place at the beach where an animated character (squirrel) asks the child to identify which number (1-9) is shown on the screen (see Fig. 5.2).

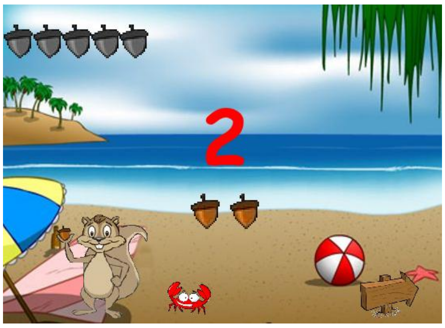


FIGURE 5.2: Number recognition task.

- The shape recognition game takes place in a theater. Each time one of the shapes (star, circle, square, rectangle, triangle, pentagon) appears on stage and the child must identify it. The animated character (teacher) provides help and guides the child through the task (see Fig. 5.3).

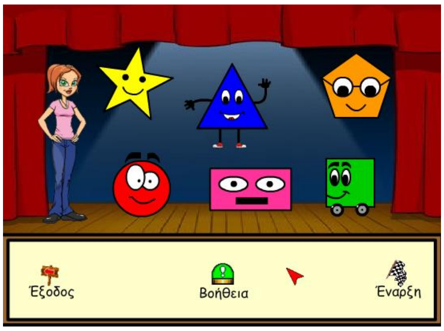


FIGURE 5.3: Shape recognition task

- For the comparison task, an animated character (rabbit) puts some items inside and some outside of a basket. The rabbit asks then the children to determine whether the items inside the basket are more (or less) than those outside (see Fig. 5.4).

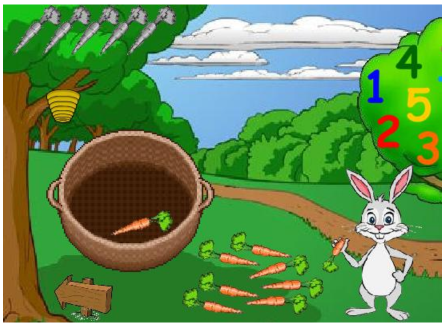


Figure 5.4: Quantity comparison task

- For the addition task, the child must help an animated character (bear) to collect some honey from the beehives. A simple addition task appears on the screen (sum up to 9) for the child to preform. For each correct answer a bee fills a honey jar with honey (see Fig. 5.5).

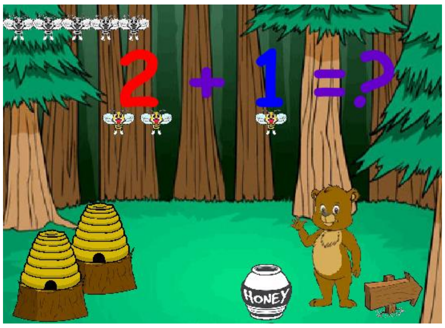


Figure 5.5: Addition task

5.3 Implementing Fantasy, Curiosity and Difficulty

In this section, we describe how variable levels of fantasy, curiosity and challenge have been implemented for each of the five tasks. According to Malone [3], fantasy can be separated into intrinsic fantasy and extrinsic fantasy. Intrinsic fantasy depends on the use of a skill in order to achieve some fantasy goal but not vice versa. In contrast in intrinsic fantasy the skill also depends on the fantasy. The second factor Malone recognizes in good games is curiosity. He defines curiosity is the motivation to learn independently of any goal seeking or fantasy-fulfillment. Specifically "games can evoke players curiosity by providing environments that have an optimal level of informational complexity". That means that those environments "should be neither too complicated nor too simple and should be novel and surprising but not completely incomprehensible". Finally in order for a computer game to be challenging according to Malone a goal must be provided whose attainment is uncertain.

5.3.1 Difficulty Implementation

The challenge element is crucial in a game. If a game is too easy, the outcome is likely to be certain, making the game predictable and boring. If it's too hard players are quickly demotivated. This is well-understood by the gaming industry and thus most computer games are playable at different levels. We have implemented three different levels of difficulty for each of the five tasks.

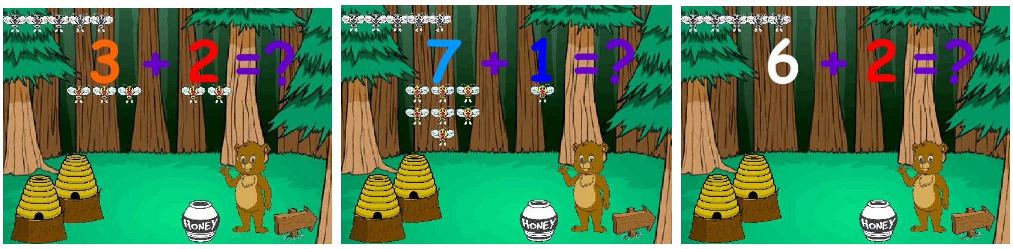


FIGURE 5.6: Addition task in different difficulty levels: sum from 1-5 with items help (a), sum from 5-9 with items help (b) and sum from 1-9 without help items(c)

For example for the addition task the system asks additions with sum from one to five at difficulty level 0, from five to nine at level 1, and from one to nine but without the helping items underneath each number at level 2 (see Fig. 5.6). The implementation of the three difficulty levels is shown in Table 5.1 for each task.

Value	**Difficulty**				
	Farm	**More/Less**	**Numbers**	**Additions**	**Shapes**
0	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	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	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 5.1: The three levels of difficulty as implemented in our application. Implementation of difficulty is task dependent.

5.3.2 Fantasy Implementation

Fantasy often makes computer games more interesting. Almost every game requires the player to take on a new role (fantasy identity), a process that is apparently very fulfilling. In our work, we use the intrinsic type of fantasy as defined by Malone [3], i.e., the use of a skill is required to achieve some fantasy goals. We have been implemented this type of fantasy by taking the existing task oriented games and adding to them a fantasy goal, namely, helping an alien that crashed to earth return to his planet (see Fig. 5.8). In order to help the alien, the child must collect 2 items that alien can use them in order to repair his spaceship.

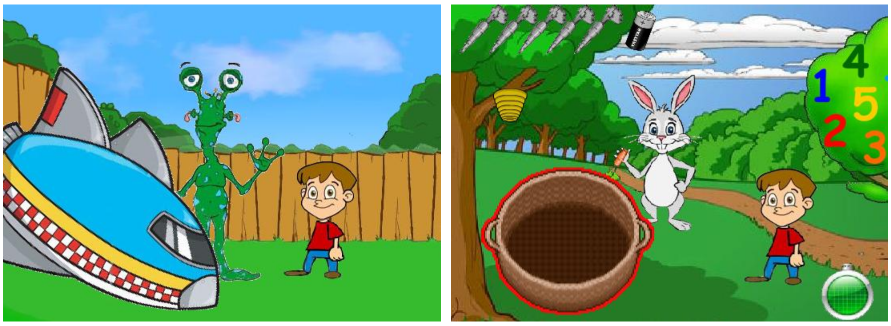


Figure 5.7: Screen-shots of the implemented fantasy goal.

In order to implement different fantasy levels, short animations were also added to each task (triggered fantasy elements). For example, for the numbers recognition task (see Fig. 5.8), the crab starts walking around making noises when the child clicks on the crab or says "crab". Thus, in our implementation the three different fantasy levels are: without story, with story but without fantasy triggers, and with story and fantasy triggers. Table 5.2 shows the implemented fantasy elements.

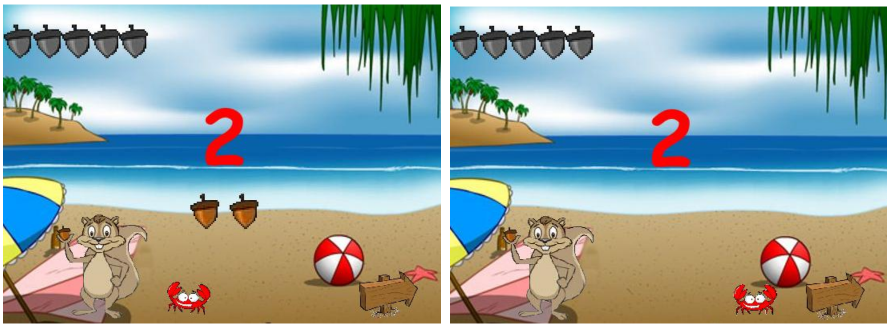


Figure 5.8: Fantasy trigger for the numbers recognition task.

5.3.3 Curiosity Implementation

Curiosity is the less obvious factor. Malone identifies two main features of curiosity: sensory curiosity, or the attraction to the environment (sounds, movement, images etc)

Value	**Fantasy**				
	Farm	**More/Less**	**Numbers**	**Additions**	**Shapes**
0	No story or fantasy triggers	No story or fantasy triggers	No story or fantasy triggers	No story or fantasy triggers	No story or fantasy triggers
1	Story but no fantasy triggers	Story but no fantasy triggers	Story but no fantasy triggers	Story but no fantasy triggers	Story but no fantasy triggers
2	Story and fantasy triggers (different item animations)	Story and fantasy triggers (bee buzzing around)	Story and fantasy triggers (crab making noises)	Story and fantasy triggers (bee buzzing around)	Story and fantasy triggers (train assembly)

Table 5.2: The three levels of fantasy as implemented in our application.

and cognitive curiosity or a desire to bring better "form" to one's knowledge structures. Some of the ways to achieve this according to Malone are: rewards, information representation system and surprising feedback. We have implemented several of these elements in our application.

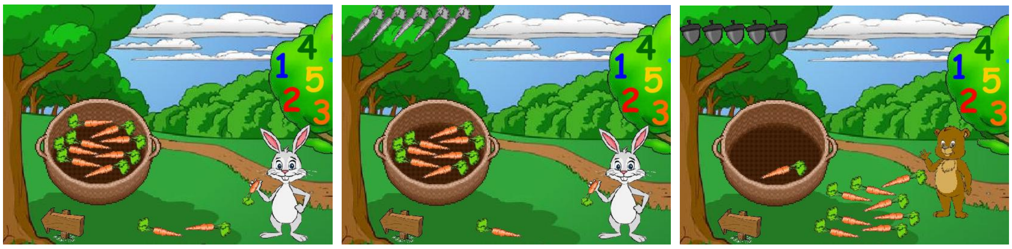


Figure 5.9: Implementation of the three levels of curiosity: No answers bar and randomness (a), with answers bar but no randomness (b), with answers bar and randomness (c).

A bar representing (progress with) correct answers has been added at the top of the screen for each task. Furthermore we have implemented the incentive of the reward. When a child wins a game task an object passes to his possession. According to Malone, the "easy" way to engage users' curiosity and have surprising feedback is using randomness. For example, the animated characters now randomly appear in each task depending on curiosity level. Also the system proposes random tasks to the children based on curiosity level and child's age. Furthermore some graphics that appear on stage (e.g., answer bar items) are now selected randomly. In Fig. 5.9 example screen-shots for the three levels of curiosity implementation are shown. Finally Table 5.3 summarizes how varying degrees of curiosity have been implemented in our application.

Value	Curiosity
0	No answers bar and randomness
1	Answers bar but no randomness
2	Answers bar and randomness

Table 5.3: The three levels of curiosity as implemented in our application.

5.4 Application Flow Diagram

5.4.1 Task Flow Diagram

All five tasks follow the same flow diagram shown in Fig. 5.10. After a small introduction children can choose to proceed to the main task or leave the game. If a child chooses to play a specific game, the system generates a question based on difficulty level and the animated agent asks the child to answer it. Then the system waits for the answer. At that stage children can provide an answer or trigger same fantasy elements.

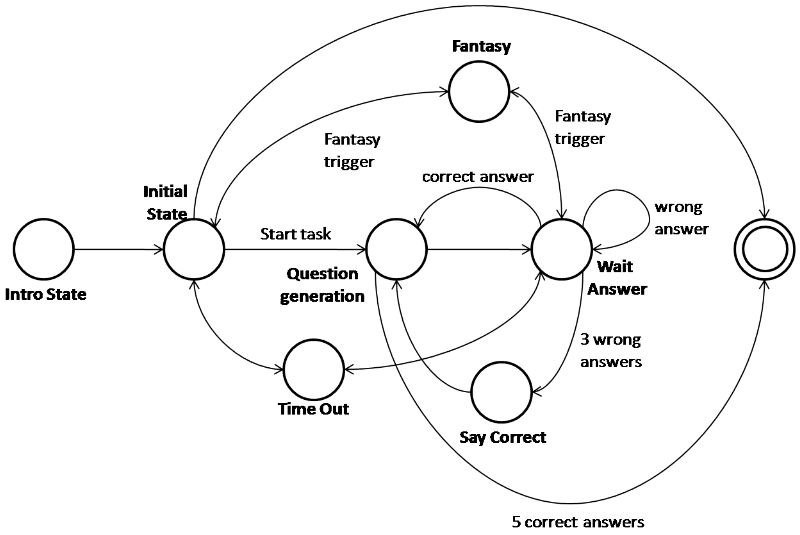


Figure 5.10: Game flow diagram.

If the child gives the correct answer then the system generates another question. When a wrong answer is given, the agent repeats the same question. After three wrong answers, the agent provides the correct answer and the system generates the next question. Each game concludes after the child gives five correct answers. The child can leave the task or trigger some fantasy elements anytime. We have also implemented a time-out; when a child delays an answer or takes no action, the agent repeats the question. Note that based on the curiosity value, the game selects the agent with whom the child will interact and displays (or not) the answer bar.

5.4.2 Interface Flow Diagram

In order to merge the existing task oriented games with the fantasy and curiosity elements and put them in a single application, a simple interface was designed as shown in Fig. 5.11.

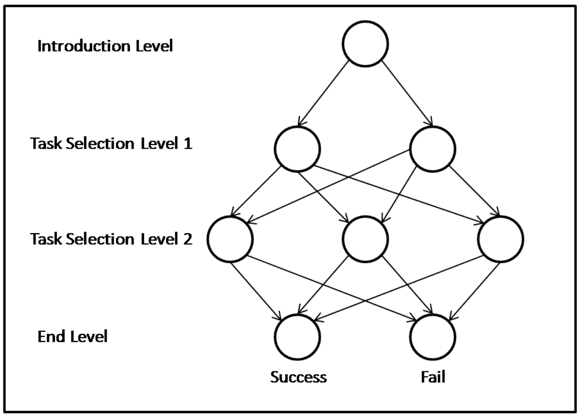


Figure 5.11: Game Interface flow Diagram.

The new interface application consists of four different levels. The first level is the introduction of the fantasy story, where children can also learn how to use the two different input modalities. (see Fig. 5.12(a)). The next two levels, they represent the two different tasks that child has to played, so that he/she collect the items that needed in order to finish the story. In each level of the interface the children can choose between the implemented task oriented games, depend on age and curiosity level. At curiosity levels where randomness exist, the proposed tasks are random, while in the curiosity levels where there is no randomness, the task propose is always the same.

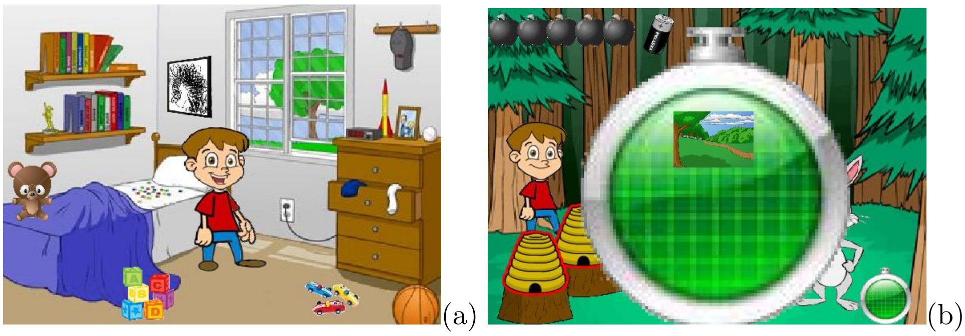


Figure 5.12: Screen-shots of the interface: tutorial scene (a) and Task selection using radar (b).

The task selection can be achieved through the interface using a radar that the alien gives to the child (see Fig. 5.12(b)). At the second level, children can choose between 2 different tasks, while at level 3 between 3 different tasks. Note that at level 3 the previous selected task can not be selected. Finally the last level has the “win” and “loose” states, where a short animation is played depend on the items that child manage to collect in the tasks that he/she have been selected.

Chapter 6

Evaluation Methodology

6.1 Introduction

An important step in the iterative development circle of an interactive system is the evaluation of the interface design and implementation [31]. Evaluation helps to ensure that the system functionality fulfills the requirements of the various tasks supported. It also allows the system designer to measure the effectiveness of the system, by measuring user performance. Finally, evaluation helps to ensure that the certain usability principles and guidelines have been followed, resulting high levels of user satisfaction.

Usability is a key quality of interactive systems. A usable system is one that can be used effectively, efficiently and enjoyably. By effectively we mean that a user can achieve goals that the system was intended to support. By efficiently we mean that these goals can be achieved with acceptable levels of resource, mental energy, physical effort or time. Finally by enjoyable we mean that the usage of the system delivers levels of enjoyment appropriate of the context of the use. In our case entertainment is our primary quality key, because we are talking about games. Thus the main goal of a successful game for preschooler is to provide fun, excitement and engagement.

6.2 Multimodal Systems Evaluation

The evaluation of multimodal dialog systems is a complex task and different metrics (objective and/or subjective) are typically used to evaluate different aspects of such systems [8]. Objective metrics such as speed, number of errors, task completion, are computed for the various system configurations and are statistically analyzed to determine the best system. Alternatively, subjective metrics can be used in order to elicit direct user feedback using either interviews or questionnaires. Subjective metrics are simple to carry out and analyze, and can provide useful information if are well designed. However the elicited information is subjective and may be less accurate than the objective evaluation methods.

6.3 Our Approach

Since we are mainly interested in investigating how fantasy, curiosity and challenge affect children satisfaction, the relation between the three factors and objective/subjective criteria is key. The following objective criteria are reported: average response time, task completion and input modality usage. Response time is defined as the time that elapses from the end of a system prompt until the child completes his/her answer (stop talking or clicks the mouse on a valid target). We separated the response time to inactivity time (end of a system prompt until first voice or mouse activity detection) and interaction time (response minus inactivity time)[18]. In addition to these objective metrics, we also report the most enjoyable system setup that each child selected for each session. At the end of each session, children participated also in an exit interview; a summary of these subjective opinions is also presented in Section 7.3.

6.4 Experimental Setup

The evaluation of the system took place in a noisy preschool environment using a WoZ experimental setup. Nine native Greek speakers, ages four to six, participated in the study by playing different versions of the application (at different values of fantasy, curiosity and difficulty). Five of them were boys and four of them girls. All of the subjects believed that they were interacting with an automated system, i.e., they had no knowledge of the existence of a wizard.

In order to familiarize themselves with the system, each child played the tasks appropriate for his/her age once, using both mouse and voice. After finishing the demo session, each child was asked to play 3 different versions (sessions) of the game and choose the one that he/she enjoyed most. To avoid overloading/tiring the child each session was played at different visits to the preschool. At each session only the value of one factor (fantasy, curiosity or difficulty) was modified, while the values of the two other factors remained constant at level 1. The order that each factor and factor level was presented to the child was randomized. Thus, at each session the child played three versions of the application (one for each of the three levels of the relevant factor). Each user played at least once all the tasks that are suitable for his/her age as discussed in Chapter 4. Overall, each child played 3 sessions, corresponding to 9 different application setups (for each application run each child played 3-5 different tasks). Note that an adult was present during the data collection (sitting next to the child) to help and guide the child through the application, as needed. After the completion of each session the children were asked to evaluate the system by participating in a subjective assessment.

Chapter 7

Evaluation Results and Analysis

7.1 Introduction

In this chapter we report the evaluation results and their implications for multimodal dialogue systems designed for preschoolers. First we present the objective metrics results: interaction time, modality selection and task completion, as well as the correlation between the various objective metrics. Then the subjective evaluation results are reported, including user's subjective system choice and the results from the exit interview. Finally we present the results of emotion classification in two categories (negative, non-negative) using audio features from the collected data.

7.2 Objective Metrics

In Table 7.1, a summary of the objective evaluation metrics is shown as a function of age and gender.

	Age			**Gender**	
	4	**5**	**6**	**M**	**F**
Av.Sess.Time(min)	4.53	3.40	3.72	3.93	3.65
Av.Rsp.Time(sec)	4.78	3.78	3.64	3.78	4.38
Av.Inact.Time(sec)	0.99	1.12	1.04	0.89	1.29
Av.Inter.Time(sec)	3.79	2.66	2.60	2.89	3.09
Mouse usage(%)	16.14	18.90	23.55	20.88	19.79
Speech usage(%)	83.85	81.09	76.44	79.11	80.20
Wrong Answers(%)	9.83	5.12	2.75	4.75	6.64
Task comp.(%)	89.65	97.14	97.37	90.32	97.62

Table 7.1: Objective evaluation results.

Specifically, for each age and gender, average response time (sec), inactivity and interaction time (sec), average session time (sec), percent number of turns of speech and mouse input, percent of task completion and wrong answers are shown. Note that the average response time is computed using only the games that are suitable for each age group (see discussion in Chapter 5).

7.2.1 Inactivity and interaction time

In Fig 7.1 the average response time (separated in inactivity and response time) for each user is shown. Users 2,3 and 4 are four years old, users 1, 5 and 6 are five years old, and users 7,8 and 9 are 6 years old.

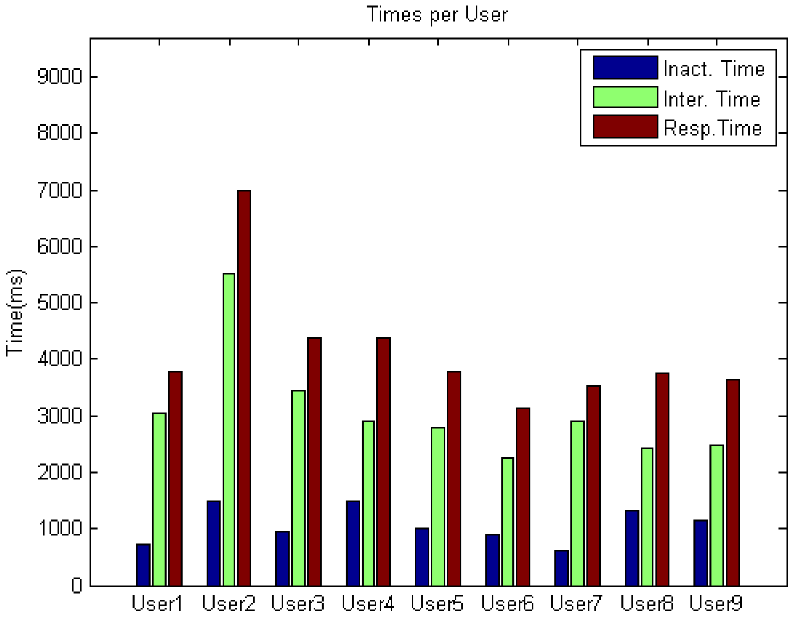


FIGURE 7.1: Inactivity, interaction and response times per user.

As shown in Table 7.1, four year-olds have higher average response time (by 1 sec) than five and six year-olds (no significant difference in response time between ages five and six). There is no significant difference in inactivity time for all three ages, thus the 1 sec difference is attributed to interaction time (Fig. 7.2 (a)). The average response time for girls is slightly higher than that of boys (both inactivity and interaction times are higher) (Fig. 7.2 (b)).

In Fig. 7.3 the average response time per task and age are shown. As expected average response time of four year-olds children is higher than five and six year-olds. Also five year-olds and six year-olds have similar response time in most tasks. The trend is consistent across tasks, with the exception of the comparison task ("More/Less"). The

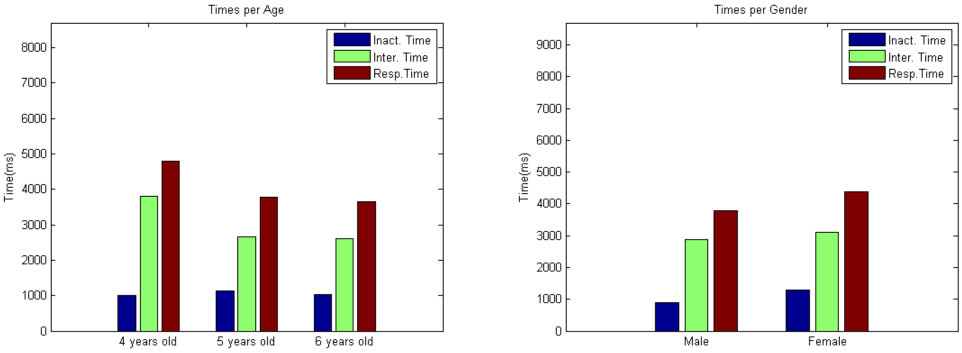


Figure 7.2: The average inactivity, interaction and response times per age (a) and gender (b).

very high response time for the "Numbers" task is due to the fact that four year-old have a hard time performing this task.

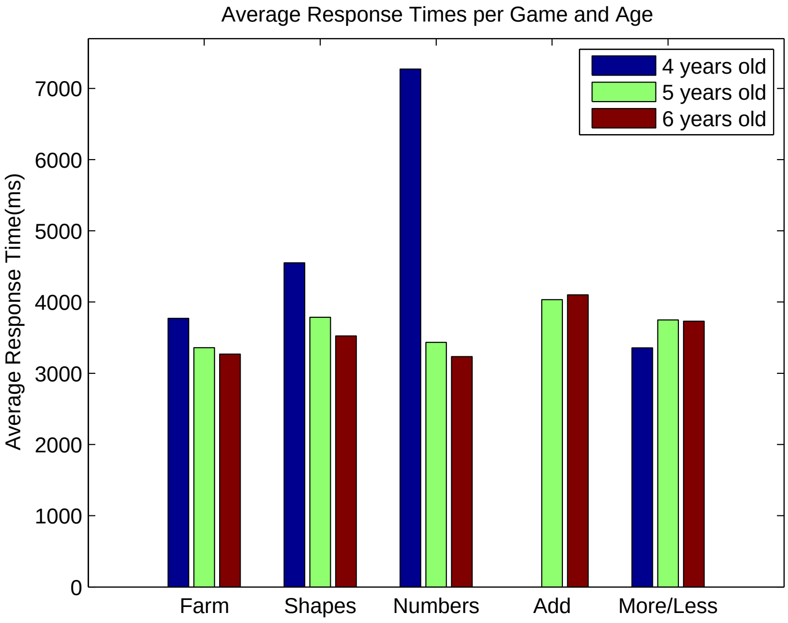


Figure 7.3: The average response time per task and age.

In the Fig. 7.4 the average response time separated in inactivity and interaction time, per task and age is shown. The most interesting trend observed in the "Numbers" task, where the inactivity and interaction time decreased as the age increased. As we said before, that is because four year-olds have a hard time performing the "Numbers" task. Also another interesting trend can be observed at the "Farm" task. Although inactivity time for the younger children is lower than the older children, the trend for the interaction time is exactly the opposite. One possible explanation for this are the fantasy elements of the game. The interaction of four year-old children contains fantasy elements that do not appear in the interaction of older children. A common interaction example of four year-old children is “horse go home”. Older children use shorter phrases (“It’s horse”). For all the other tasks the trends are the same for all ages.

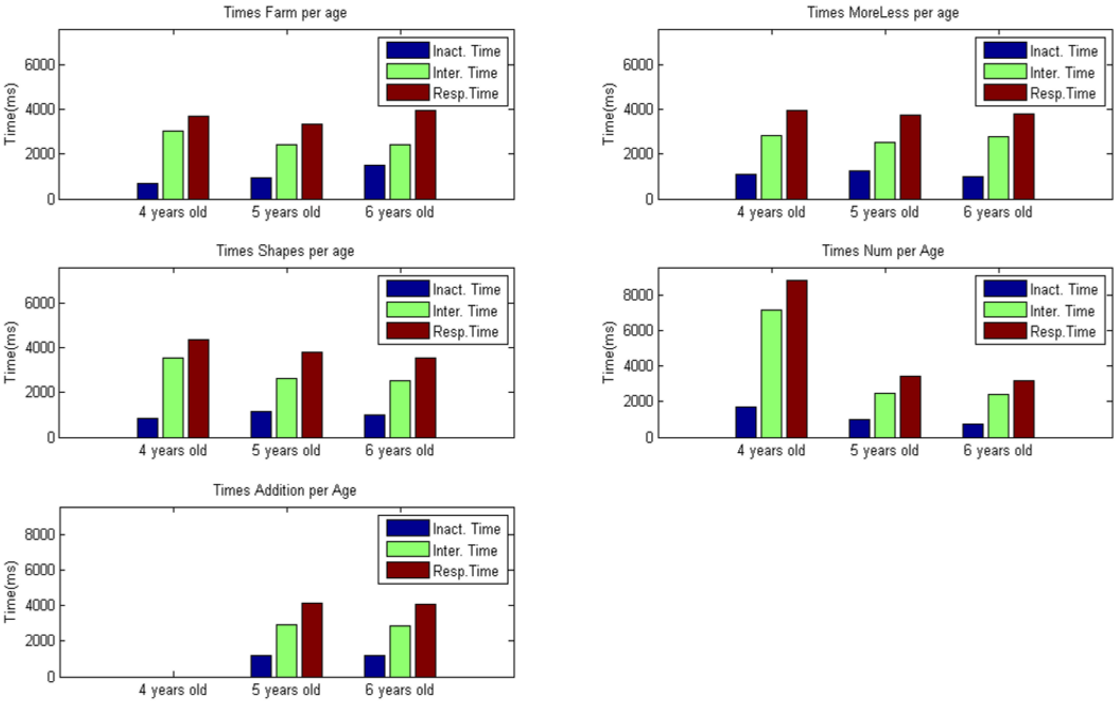


Figure 7.4: Inactivity, interaction and response time per task and age.

7.2.2 Modality Usage and Task Completion

Five and six year-old children have significantly better task completion statistics (around 97%), while 4 year-olds are close to 90%. This is mainly due to the fact that younger children, when facing a difficulty in a task, they often chose to play another task. Older children are usually more persistent and insist until they complete the task at hand.

Task completion percentage for girls is significantly higher than that of boys (97.62% and 90.32% respectively). Again this difference can be attributed to persistence. In terms of modality usage (see Fig. 7.5), we observe a drop of speech usage with increasing age. At the age of four the mouse input usage is close to 16%. At the age of five 19% and at the age of six 23%. This is partly due to the familiarity that older children have with the mouse input device, as explained next. However, speech remains their main input modality for all age groups. Modality usage is similar for boys and girls.

7.2.3 Correlation between various objective metrics

In Table 7.2, the correlation between various objective factors, such as age and gender is shown. As expected, there is a negative correlation between response time and age,

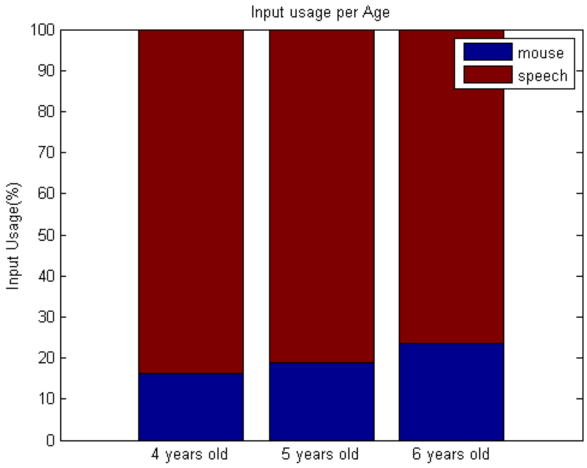


FIGURE 7.5: Inactivity, interaction and response time per task and age.

i.e., as the children grow their response time improves. Also there is positive correlation between mouse skill (the mouse skill was evaluated by the person performing the exit interview) and age. Thus as the children grow, the mouse skill improves and they use mouse input more. Finally there is correlation between gender and inactivity time (girls have on average higher inactivity time than boys) and between gender and task completion, as we discuss in 7.2.2.

Factor pair	**Corr. Coef.**	**p-value**
Resp. Time/Age	-0.2524	0.0378
Inter. Time/Age	-0.2997	0.0130
Mouse Skill/Age	0.6272	0
Mouse usage/Age	0.2580	0.0338
Gender/Inact. Time	0.4041	0.0006
Gender/Task Compl.	0.2603	0.0321

TABLE 7.2: Correlation between various objective metrics, age and gender.

Table 7.3 shows the correlation between various objective metrics and speech usage. As expected there is negative correlation between speech usage and age. As the children grow their mouse skill improves so they use both mouse and speech. Also there is correlation between speech usage and response time. This is obvious, since speech modality is slower than the mouse modality for the specific tasks, thus the interaction time is higher. Finally there is correlation between speech usage and task completion. That is because as we discuss later, speech modality is very popular among the children, and that was another motive for them to complete the tasks.

Finally in Table 7.4 the correlation between time metrics (inactivity, interaction and response time) and various objective metrics is shown. There is correlation between inactivity and interaction time, with the given wrong answers. As the number of wrong

Factor pair	**Corr. Coef.**	**p-value**
Speech usage/Age	-0.2579	0.0338
Speech usage/Resp. Time	0.1996	0.1026
Speech usage/Inter. Time	0.1816	0.1287
Speech usage/Wrong answers	0.2726	0.0245
Speech usage/Task completion	0.4061	0.0006

Table 7.3: Correlation between various objective metrics and speech usage.

answers increase, the inactivity and interaction time are also increases, thus the response time increases too. This is because when children make mistakes, they think more before answering again, thus the inactivity time increases. Also children are more cautious and uncertain when they give their answer. Also note the negative correlation between average response time and mouse skill, i.e., as the mouse skill increases the average response time falls.

Factor pair	**Corr. Coef.**	**p-value**
Resp. Time/Mouse skill	-0.2540	0.0366
Inact. Time/Wrong Ans.	0.3002	0.0129
Inter. Time/Wrong Ans.	0.3696	0.0019
Resp. Time/Wrong Ans.	0.4143	0.0005

Table 7.4: Correlation between various objective metrics and speech usage.

7.3 Subjective Metrics

Subjective evaluation is the second part of our analysis. In this section we discuss user preferences regarding fantasy, curiosity and difficulty, and we examine if these factors are correlated with entertainment. Furthermore we examine if there is correlation between these factors and the various objective metrics. Finally we report the results from the exit interview.

7.3.1 Users preferences and Entertainment

First we evaluate how fantasy, curiosity and difficulty/challenge affect the user experience. As shown in Fig. 7.6(a), most children preferred the application with higher levels of fantasy and curiosity. Specifically, six out of the nine children picked the version of the game with story and fantasy triggers (fantasy level 2). Also six out of the nine children chose the game version with randomly created characters, random task proposals and answer bar (curiosity level 2). In Fig. 7.6(b) the selected "best" system configuration is shown. Systems with high values of fantasy, curiosity and difficulty were the most popular among the children.

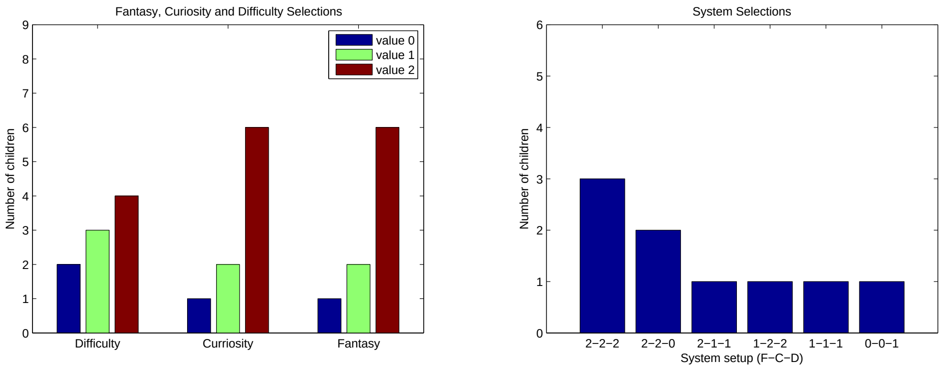


Figure 7.6: (a) Histogram of subjective optimal levels of fantasy, curiosity and difficulty. (b) Histogram of system picked best overall (by the user), e.g., 2-1-0 corresponds to fantasy level 2, curiosity 1, difficulty 0.

In order to compute the correlation between the three factors and entertainment, we have labeled each system version as "entertaining" or "not entertaining" based on the child's preferences, i.e., for each session/factor one system setup (the one picked by the child) is labeled "entertaining" and the other two "not entertaining". Table 7.5 shows the correlation coefficients (and their corresponding p-values) between the level of each factor (fantasy, curiosity, challenge) and entertainment (binary variable defined above). Both fantasy and curiosity are positively correlated with child's entertainment.

Factor	**Corr. Coef.**	**p-value**
Fantasy	0.2778	0.0120
Curiosity	0.2778	0.0120
Difficulty	0.1667	0.1370

Table 7.5: Correlation between entertainment and the three factors.

Per the challenge factor, it seems that the preferred level of difficulty is very much child-dependent. Two children chose easy difficulty, three medium and four children selected the games with high difficulty level. As a result the correlation between difficulty and entertainment is modest.

Table 7.6 shows the correlation between the three Malone factors and various objective metrics. The results show that there is correlation between fantasy and speech usage. Higher levels of fantasy increases speech usage. This is due to the fact that when children absorbed playing the game, they become more spontaneous. Thus they use more the speech modality which is more natural communication modality.

Also note the correlation between curiosity and time (response and interaction); for high levels of curiosity there is more information on screen and the cognitive load increases. Despite this, children prefer higher levels of curiosity. The correlation between task completion and curiosity, indicates that the high levels of curiosity, is another motivation

Factor pair	**Corr. Coef.**	**p-value**
Fantasy/Speech usage	0.2236	0.0668
Curiosity/Resp. Time	0.1888	0.1231
Curiosity/Inter. Time	0.1910	0.1186
Curiosity/Task Compl.	0.2850	0.0185
Difficulty/Wrong Ans.	0.1909	0.1190
Difficulty/Inact. Time	0.1985	0.1046

Table 7.6: Correlation between the three factors and various objective metrics.

for children to complete the selected task. Furthermore, as expected there is correlation between difficulty and wrong answers (as the difficulty level increase more wrong answers are given). Finally the correlation between difficulty and inactivity time is because more thinking is needed for more difficult tasks.

7.3.2 Questionnaire Results

The results from the exit interview are shown in Table 7.7. Note that for the last three questions only the most popular answers are shown at the Table. Most children enjoyed interacting with the system using speech, liked the graphics, sounds and animation of the games, and enjoyed the underlying story. Finally, most children would like to interact again with the application in the future. Also it is interesting that children found wearing a headset consistently annoying.

7.4 Audio Data analysis

Besides the objective and the subjective metrics that we discuss in this Chapter, we are also analyzed the audio data that we have collected. Using the audio data, we have tried to find the correlation between various audio feature statistics and the three Malone factors. First we discuss the audio features that we have used and their correlation with fantasy, curiosity and difficulty. Then we are investigate how to classify the child's emotions into two categories (negative and non-negative) using these audio features.

7.4.1 Audio Feature Extraction

Thirteen different pitch and energy audio features are extracted from the collected audio data using the Praat toolkit [48]. Minimum, maximum, mean, median and standard deviation of pitch and intensity are computed, as well as energy, F0 points and duration.

Question	YES	NO
Did you like that you can speak to game characters?	96.3%	3.7%
Did you like the game graphics, sound and animations?	88.8%	11.2%
Did the characters listen to you when you talk to them?	92.6%	7.4%
Did you understand what they said to you?	92.2%	7.8%
Did you like the story?	88.8%	11.2%
Would you like to play different story in the future?	96.3%	3.7%
Does the headset annoying you?	11.2%	88.8%
Would you like to play the games again in the future?	85.2%	14.8%
What was your favorite character?	Rabbit, Farmer	
What was your favorite game?	Farm, Additions	
What do you not like about the games?	Bees, Numbers recognition task	

Table 7.7: Exit interview results.

Using these audio features, we have tried to investigate if there is correlation between them and the three Malone factors, as well as the reported entertainment value. Moderate correlation with optimal levels of fantasy, curiosity and difficulty was found. For example correlation values between pitch statistics (average session pitch minus average speaker pitch) and fantasy was 0.1621. Although more research work (and more data) are needed to identify good predictors of user preferences, in order to maximize the engagement and enjoyability of child-computer interaction.

7.4.2 Emotion Classification

Using the extracted audio features we have build three classifiers that classify emotions into two categories (negative, non-negative). Three post graduate students labeled the collected data and assigned them into 5 different emotion categories namely anger, happy, neutral, boring and sad. In our experiments we have use only the data that all the three labelers agreed on.

The selected data set consists of 234 instances as shown in Table 7.8. Due to the fact that for some categories we don't have enough data, we organize the labeled data into two different categories. Non-Negative emotions (neutral, happy) and negative emotions

Label	Instances
Happy	19
Neutral	170
Anger	13
Sad	17
Boring	15

Table 7.8: Emotion Classification Dataset.

(angry, sad, boring). Then we build and evaluate three different classifiers (Bayes, NNR-3 and LDA) using leave one out cross validation. The results are shown in Table 7.9.

Classifier	Possitive	Negative	Total
Bayes Accuracy (%)	90.09	95.45	90.6
NNR-3 Accuracy (%)	94.81	31.82	88.8
LDA Accuracy (%)	95.28	95.45	95.3

Table 7.9: Positive/Negative emotion classification results.

Bayes classifier has manage to classify correct 90.09% of the positive and 95.45% of the negative samples. LDA classifier has the higher accuracy with 95.28% at the positive samples and 95.45% at the negative samples. Finally NNR-3 has low accuracy (31.82%) in negative samples and high (94.81%) in positive samples. Compare to older children (ages 7-15) and adults [14, 38], the accuracy of emotion classification for younger children is higher. This is probably due to the fact that younger children are more spontaneous, and as a result it is easier to identify their emotional state.

7.4.3 Emotion and objective metrics correlation

In Table 7.10 the correlation between the percentage of negative and positive utterances per session, inactivity time and wrong answers are shown. In order to compute the percentage of positive and negative utterances, we labeled each sample using the LDA classifier.

Factor pair	Correlation	p-value
Positive Emotions/Inact. time	-0.2265	0.0632
Negative Emotions/Inact. time	0.2265	0.0662
Positive Emotions/Wrong ans.	-0.2418	0.0470
Negative Emotions/Wrong ans.	0.2404	0.0483

Table 7.10: Correlation between emotional state of the children, inactivity time and wrong answers.

The results show that there is correlation between inactivity time and the emotional state of the children. Long inactivity time is indicator of negative emotions while short inactivity time is indicator of positive emotions. Also correlation between the wrong answers and the emotional state of children is found. As shown in the table, wrong answers cause negative feelings to the children.

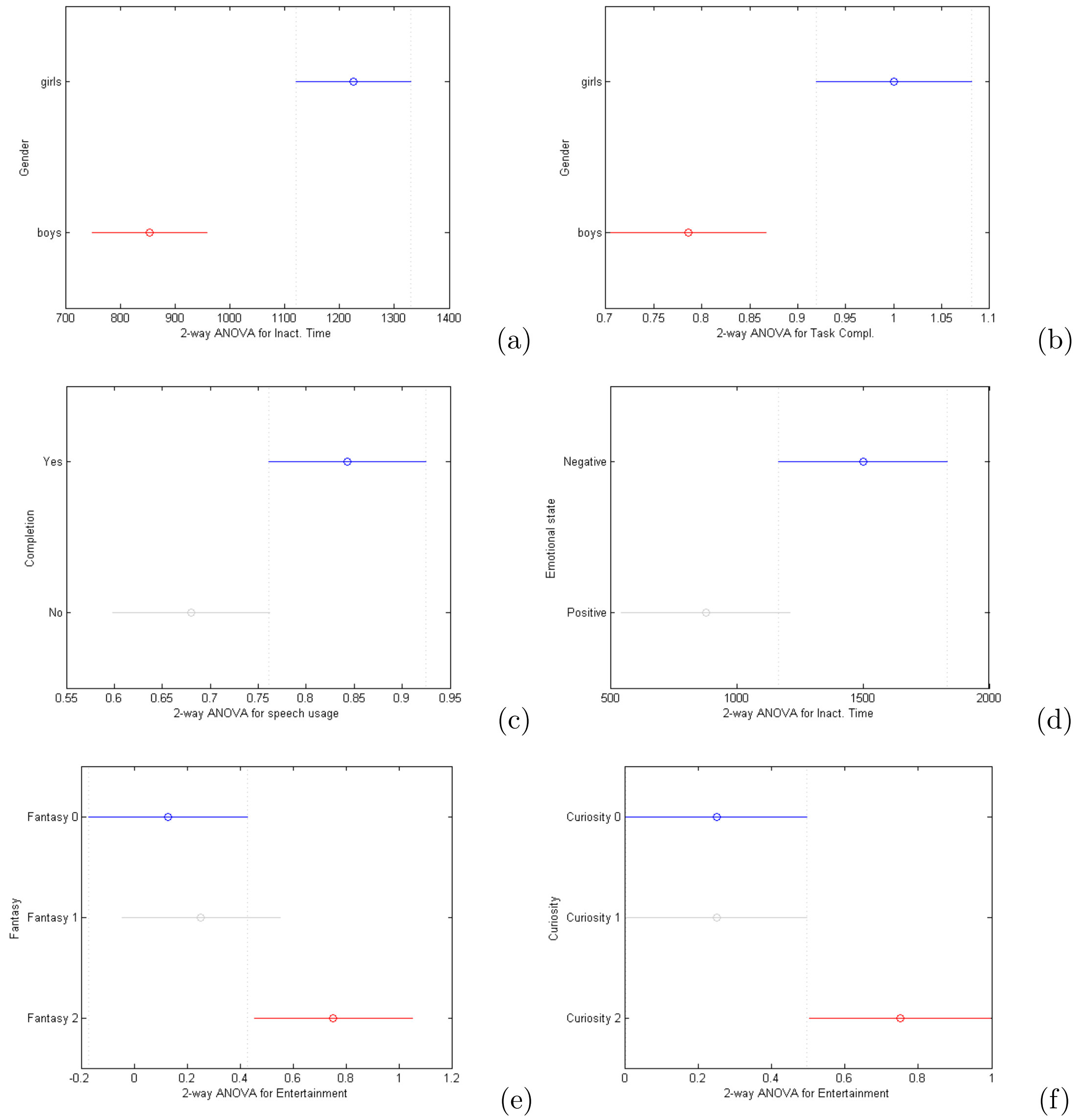
7.5 ANOVA Analysis

In order to confirm the statistical significance of our results, 2-way ANOVA statistical analysis was performed at the collected data. Specifically, there is a significant reduction of inactivity time between boys and girls (see Fig. 7.7(a)). Also there is a significant difference between boys and girls in the task completion. As we said before, girls are usually more persistent and insist until they complete the task at hand (see Fig. 7.7(b)).

Fig. 7.7(c) shows that the usage of speech compare to task completion is statistical significant. It seems that speech usage is an extra motive for children in order to complete the task at hand. Furthermore Fig. 7.7(d) shows that the relation between emotional state and inactivity time is almost significant.

Finally Fig. 7.7(e) and Fig. 7.7(f) shows the significant relation between entertainment and fantasy and curiosity factors. Specifically the entertainment is statistical higher for curiosity and fantasy level 2 compare to curiosity and fantasy level 0.

Although, because of the small number of participants in our experiments, more data are needed in order to find more significant results. A second series of experiments with more children could help in that direction.



(a)

(b)

(c)

(d)

(e)

(f)

Figure 7.7: (a) 2-way ANOVA results (gender and inactivity time), (b) 2-way ANOVA results (gender and task completion), (c) 2-way ANOVA results (speech usage and task completion), (d) 2-way ANOVA results (emotional state and inactivity time), (e) 2-way ANOVA results (Fantasy and entertainment), (f) 2-way ANOVA results (curiosity and entertainment).

Chapter 8

Conclusions and Future Work

8.1 Conclusions

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 starting point to develop web-based speech-enabled applications for children. Five preschool games were implemented based on popular preschool activities and evaluated by nine children of ages 4-6.

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

Moreover experiments showed that high levels of fantasy encourage the use of voice modality. When children interact inside a fantasy world they are more spontaneous, thus they use the most natural modality, in our case speech. Similarly, speech usage increase the task completion which indicates that speech usage is another motive for children to complete the task. Also high levels of curiosity could increase the cognitive load, although they can also give an extra motive to the child in order to finish the task in hand.

Preliminary experiments also showed that interaction patterns and acoustic features are indicators of (subjectively) optimal levels of fantasy, curiosity and difficulty. Moreover we showed that emotion classification is possible in such ages. The classification results was better compare with adults and older children and that is due to the fact that younger children are more spontaneous in their interaction with the computer. Furthermore wrong answers and inactivity time are good indicators for recognize negative emotions.

The WoZ experiment in a gaming environment provided data for further use in the creation of novel language models, understanding strategies for dialog systems and help the identification of more indicators towards adapting multimodal dialogue systems for children.

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.

8.2 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, the Wizard of Oz should be replaced with an Automatic Speech Recognition System. With the collected data, the necessary acoustic and language models could be build in order to replace the Wizard of Oz module with an Automatic Speech Recognizer. The ASR module is already exist and it is functional for adult users.

Furthermore there are many thing that can be done regarding the adaptation of fantasy, curiosity and difficulty. By improving the existing emotion classifier and using the average response time and the correct/wrong answers as indicators, we can easily implement the difficulty adaptation in the system, in a way that the user during the gameplay stays in “flow”. When the children is negative excited (Anxiety) the difficulty level could be drop down. On the other hand when negative calm (Boredom) is detected the difficulty level could be increase. By adapting the difficulty value and by offering high values of curiosity and fantasy in our games, we are going to achieve high entertainment.

Finally more work should be done in order to investigate the exploration and exploitation elements in such games. I our work we give more emphasis in the exploitation part, but exploration is a parameter that should be investigate. This could be done by building a more sophisticate interface with exploration elements. Different stories could also be implement. Different people will find different fantasies appealing, so we should be aware of likes and dislikes relating to social groups such as gender, race, age, etc.

Appendix A

System setup manual

In this Appendix the installation manual of the platform is provided.

Instructions:

- Install the Java Development Kit (JDK) 1.5 or later. In this work we used JDK 1.6.0.11

- Install the Apache Tomcat web server. Tomcat 5.5 is recommended.

- Install the My SQL Database and run the children.db script from the thesis cd in order to create the database.

- Copy the files from the JSP directory of the thesis cd to the /Webapps/ROOT/children folder in the tomcat directory.

- Compile the java beans using the following commands:

  javac web/persistence/*.java

  javac web/children/*.java

  You have to be sure that the username and password variables in the javabeans are the same with the ones you put during the database installation.

- In the /Webapps/ROOT/WEB-INF/ folder of the Tomcat, create a directory with the name “classes” and copy the web folder with the compiled javabeans in it.

- Copy the mysql connector from the thesis cd to the /Webapps/ROOT/WEB-INF/lib and the web.xml to the Tomcats conf folder.

- Compile the Controller Applet using the following commands: javac Controller.java

  jar cvf controller.jar *.class

jarsigner -keystore myKeyStore controller.jar me

Copy the controller.jar file to the /Webapps/ROOT/children folder.

- Startup the web server. In order to be sure that it is working, open the following address in your browser: http://localhost:8080/. In order to have access to the platform open the following address: http://localhost:8080/children

Appendix B

Data Analysis Readme

In this Appendix the data analysis scripts are explained.

B.1 Instructions

B.1.1 Database foldering

The collected data are stored hierarchically based on date, user and session id. In each session folder the log files for each game are exist and the collected audio data. All the data during the experiments are stored in the WoZ directory.

B.1.2 Extracting information from log files

In order to extract the necessary data from the log files, run the stats.pl script. All the data are stored in a new file called datastats.txt

To extract the audio features from the collected data run the features.praat script. In order to run the script the Praat Toolkit is needed.

To extract the correlations between the various objective and subjective metrics, execute the correlation.m file in the MATLAB environment. For Emotion Classification run emotions.m script. Finally for ANOVA analysis run the anova.m script.

More detailed analysis could be found in the scripts.

Bibliography

[1] A. Potamianos and S. Narayanan, "Robust recognition of children's speech," *IEEE Transactions on Speech and Audio Processing*, vol. 11, pp. 603-616, Nov. 2003.

[2] Bruckman, A. and Bandlow, A, "Human-computer interaction for kids," In *the Human-Computer interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications*, J. A. Jacko and A. Sears, Eds. Human Factors And Ergonomics. L. Erlbaum Associates, Hillsdale, NJ, 428-440, 2003.

[3] Malone, T. W, "What make things fun to learn? A study of intrinsically motivating computer games," In *Proceedings of the 3rd ACM SIGSMALL Symposium and the First SIGPC Symposium on Small Systems*, Palo Alto, California, United States, September, 1980.

[4] Malone, T. W, "Heuristics for designing enjoyable user interfaces: Lessons from computer games", In *Proceedings of the 1982 Conference on Human Factors in Computing Systems,* Maryland, United States, March, 1982.

[5] Malone, T., Lepper, M., "Making learning fun: A taxonomy of intrinsic motivations of learning". In R. E. Snow & M. J. Farr (Eds.), *Aptitude, learning, and instruction: Vol. 3. Conative and affective process analyses* pp. 223-253, Hillsdale, NJ: Lawrence Erlbaum, 1987.

[6] S. Narayanan and A. Potamianos, "Creating conversational interfaces for children," *IEEE Transactions on Speech and Audio Processing*, vol. 10, pp. 65-78, February, 2002

[7] S. Lee, A. Potamianos, and S. Narayanan, "Acoustics of children's speech: Developmental changes of temporal and spectral parameters," *The Journal of the Acoustical Society of America*, vol. 105, pp. 1455-1468, March, 1999.

[8] N. Beringerb,U. Kartal, K. Louka, F. Schiel, and U. Turk, "Promice: A procedure for multimodal interactive system evaluation," In *Proceedings of the LREC Workshop on Multimodal Resources and Multimodal Systems Evaluation*, Las Palmas, Spain, pp. 271-280, 2002.

[9] Lazzaro N, "Why We Play Games: Four Keys to More Emotion Without Story," *Technical Report*, XEO Design Inc., 2004, available at: http://www.xeodesign.com.

[10] Csikszentmihalyi M, "Flow: The Psychology of Optimal Experience," New York: Harper & Row, 1990.

[11] Yannakakis, G. N., and Hallam, J., "Evolving Opponents for Interesting Interactive Computer Games," In *Proc. of the 8th International Conference on Simulation of Adaptive Behavior.* The MIT Press, pp. 499-508, 2004.

[12] G. N. Yannakakis, J. Hallam, and H. H. Lund, "Comparative Fun Analysis in the Innovative Playware Game Platform," In *Proceedings of the 1st World Conference for Fun n' Games*, pp. 33-37, 2006.

[13] Xiao Benfang, Girand Cynthia, and Sharon Oviatt, "Multimodal integration patterns in children," In *Proc. ICSLP-2002*, pp. 629-632, 2002.

[14] Yildirim Serdar, Lee Chul Min, Lee Sungbok, Potamianos Alexandros, Narayanan Shrikanth, "Detecting Politeness and frustration state of a child in a conversational computer game," In *Proc. European Conf. on Speech Communication and Technology*, pp. 2209-2212, Lisbon, Portugal, 2005.

[15] Narayanan Shrikanth, Potamianos Alexandros, Wang Haohong, "Multimodal systems for children: building a prototype," In *Proc. European Conf. on Speech Communication and Technology*, pp. 1727-1730, Budapest, Hungary, September, 1999.

[16] Chanel G., Rebetez C., Betrancourt M., and Pun T., "Boredom, Engagement and Anxiety as Indicators for Adaptation to Difficulty in Games," In *Proceedings of the 12th international Conference on Entertainment and Media in the Ubiquitous Era*, Tampere, Finland, October, 2008.

[17] G. N. Yannakakis, J. Hallam and H. H. Lund, "Capturing Entertainment through Heart-rate Dynamics in the Playware Playground," In *Proceedings of the 5th International Conference on Entertainment Computing, Lecture Notes in Computer Science*, vol. 4161, pp. 314-317, Cambridge, UK, September, 2006.

[18] M. Perakakis and A. Potamianos., "A study in efficiency and modality usage in multimodal form filling systems." *Audio, Speech, and Language Processing, IEEE Transactions on,* 16(6):1194-1206, Aug. 2008

[19] M. Perakakis and A. Potamianos., "Multimodal system evaluation using modality efficiency and synergy metrics," In *Proceedings of the 10th international Conference on Multimodal interfaces,* Chania, Crete, Greece, October, 2008.

[20] G. N. Yannakakis, and J. Hallam., "Entertainment Modeling in Physical Play through Physiology beyond Heart-Rate," In *Proceedings of the Int. Conf. on Affective Computing and Intelligent Interaction, Lecture Notes in Computer Science*, vol. 4738, pp. 256-267, Lisbon, Portugal, September, 2007.

[21] G. N. Yannakakis, and J. Hallam, "Feature Selection for Capturing the Experience of Fun," In *Proceedings of the AIIDE'07 Workshop on Optimizing Player Satisfaction*, AAAI Press Technical Report WS-01-01, pp. 37-42, Stanford, USA, June, 2007.

[22] Ruff, H. A. & Lawson, K. R., "Development of sustained, focused attention in young children during free play," *Developmental Psychology,* pp. 85-93, 1990.

[23] Piaget J., "Science of Education and the Psychology of the Child," *Published by Orion Press*, New York, 1970.

[24] E. F. Strommen and F. S. Frome, "Talking back to big bird: Preschool users and a simple speech recognition system," *Educational Technology Research and Development*, vol. 41, pp. 5-16, 1993.

[25] J. Mostow, A. G. Hauptmann, and S. F. Roth, "Demonstration of a reading coach that listens," in *Proceedings of the ACM Symposium on User Interface Software and Technology*, pp. 77-78, 1995.

[26] M. Russell, B. Brown, A. Skilling, R. Series, J. Wallace, B. Bonham, and P. Barker, "Applications of automatic speech recognition to speech and language development in young children," in *Internat. Conf. Speech Language Processing,* (Philadelphia, PA), October, 1996.

[27] A. Hagen, B. Pellom, and R. Cole, "Children's speech recognition with application to interactive book and tutors," in *Proc. ASRU Workshop,* 2003.

[28] J. Cassell and K. Ryokai, "Making Space for Voice: Technologies to Support Children's Fantasy and Storytelling," *Personal Technologies*, vol. 5, 2001.

[29] Gustafson, J., Bell, L., Boye, J., Lindstrom, A. and Wiren, M., "The NICE Fairytale Game System," *Proceedings of SIGdial 04*, Boston, April, 2004.

[30] Busso, C., Deng, Z., Yildirim, S., Bulut, M., Lee, C. M., Kazemzadeh, A., Lee, S., Neumann, U., and Narayanan, S., "Analysis of emotion recognition using facial expressions, speech and multimodal information," In *Proceedings of the 6th International Conference on Multimodal interfaces* (State College, PA, USA, October, 2004).

[31] Dix A., Finlay J. and Beale R., "Human-Computer Interaction." Prentice Hall, 2004.

[32] Oviatt, S. , "Taming recognition errors with a multimodal interface," *Communications of the ACM 43,* p. 45-51, Sep. 2000).

[33] C. Naas, L. Cong. "Ten principles for designing human-computer dialog systems," *In (D.A Dahl,editor) Practical Spoken Dialogue Systems,* p. 143-163. Kluwer Academic Publishers, 2004.

[34] S. Das, D. Nix, and M. Picheny, "Improvements in children's speech recognition performance," in *Proc. ICASSP,* pp. 433-436, 1998.

[35] Gustafson, J., Sjolander, K., "Voice transformations for improving childrens speech recognition in a publicly available dialogue system," in *Proc. ICSLP-2002,* pp. 297-300, 2002.

[36] Ermi, L., Mayra, F., "Fundamental components of the gameplay experience: Analysing immersion," In: S. de Castel & J. Jenson (eds.), *Changing Views: Worlds in Play., Selected papers of the 2005 Digital Games Research Association's (DiGRA) Second international Conference,* 2005.

[37] Jameson, A., "Adaptive interfaces and agents," *In the Human-Computer interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications,* J. A. Jacko and A. Sears, Eds. Human Factors And Ergonomics. L. Erlbaum Associates, Hillsdale, NJ, pages 305-330, 2003.

[38] Lee C. M., Narayanan, S.S., Pieraccini, R., "Classifying emotions in human-machine spoken dialogs," In *Proceedings of ICME '02. Proceedings.* pages 737-740, August 2002.

[39] C. C. Chiu, Y. L. Chang, and Y. J. Lai, "The analysis and recognition of human vocal emotions," in *Proc. International Computer Symposium 1994,* pp. 8388, NCTU, Hsihchu, Taiwan, R.O.C., December, 1994.

[40] P. Ekman, W.V. Friesen, "Facial Action Coding System: Investigator's Guide," *Consulting Psychologists Press,* 1978.

[41] Lee, Chul Min, Narayanan, Shrikanth S., Pieraccini, Roberto "Combining acoustic and language information for emotion recognition," In *ICSLP-2002,* pp. 873-876, 2002.

[42] Rosalind W. Picard, Elias Vyzas, Jennifer Healey, "Toward Machine Emotional Intelligence: Analysis of Affective Physiological State," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, pp. 1175-1191, 2001.

[43] Young, R.M., “An overview of the Mimesis Architecture: Integrating Intelligent Narrative Control into an Existing Gaming Enviroment” Working Notes *of the AAAI Spring Symposium on Artificial Intelligence and Interactive Entertainment*, 2001.

[44] Cavazza, M.F. Charles and S. J. Mead, “Character based interactive storytelling,” *IEEE Intelligent Systems, Special issue on AI in Interactive Entertainment*, pp. 17-24, 2002.

[45] Meliopoulos Spyros, “Multimodal System for Preschool Children,” Diploma Thesis, Technical University of Crete, October, 2008.

[46] “CSLR Reading Tutor Project (2002)”. Available at: http://cslr.colorado.edu/beginweb/reading/reading.html/.

[47] “Colorado Literacy Tutor (2002)”. Available at: http://colit.org.

[48] Boersma, P., Weenink, D., “Praat: doing phonetics by computer (Version 4.6)”. Available at: http://www.praat.org/.

[49] Macromedia Flash support page. Available at: http://www.adobe.com/products/.