Università degli Studi di Padova

HEAD OFFICE: Università degli studi di Padova
Department of Information Engineering (DEI)

PH.D COURSE: Information Engineering
CURRICULUM: Information Communication Technology (ICT)
SERIES: XXXII

Models and methods for sound-based input in Natural User Interfaces

Coordinator
Ch.mo Prof. Andrea Neviani

Supervisor
Ch.mo Prof. Antonio Rodà

Ph.D. Student
Edoardo Micheloni

Contents

Abstract

In the last years, the Multimodal Interfaces and the Natural User Interfaces (NUIs) are finding more and more applications, thanks to the diffusion of mobile devices and smart objects that do not allow a traditional WIMP interaction. In these contexts, the interaction modes most used are the natural language and the gestures recognition. The objective of this thesis is to explore innovative interfaces based on non-verbal sounds, produced by the interaction of the user with common objects. The potentialities and the problems related to the design and implementation of this type of interfaces will be discussed through three case studies, in which non-verbal sounds are used for interaction with embedded systems developed for the valorization of cultural heritage. The sounds analysed in these projects are i) broadband noises, ii) impulses and iii) pitched sounds. The obtained results, thanks to a strong multidisciplinary approach, opened to a fruitful technology transfer between university and companies/institutions involved. First of all, the study of broadband noisy sounds was addressed through the interpretation of air blown signal. The resulting sensors equipped system was included in a multimedia installation for the valorization of an ancient Pan flute preserved at the Museum of Archaeological Sciences and Art of Padova (Italy). Secondly, the impulsive sounds were studied from footsteps detection on a wooden runway in order to realize a real-time position mapping technology. The resulting system was used for the 3D exploration of a usual 2D painting exposed during "The European Researchers' Night 2018" in Padova (Italy). Finally, pitched sound signals were studied analysing notes produced by an acoustic piano. The resulting algorithm for real time note detection was applied to the video-game *Musa*, which had the goal to teach children how to play the piano. In these projects, both the algorithms, by means of quantitative analysis, and the interfaces

between user and computer, by means of qualitative analysis, were validated to assess the "naturalness" of the interaction.

Sommario

Negli ultimi anni, le Interfacce Multimodali e le Interfacce Utente Naturali (NUIs) stanno trovando sempre più applicazioni, grazie alla diffusione di dispositivi mobili e oggetti intelligenti che non consentono l'interazione WIMP tradizionale. In questi contesti, le modalità di interazione più utilizzate sono il linguaggio naturale e il riconoscimento dei gesti. L'obiettivo di questa tesi è esplorare interfacce innovative basate su suoni non verbali, prodotti dall'interazione dell'utente con oggetti comuni. Le potenzialità e i problemi relativi alla progettazione e all'implementazione di questo tipo di interfacce saranno discussi attraverso tre casi di studio, in cui i suoni non verbali vengono utilizzati per l'interazione tramite sistemi integrati sviluppati per la valorizzazione del patrimonio culturale. I suoni analizzati in questi progetti sono: i) rumori a banda larga, ii) impulsi e iii) suoni intonati. I risultati ottenuti, grazie a un forte approccio multidisciplinare, hanno aperto al trasferimento tecnologico tra l'università e le aziende/istituzioni coinvolte. Nel primo progetto, lo studio dei suoni rumorosi a banda larga è stato affrontato attraverso l'interpretazione del segnale emesso da un soffio. Il sistema di sensori risultante è stato incluso in un'installazione multimediale per la valorizzazione di un antico flauto di Pan conservato presso il Museo delle Scienze Archeologiche e d'Arte di Padova (Italia). Nel secondo progetto, i suoni impulsivi sono stati analizzati mediante il rilevamento dei passi di un utente su una pedana di legno, al fine di realizzare una tecnologia di mappatura della posizione in tempo reale. Il sistema risultante è stato implementato per l'esplorazione 3D di un dipinto contemporaneo esposto durante "La notte europea dei ricercatori 2018" a Padova (Italia). Come ultimo progetto, i segnali audio intonati sono stati analizzati per mezzo di un pianoforte acustico. L'algoritmo risultante per il rilevamento multi-pitch in tempo reale delle note è stato applicato al videogioco *Musa*, avente

l'obiettivo di insegnare ai bambini a suonare il pianoforte. In questi progetti, sia gli algoritmi, mediante analisi quantitative, sia le interfacce tra utente e computer, mediante analisi qualitative, sono stati analizzati per valutare la "naturalezza" dell'interazione.

Acknowledgements

I would simply like to thank you all the people who allowed the achievement of this Ph.D.: first of all, my supervisor Prof. Antonio Rodà, Prof. Sergio Canazza and all the people that I had the pleasure to meet and work with at the CSC. It's been three busy years but full of experiences. Thanks to my family and my friends, Padova and Via Marzolo. Never stop!

Publications

Published Journal Paper

- Niccolò Pretto, Carlo Fantozzi, Edoardo Micheloni, Valentina Burini, Sergio Canazza, "Computing Methodologies Supporting Preservation of Electroacoustic Music from Analog Magnetic Tape", Computer Music Journal (CMJ), December 2018.

- Edoardo Micheloni, Marco Tramarin, Antonio Rodà, Federico Chiaravalli, "Playing to play: a piano-based user interface for music education video-games", Multimedia Tools and Applications, Springer, December 2018.

Accepted Journal Paper

- Niccolò Pretto, Edoardo Micheloni, Silvia Gasparotto, Carlo Fantozzi, Giovanni De Poli, Sergio Canazza, "Technology-enhanced interaction with cultural heritage: an antique Pan flute from Egypt", ACM Journal on Computer and Cultural Heritage (JOCCH).

International and national conference paper

- Giulio Pitteri, Edoardo Micheloni, Antonio Rodà, Carlo Fantozzi, Nicola Orio, "Listen By Looking: a mobile application for augmented fruition of live music and interactive learning", Proc. of 5th EAI International Conference on Smart Objects and Technologies for Social Good (GOODTECHS 2019), 2019.

- Edoardo Micheloni, Marcella Mandanici, Antonio Rodà, Sergio Canazza, "Interactive painting sonification using a sensor-equipped runway", in Proc. Int. Conf. Sound and Music Computing (SMC 2017), July 2017

- Edoardo Micheloni, Niccolò Pretto, Sergio Canazza, "A Step toward AI Tools for Quality Control and Musicological Analysis of Digitized Analogue Recordings: Recognition of Audio Tape Equalizations", in Proc.of 11th workshop on Artificial Intelligence for Cultural Heritage (AI*CH 2017), November 2017

- Matteo Lionello, Marcella Mandanici, Sergio Canazza, Edoardo Micheloni, "Interactive Soundscapes: Developing a Physical Space Augmented through Dynamic Sound Rendering and Granular Synthesis", in Proc. Int. Conf. Sound and Music Computing (SMC 2017), July 2017

- Federico Avanzini, Federico Avanzini, Sergio Canazza, Giovanni De Poli, Carlo Fantozzi, Edoardo Micheloni, Niccolò Pretto, Antonio Rodà, Silvia Gasparotto, Giuseppe Salemmi, "Virtual reconstruction of an ancient Greek pan flute", in Proc. Int. Conf. Sound and Music Computing (SMC 2016), June 2016

- Edoardo Micheloni, Niccolò Pretto, Federico Avanzini, Sergio Canazza, Antonio Rodà,"Installazioni interattive per la valorizzazione di strumenti musicali antichi: il flauto di pan del Museo di Scienze Archeologiche e d'Arte dell'Università degli Studi di Padova" in XXI Colloquio di Informatica Musicale (CIM 2016), June 2016.

List of Figures

List of Tables

Chapter 1

Introduction

For a long time, it seemed that the human-computer interfaces were to be limited to working on a desktop computer, using a mouse and a keyboard to interact with windows, icons, menus, and pointers (WIMP). Nowadays, physical input devices that draw on users' skill of interaction with digital world gain increasing popularity. For example, Apple has developed wireless stylus pen accessories used to draw and write on iPads. Simultaneously, the 5G revolution takes place: interaction with "computers" becomes pervasive in everyday objects and environments thanks to integrated computational and mechatronic components. Starting from pervasive computing, today new researches as Internet Of Things give life to the study of new way of interaction between connected objects. As computers have become more ubiquitous, the way we interact with computing has developed in different directions. For personal computers, the dominant paradigm remains the GUI (Graphical User Interface), which is still in many ways the same as the WIMP developed by researchers at Xerox PARC in the 1970s. For mobile devices, touchscreens have become the dominant form of interaction, and they are also making inroads to the productivity domain, in particular on ultralight laptops and tablets. At home, vocal interaction has taken great strides recently, especially due to advances in artificial intelligence and data collection, with a number of "smart speakers" available from various manufacturers (i.e. Google Home, Amazon Echo). In entertainment and training, interest in virtual and augmented reality is spreading, although usage still remains largely limited to specialists and enthusiasts.

Interaction is the basis of "multimodal interaction" [18], "natural user interface"[195], "tangible interfaces"[91] and "physically-based interaction"[95] which draw upon the human urge to be active and creative with one's hands [197], and can provide a means to interact with computational applications in ways that leverage users' knowledge and skills of interaction with the everyday, non-digital, world [95]. Multimodal interaction aims to supporting the recognition of naturally-occurring forms of human language and behaviour through the use of recognition-based technologies [147, 192] exploiting different tools for input and output of data. Natural user interfaces (NUI) are a type of user interface designed to be perceived as natural as possible by the user. The goal of a NUI is to create seamless interaction between the human and machine, making the interface itself seem to disappear [195]. Indeed, they are strongly multidisciplinary, since their evaluation and development is based on a strong collaboration and exchange of information between different fields of research [174]. This attention is necessary to ensure natural user interface to be "really" natural [138]. On the other hand, Tangible User Interface (TUI) emerged as a new interface and interaction style where the vision centered on turning the physical world into an interface by connecting objects and surfaces with digital data.

Engineering/information technology is the core of these fields and the starting contribution to Human Computer Interaction (HCI) [153] that nowadays, along with psychology and design, plays a paramount role [172, 141, 126]. HCI is a multidisciplinary field of study which researches the use of computer technology (machine in general), focusing on interfacing the user with the system [44]. Initially, in the 1980s, it was concerning only computers, while today HCI has expanded to cover almost all forms of information technology. Much of the research in the field seeks to improve HCI taking interest in: i) Models and theories of human-technology use, ii) frameworks for the design of interaction interfaces, iii) optimization of systems property as usability, iv) new technologies to interpret user actions. The development of an ideal systems would require expertise in a wide variety of topics which include but are not limited to psychology, sociology, ergonomics, computer science and engineering, art and graphic design. However, in the academic research, there are three main areas working in the field and they are Computer

Science, Design and Psychology.

In HCI, "interaction" is the main topic of research and real-time applications are one of the main focuses studied in the field, developing fast, scalable and adaptable algorithms [104]. The aim is to enhance interaction between a user and the technology, in a most fruitful way, avoiding basic issues that may occur such as interruption [168, 125] of the system or wrong matches between user action and system reaction [132], becoming a virtuous example of usability. A lot of these example can be found in multimedia applications, as in visual/3D interactions [94, 12, 188, 162] or audio processing [26, 34, 43, 170]. Indeed, usability is part of the broader term "user experience" and it refers to the ease of access and/or use of a product/system. A system is not usable or unusable *per se*; its features, together with the context of the user, determine its level of usability.

As engineers, the main approach to the study and the development of models for this purpose should be to use *problem-solving capacity* as a criterion to determine the progress of asking solutions: how do these solutions advance our capacity to solve problems in human use of technology? From the engineering point of view, this can be seen as *Constructive Problem* [145] which covers some of the areas of HCI showing the most of its vitality at conferences, including interactive systems, interactive applications, interface and sensor technology, interaction techniques, input devices, user-interface design, interaction design, and concept design. More importantly, this problem type cuts across design and engineering, both extensive topics. The assessment of these technologies has a paramount importance in order to validate the problem solving capability and to encourage the development and the research in technologies that could be a push forward for scientific progress [102, 201] for multimodal applications and natural user interface.

Along with the technology involved, the interfaces through which the interaction occurs allows a straightforward interpretation from the user on how to perform the necessary actions [174, 186]. One example could be a door for emergency exit: if the handle is not easy to identify or use, whatever technology is behind it, it will not reach the purpose for which it was developed. Today, the most functional and common interfacing technologies are spreading around our daily activities. The most common examples are the touch screen [173] as for smart phone and tablet, speech recognition [70, 80] allowing

to give instructions to virtual assistants and gestures recognition [130, 156] through which, for example, it is possible to raise and lower the radio volume in modern cars. The profit and interest of industries around these technologies are pushing forward the research for further improvement. However, there is still room for new researches to increase the performances of technologies involving multimodal interaction exploiting behaviours of sounds different from voices. This thesis especially focuses on body and sound interaction [183, 68], trying to enhance the connection between our tangible world (i.e. body [149, 171, 185]) and the intangible nature of sound waves. Furthermore, a special focus is the interaction design between user and the interaction systems developed. Interaction design aims at designing interactive digital products, environments, systems, and services [36]. Interaction design is also used to create physical (non-digital) products, exploring how a user might interact with it. Usability testing is the most common method to validate interactive systems. The basic idea behind it is to check whether the product or brand works well with the target users.

In this thesis, along with this research field, three main case studies facing natural user interaction are presented. One of them provides a slightly different paradigm of interface where the natural interaction with the system is the content itself more than the application at which the interface is applied. This paradigm could be seen as slightly different from a natural user interface, which is meant to "disappear" between the user and the system. However, the employed interaction design paradigm was necessary in order to let users "naturally" learn to use the interface.

1.1 Case studies

This thesis tried to open new research lines exploring sound behaviours, going beyond the usual verbal communication, and user interaction encountering two main difficulties. First of all, the interpretation of sound signals behaviours, giving relevance to the interaction. This issue can be seen as the development of a proper semantic that describes events and specific features of the signals. Secondly, the modelling of real-time interaction design with pervasive systems technology and embedded systems to ensure a scalable implementation. The results were promising and opened to a productive technol-

ogy transfer between university and companies/institutions involved. In this thesis, the research conducted on the three types of sounds is presented and applied to case studies of cultural heritage enhancement, where sound (see Sound Music Computing field) and arts are largely studied [49], especially when engineering HCI approach is needed to rely on efficient and usable models of technology, devoted to real-time interactions, applied as interfaces between user and the technology [172, 81, 82]. The works developed have been numerically tested to quantify the performances and tested from the user point of view. This was necessary in order to ensure "problem solving" capability together with a push forward of innovation for sound-based input interaction.

As far as broadband noises are concerned (usually considered something to be avoided during sound analysis [200, 25]), they are not common in nature even though the sound of waterfall is considered very close to white/pink noises. Furthermore, they are not easy to be created by humans. However, exhaling/blowing can be a simple way to produce this type of sound, indeed the spectrum analysis shows the behaviour of a broadband noise. In literature there are several studies about breathing sensors [100, 124] and some are available on sale. Most of them are medical equipment allowing very high precision of measurement and reliability. The result is expensive products that leave no room for the use of breath for scalable applications. In the related case study of this thesis, a sensor based on microphone capsule, able to detect modulations of blow intensity have been developed and than used in a multimedia installation, the aim of which was the enhancement of a Pan flute from ancient Egypt. The installation is now at the museum of Archaeological Science and Arts of Padova. The development of this installation is the result of a strong and productive collaboration between Information Engineering and Cultural Heritage departments of University of Padova. The installation usability, engagement and further parameters have been assessed by a group of experts in the filed of information engineering, musicology and archaeology.

As far as impulsive sounds are concerned (generally considered "impulsive sounds" [48] studied for years from loudness point of view [17, 63]), they are very simple to be created by nature and humans. For instance, clapping hands is a simple example that for years has been seen, in the collective imaginary, has a technological way to turn

on and off lights. Another impulsive sound produced by the human body, and studied in this thesis, are footsteps. In the related case study, a real-time position mapping algorithm based on localization through piezoelectric sensors placed below a wooden runway has been developed. This structure has been used for the "3D interaction" of users with a painting. The interaction design have been evaluated by visitors of "European researchers' night 2018". The results of this research were transferred and used by Microtec s.r.l., a company leader in the sector on diagnosis of wood logs.

As far as pitched sounds are concerned [27, 13], the phonatory system and musical instruments are very good in producing harmonic or pitched sounds. For example, singing is the easiest example of a tuned sounds reproduced by vocal cords and vowels [180] are characterized by harmonic components, feature that can be found in musical instruments like piano and guitars. The study of pitched sound has been pursued for several years but there is still room for further studies where the sound analysed is not coming from a controlled environment. Furthermore, in the related case study of this thesis, the interaction design focused on an interface that wants to be "naturally" learned by the user, becoming the content itself more than a way of interacting with the system. The case study is a video-game, called Musa (and developed by the homonymous company), the means through which the user is induced to learn how to use a complex and not-intuitive control interface like the keyboard of a piano (compared to the usual controller like the joystick) to interact with the video-game. The multi-pitch algorithm developed is now used in the app to detect the notes played on an acoustic piano or a digital keyboard with integrated speakers.

1.2 Multimodal interaction

Human interaction with the world is inherently multimodal [155, 24]. We employ multiple senses, both sequentially and in parallel, to passively and actively explore our environment, to confirm expectations about the world and to perceive new information. Multimodal interaction systems aims to supporting the recognition of naturally occurring forms of human language and behaviour through the use of recognition-based technologies [147, 192]. Multimodal systems represent a research-level paradigm shift

away from conventional WIMP interfaces toward providing users with greater expressive power, naturalness and portability. The importance of this filed of study is highlighted by the recent acquisition of *Facebook* of the start-up *Ctrl-Labs*, estimated in 1 billion dollars, specialized in the development of neural interaction system [1]. The aim of this acquisition seams to be the development of a bracelet that would allow the user to interact with the computer. Well-designed multimodal systems integrate complementary modalities to yield a highly synergistic blend in which the strengths of each mode are used to overcome weaknesses of the others. Such systems can potentially function more robustly than uni-modal systems that involve a single recognition-based technology such as speech, pen, or vision [146]. In contrast to human experience with the natural world, HCI has historically been focused on uni-modal communication (i.e., information or data communicated between human and computer primarily through a single mode or channel, such as text on a screen with a keyboard for input). However, almost all interaction with computers has been multimodal to some degree (i.e. combining typed text with switches, buttons and providing various visual and auditory output signals) for much of interactive computing's history, the model of a single primary channel for data input, and perhaps a different primary channel for data output, has been the norm [187]. The model of Fig.1.1 represents the data flow from and toward the user through the system. Users perform actions as speak or move an arm, and these are perceived through sensors like microphones or video cameras from the system. The information collected are then analysed at a features level and through the fusion process (described below) and interpreted from the system. Then a strategy for the interaction is applied, result of the interaction design developed. Afterwards, throughout fission, the way back is performed with actions like text-to-speech or image rendering and perceived by the user thanks to sight, hearing etc. Multimodal integration, also referred to as the fusion engine, is the key technical challenge for multimodal interaction systems. In general, the meanings of input streams can vary according to context, task, user, and time. In [134] multimodal interfaces have been classified depending on the fusion method (combined or independent) and the use of modalities (sequential or parallel). In an exclusive multimodal system,

[1] https://www.cnbc.com/2019/09/23/facebook-announces-acquisition-of-brain-computing-start-up-ctrl-labs.html

the modalities are used sequentially. In an alternative multimodal system, modalities are used sequentially but they are integrated together. In a concurrent multimodal system, modal information is available in parallel. Finally, in a synergistic multimodal system, the modes are available in parallel and fully integrated. Synergistic mode should be the goal to be reached, however there are still possible benefits of the other multimodal interfaces. The focus of fusion engine is how and when different input should be integrated [178, 152]. This model is applied to multimodal interaction but it can easily collapse in an unimodal interaction without having multiple input or output.

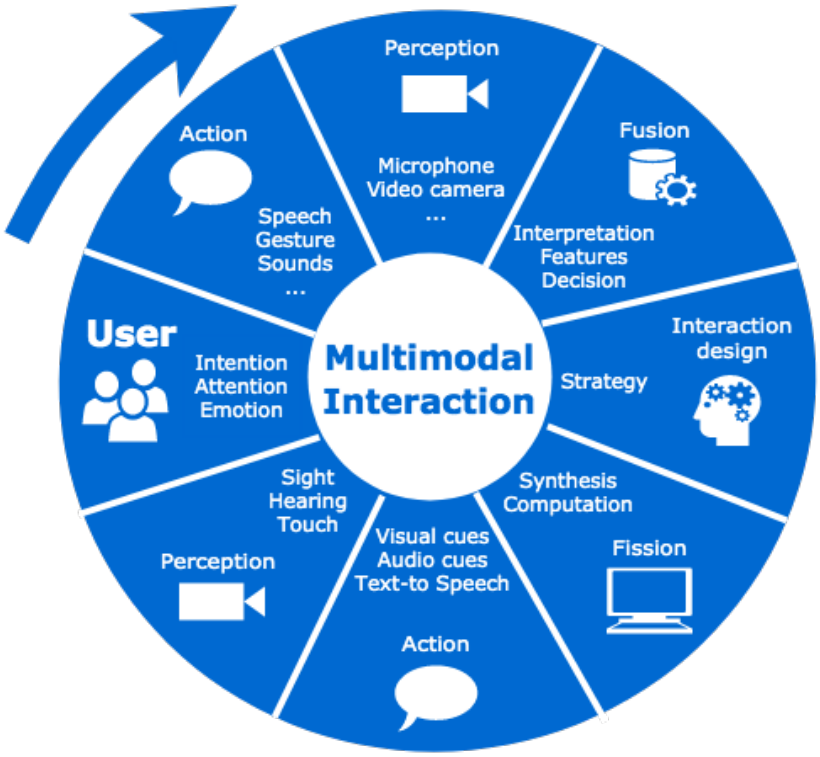


Figure 1.1: Multimodal interaction Model: representation of user interaction model in multimodal contest.

Modern HCI interfaces [96], thanks to the presence of computing power ranging from intelligent walls to hand held interfaces, demand intuitive ways of interaction and human-centered design approaches (approach to problem solving, that develops solutions to problems by involving the human perspective during the entire process). Nowadays, devices such as smartphones and e-books [3] are equipped with advanced signal process-

ing tools and improved hardware, making real-time fusion of data and multimodal interaction possible. Effective interactions are considering different combinations of modality and multisensory approaches such as hand gesture [129], tooth clicks [176], eye blink [69]. Despite the significant progress on multimodal interaction systems in recent years, much work remains to be done. Each unimodal technology (vision-based tracking and recognition, speech and sound recognition, haptics, touch-based gesture, etc.) is an active research area in itself. In the context of multimodal interaction, natural user interfaces have a paramount role, in order to avoid issues relating to cognitive load in multimodal systems, both in terms of what multimodal systems can indicate about a user's cognitive load [31], when people naturally interact multimodally [148]. In the next sections will be presented natural user interfaces and two of the main classes in which is possible to divide unimodal technology: tangible and gestural interfaces.

1.3 Natural user interface

Natural user interfaces (NUI) have been researched since the 1980s, for instance in the project "put-that-there" [18], the author used gestures and voice commands for the control of a GUI. NUI is a type of user interface that is designed to feel as natural as possible to the user. The goal of a NUI is to create seamless interaction between the human and the machine, making the interface itself seem to disappear [195] and enabling it not weigh on users [138]. Natural user interfaces allow a user to use an interface with very little training, as they can draw from experiences from other activities or interfaces. A common example is a touchscreen interface, which allows you to move and manipulate objects by tapping and dragging your fingers on the screen [78]. The digital objects on the screen respond to user touch, similarly as physical objects would. This direct feedback provided by a touchscreen interface makes it seem more natural than using a keyboard and mouse to interact with the objects on the screen [118]. Another modern example of an NUI is a motion-based video game. Microsoft's Xbox Kinect allows you to control your on-screen character by simply moving your body. Today, this technology can be found in other sensors like the Inter RealSense d435, which is even able to detect everyday objects. This motion-based interfaces is considered natural user interfaces

since they respond to your natural motions. Touch screen and body tracking are a good example of tangible and gestural interfaces. If the focus is on combinations of input and output that are experienced as natural, the collection of natural user interfaces includes modes such as gesture and body language, location, eye gaze and the full spectrum of audio and visual output, tactile and other experiences on the "output". A given combination of type of input and output is a multimodal experience that refers to either a combination of more than one output or more than one input as part of the natural interaction. Human interaction with the world is multi-modal, and rich multi-modal interaction is part of what defines a natural experience [97].

The increasing interest in NUI over the past decade has drawn some critical attention to that subject matter. Donald Norman, Northwestern University professor, points out that natural user interfaces are actually not always natural and that sometimes using gestures can in fact become a problem [138]. The nature of gestures is fleeting/invisible and they do not leave any trace, which proves to be a considerable drawback when the user receives wrong response from the system or when there is no response at all. In such situations, it is difficult to reconstruct the gesture as it was read and interpreted by the system. Furthermore, if an interface of an application is based only on techniques which are the components of NUI, the user will need a lot more time to start working with such an application than in the case of a traditional one based on WIMP. In the case of NUI there is no guarantee that graphical elements like menus or icons will be visible in the application. Depriving users of well-known elements could make them reluctant to use such applications, whereas gradual implementation of NUI will allow the creation of new habits and finally it will be possible to replace traditional elements of user interfaces [1].

Tangible interfaces

Based on the work "Tangible Bits" [92] of 1997, the Tangible User Interface (TUI) emerged as a new interface and interaction style. The vision centered on turning the physical world into an interface by connecting objects and surfaces with digital data. Also known as a "graspable user interface" [60], a term that is rarely used today, its primary purpose is to empower collaboration, learning and design through the use of physical

elements. The idea is that by integrating physical elements into a user interface, it promotes a natural user interaction [46]. Tangible interfaces can be described under the definition of Donald A. Normann *Phisicality*: the return to mechanical controls, coupled with intelligent, embedded processors and communication [137]. The simplest example of tangible interfaces are the mouse [196] and the keyboard [140]. More recent interfaces are: i) SandScape [91], which allows the user to form a landscape out of sand on a table; the sand model represents the terrain, which is projected on the surface; ii) Sifteo [127], which allows the user to play games in a mixed reality, on the surface of a set of cubes or iii) WOWCube [144], which allows games to be played in a mixed reality, on the surface of a three-dimensional puzzle.

Nowadays, there are different research areas that are related to and overlap with TUIs. They are part of an emerging generation of HCI, and are Tangible Augmented Reality [114], Tangible Tabletop Interaction [99] with the ReacTable and Microsoft Surface [199], Ambient displays [71], and Embodied Interaction [59]. One area that has received much interest from tangible interface designers is learning (i.e. [157, 203]). This interest is related to the more general view within education that hands-on activity or manipulation of physical objects can be of particular educational benefit [131]. For example, three-dimensional forms might be perceived and understood more readily through haptic and proprioceptive perception of tangible representations than through visual representation alone [67]. While the academic legacy is obvious, the commercial influence is less clear. Several products and interfaces have been launched that incorporate tangible interaction, although industry adoption is not as prominent as the research volume would indicate. One example in recent years is Microsoft's Surface Dial, which was introduced in 2017. However, compared to other forms of interaction such as touchscreens, the amount of successful tangible interfaces is comparatively small. This could have multiple reasons, including the relatively high cost of producing *ad hoc* tangible devices, the lack of general-purpose interaction paradigms, or even simply not enough marketing to the consumer market [84].

Gestural interfaces

Gestures are the motion of the body which is used with an intention to communicate with others. Gestures are a body motion which expresses meaning through movement of body parts such as fingers, arms, hands, head and face. They constitute one interesting small subspace of possible human motion. A gesture may also be perceived by the environment as a compression technique for the information to be transmitted elsewhere and subsequently reconstructed by the receiver [101]. Gestural interfaces have a much wider range of actions with which to manipulate a system with respect to traditional interfaces as mouse, keyboard, etc. In addition to being able to type, scroll, point and click, and perform all the other standard interactions available to desktop systems, gestural interfaces can take advantage of the whole body for triggering system behaviors. For example, the sweep of an arm can clear a screen, a person entering in a room can change the temperature set for the room [164]. However, this kind of interaction systems may lead to some difficulty in understanding the rules to activate specific actions. One example can be the faucet in a public bathroom with a proximity sensor to activate water. Both a design problem and a sensor features can cause misunderstanding between user and how to perform the water activation. So multidisciplinary competence are necessary in order to obtain the needed result [139]. Due to this difficulty, unlike touch gestures, touch-less gestures remain largely a notion developed in science fiction (as in the movie *Minority Report*[2]) and have only been implemented to simpler concept degree in research applications [62, 158, 121, 103], video games (e.g. Microsoft Kinect) and commercial technology [79]. Among the applications based on gesture recognition there are numerous that are based on actions performed by the entire body. These applications are mostly implemented in games (i.e. Nintendo Wii). Thanks to NUI, players have to prove being physically fit, which may have a positive influence on their health. However, the expectations connected with naturalness of the gestures may bring the opposite outcome (i.e. injuries and the destruction of the equipment in the room where the game is being played). This example can be seen as too exaggerated, but it helped to understand the difficulty in NUI development, where the context of use and

[2]Steven Spielberg, 2002

the technology involved play a key role.

1.4 Real-time sound-based input analysis

Nowadays, we have different ways to represent sound by converting its original domain to a simplified one. An example is the translation of difference in air pressure, due to air compression and rarefaction, in an electrical signal. Most microphones functioning is based on Faraday Law which states that "the electromotive force around a closed path is equal to the negative of the time rate of change of the magnetic flux enclosed by the path" [194]. Starting from the signal acquired, it is then possible to visualize it and later perform processing and analysis that allow interaction between the user and the system [116, 32]. Furthermore, this task results even more complicated if we are dealing with impulsive sound, broadband noises or pitched sound if the interaction design process is considered during the interaction development. Indeed, in a context of natural user interfaces, modelling these signal behaviours highlights the difficulty, from the engineering point of view, of extracting the more representative and scalable signal features and developing interfaces that can address the interaction with users.

In recent years, several tools able to extract different audio features, both in time and frequency domain, were developed as MIRtoolbox [111] and Opensmile [54]. Feature extraction starts from an initial set of measured data and builds derived values (features) intended to be informative and non-redundant, facilitating the subsequent learning and generalization steps, and in some cases leading to better human interpretations [75]. Today, Deep learning (DL) is one of the hottest trends in the rapidly growing digital world. It has gained huge successes in a broad area of applications such as speech recognition, computer vision, and natural language processing. [14] DL allows a machine to be fed with raw data and to automatically discover the representations needed for detection or classification. The key aspect of deep learning is that these levels/layers of features are not designed by human engineers: they are learned from data using a general-purpose learning procedure [113]. Even in multimodal and audio applications, the results are impressive [133, 136]. However, even if digital data, in all shapes and sizes, is growing at astonishing rates, in niche fields the are some issue to be addressed: i) data

collection, ii) data labeling, iii) "feature recognition". In supervised learning approach, a significant amount of data are necessary to train the system, this is not always possible in niche fields, since there is no interest in collecting this information. Due to lack of knowledge, the system is not able to correctly generalise the classification keeping high score of correct recognition. Deep learning is not low-cost process [64]. Furthermore, due to the considerable time and expense required in labelling, niche fields should use different approaches like [107], in which learning is obtained by unlabeled data. Finally, in real-time applications, the features extracted are the basis for the interaction more than just the generalization or interpretation of a behaviour. In sensory-based input interaction is crucial to understand which are the features used by the system in order to manage the interaction with the user and tune the right parameters. Therefore, a signal-processing-based approach is preferable for a real time optimization.

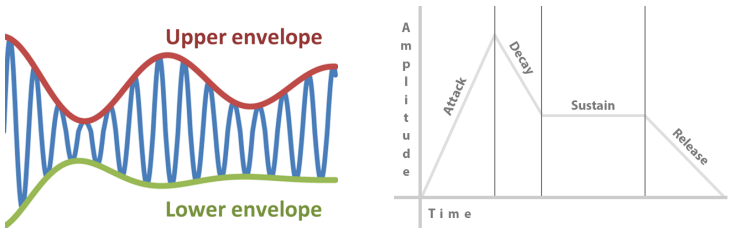
Chapter 2

Blow sensor and Pan Flute Installation

According with human experience, blow is a natural human action when interacting with objects. The blowing action is used by humans to push objects, move them away, to fill something with air, to blow out candles, etc. This action can be extended to interact with natural user interfaces and it allows to interpret actions in order to transmit and transform it in the digital world. There are many benefits of using blowing and other paralinguistic vocal controls, this includes the potential of cross-cultural use, the relatively low cost and the possibility of use by people with motor impairments especially if their impairments are accompanied by speech impairments. Blow signal have been studied in the field of non-verbal input in the HCI field. However, while the non-speech sound inputs have been successfully employed in some applications, in the mainstream market, these applications remain unused. Leaving aside the medical context, the applications of the non-speech sound input can be roughly divided into two categories, depending on whether they are based on the analysis of the acoustic signal in the real time. The real-time applications allow the user to receive feedback while still producing the sound. This is the most interesting case in the context of interaction. The typical use cases are computer games, interactive art installations, or control of the pointing devices. In this last application, several project have been carried on, in order to replace common tangible interfaces, such as the mouse, with accessible technologies. Igarashi and Hughes

claimed that non-speech sounds could be used as a means of intuitive specification of different numeric parameters [89], such as tv remote controls. Furthermore, Adam Sporka proposed U^3I [179], an interface to control a mouse which moves as long as the tone of the gesture is being held in a desired direction, along horizontal or vertical axis at a time. *Vocal joystick* [15] is a research project aimed at investigating non-speech sounds as a virtual input device in different contexts of use (i.e. mouse movements, robot arms control). In the context of cultural heritage and games, sounds, such as pitch, volume, or timbre, can be analyzed and extracted from the signal and studied how they develop over time, then affect the behavior of a game object or an interactive piece of artwork. For example, Sama'a Al-Hashimi created the artworks *sssSnake* and *Blowtter* [2], which are based on a multichannel non-speech sound input. *Blowtter* is a plotter controlled by blowing. The user moves the stylus on the paper by blowing into one of the four microphones representing the cardinal points. Instead, *sssSnake* is a game for two players, where the character of one have to run from the other.

In this project, the aim was to develop a blow signal sensor that could be easily implemented and adapted to different context, scalable and easy to use when trying to connect the precision of medical sensors with not expensive technology (in this project Arudino platforms have been used). Furthermore, the so-developed system have been applied to the context of valorization of museum finds and cultural heritage. In particular, a multimedia installation has been realized through a design process mainly centered on Design Thinking. This installation was realized to valorize an antique Pan Flute from Egypt, approximately dated back to 700 A.D., and exhibited at the Museum of Archaeological Sciences and Art (MSA) of the University of Padova. The installation is presented as a unique wooden furniture outlining two different sections: a first part is devoted to the interaction of the user with the Pan flute sounds through an interface based on "blow sensors" able to recognise the breath; in a second part, where a 42inch screen allows the user to navigate through the retrieved content of the Pan flute. The first part consisted of a stylized Pan flute carved in the surface of the wooden furniture. For each pipe of the artifact a hole has been added. Inside the holes, representing the mouthpiece, microphone sensors were placed in order to detect the interaction of the



(a) (b)

Figure 2.1: Envelope and ADSR of a sound signal.

visitors with the installation, whereas in each pipe is inserted a LED strip in order to provide an additional visual feedback while playing the flute. This is meant to simulate the natural user interaction with the musical instrument: a visitor can blow in each pipe playing the reconstructed sound of the original Pan flute. This virtual instrument was developed starting from (i) the acquisition of the blow signal; (ii) the recognition of the events that defined the flute behavior; (iii) the synthesis of the sound. The architecture of such a system may be used and easily adapted to other projects that aims at translating or interpreting the behavior of an air blow.

2.1 Blow sensor

First of all, it is necessary to outline the main problem and the needs: in order to allow users to interact with a system using their blow, it is necessary to analyse some features that can describe its behaviours. One of the most useful can be the envelope of the signal, which in physics is smooth curve outlining the extremes of an oscillating signal [98]. The envelope of a signal allows to identify four main sections of a sound: attack, decay, sustain and release. A "blow signal" is characterized by these trends and the system must be able to identify them.

Attack the time reaction of the system for rapid changes in input signal is an important aspect for real-time responses;

Decay time that signal takes to pass from attack to sustain volume;

Sustain in a long sound it represents the most percentage of a signal;

(a) Medical breath sensor

(b) Propeller for anemometer

Figure 2.2: Examples of breath and wind sensors.

Release time to stop signal after the end of an excitation, as the previous one, it is an important indicator of will user fidelity.

The system must be able to act correctly according to these four trends achieving the best trade off.

The most common sensors that can be used for this application is a breath sensors and anemometer. The first one is used in a medical context to accurately analyse features of patients' breath. Its purpose could bring high fidelity to the system. Some of these sensors are used to monitor the breaths of people in a coma. In spite of measuring precision, the main drawbacks are price and usability. Indeed, high precision and accuracy in the decision of materials rise the price of sensors. Furthermore, usability issues are due to the need of keeping a thin cane between lips. This rises the fidelity but represents an hygienic issue that can be an obstacle to scalability of the sensor. Anemometer does not have this issue, since propellers just need an air propulsion. Of course the precision decreases due to this trade off however, the propellers continue to spin after the end of the propulsion, causing the impossibility to precisely determine the end of the blow (problem with the distinction between sustain and release).

Piezoelectric sensors are based on the principle of creating a voltage at the terminals of a circuit in response to applied mechanical stress. The hypothesis were to identify the pressure on the surface generated by the blow, and the results were very surprising. After applying an amplification of $gain = 100$ in the circuit, it was possible to identify either the right begin and the end of the events. The sensor is not expensive (cents of Euro)

but the accuracy is less than medical sensors. According to the applications in which the sensor can be used, the sensitivity of the sensor was representing the right trade off. The signal acquired from the piezoelectric sensor is shown in Fig.2.4. One of the main issue was the necessity of a time varying force, that was causing a continuous variation on the field generated from the material of which it is composed. During constant forces, ΔV was not generated, therefore in Fig.2.4 it is possible to see values equals to 0 due to the stability of air pressure that can occur during a blow. This issue prevented a right representation of decay and sustain sections of the signal, since it was very difficult to figure out the envelope of the signal during time. The circuit used to amplify the signal was the one in Fig.2.3 that consisted in inverting amplifier circuit with a single power supply operational amplifier. To avoid the "holes" in the signal of Fig.2.4, another

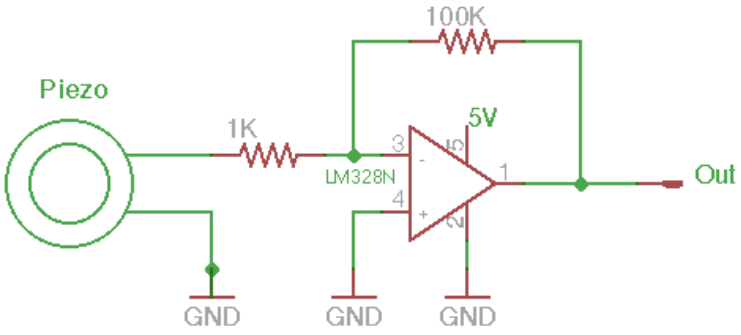


Figure 2.3: Circuit used for piezoelectric signal amplification.

kind of piezoelectric sensor was tested. Fig.2.5 shows a *film* piezoelectric sensor and it generates a voltage difference at the terminal of a circuit when its structure is bended. This sensor was chosen for raising the "density" of values different from 0, since its physical structure allows a more reactive response to small variations in pressure. The results of the first test confirmed the hypothesis but introduced some non negligible factors: size and small rejection external strokes. For example in Fig.2.5 there are shown the dimensions of the sensor that are in the order of few centimetres, very big dimensions compared to usual sensors sizes.

Microphone sensors can be really small (few millimetres) or in the order of 1 *cm*. MEMS microphones were used in a preliminary study, but one of the main issues turned

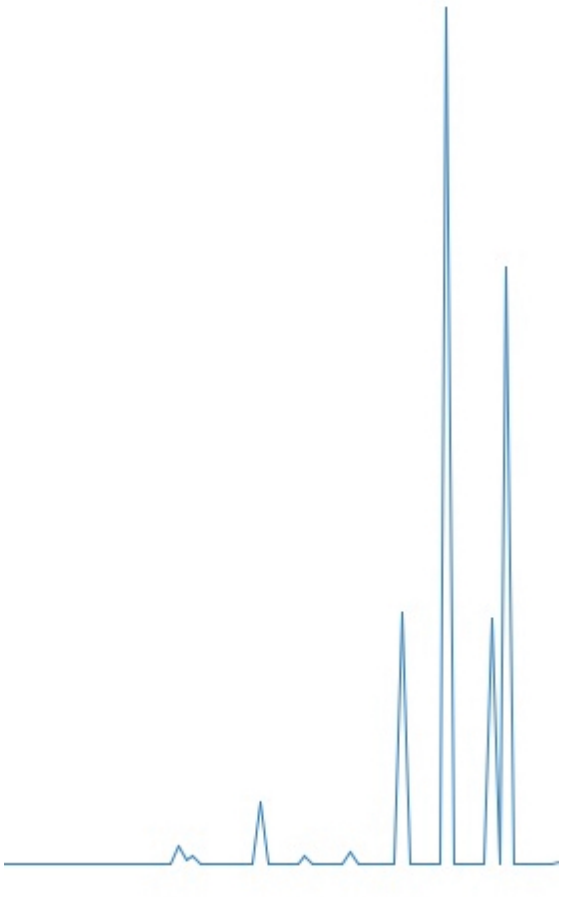
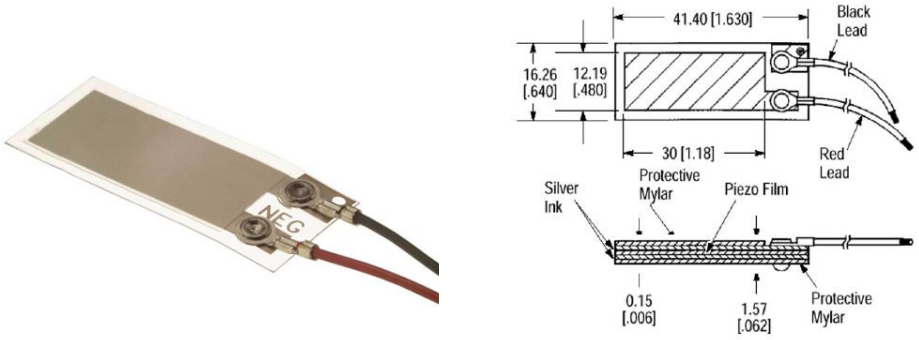


Figure 2.4: Signal received from Arduino from a piezoelectric sensor.

out to be the obstruction of the microphone hole. To solve this potential problem, a bigger microphone ($\sim 1cm$ of diameter) was tested. Usually, a microphone needs a double power supply in order to exploit either positive or negative voltage of the capsule. Alternatively, it is possible to use a single power supply, rising to $V_{IN}/2$ the voltages at the positive terminal of the operational amplifier using a voltage divider configuration (same value for the two resistances). Arduino can provide $5V$ or $3.3V$ and this dictates the use of this technique. The signal has an average of $2.5V$ but the value is not constant since it oscillates with random values and it can introduce a non-negligible uncertainty in recognizing small signals. Furthermore the input range of the signal is halved since now the signal can vary from $+2.5V$ to $-2.5V$ (see Fig.[2.6]).



(a) Film piezoelectric sensor (b) film piezoelectric sensor dimension

Figure 2.5: Example of film sensor.

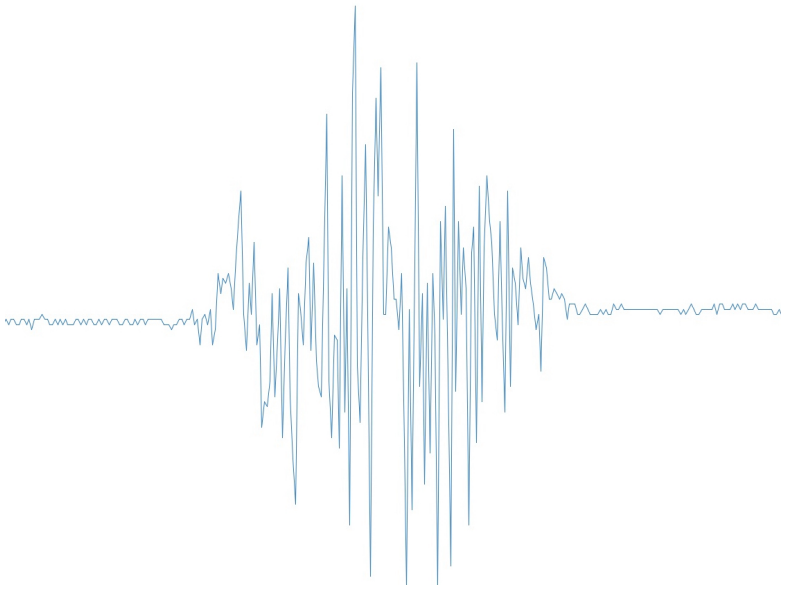


Figure 2.6: Signal of the microphone blowing on it. In the conditional circuit, a voltage divider has been applied.

2.2 Blow sensors applied to the Multimedia Installation

As previously described, the four trends/sections of a sound signal must be found and interpreted. In our case study, they were represented and synthesised using three main messages of MIDI protocol: NoteOn, ControlChange, NoteOff. The first started the reproduction of the sound with a value of velocity, the second varied the interpretation of the blow and the third stopped the reproduction of the sound. After testing the different sensors, either with piezo or microphone sensor, the signal needed to be processed to clarify the transients describing the begin and the end of the interaction and the modulations during the blow. In fact, in Fig.2.4 and Fig.2.6 the signal varies rapidly and with big variance, so it was very difficult to identify the envelope of the signal. Identifying the beginning and the end of the interaction using a threshold was a simple solution (when the signal passes over the threshold and when it returns below), however the nature of the signal prevented it. Finally, a moving average window and a rectifier to reduce the oscillation and interpret the envelope of the signal have been used. The resulted signal was constant to 0, when no interaction occurred, and displayed only the positive values when blowing on the microphone. Applying this processing to the signal, the result is shown in Fig.2.7. The result obtained with the microphone during experimental tests allowed to use a smaller window size with respect to the one used with piezoelectric sensor during the digital processing of the signal. This helped to have a faster response of the system since longer windows imply longer time to fill the buffer. The sensor used for the project case study was the electret microphone and the circuit used is shown in Fig.2.8.

2.3 The Pan Flute

The artifact (shown in Fig.2.9) arrived in Padova thanks to Carlo Anti, who directed the Italian Archaeological Mission in Egypt since 1928. Carlo Anti held the position of Rector at the University of Padova, and contributed to renovate and modernize the university and its buildings. Among them, Palazzo Liviano, which is the current location of the MSA. The archaeologist also led excavations in the ancient village of Tebtynis in

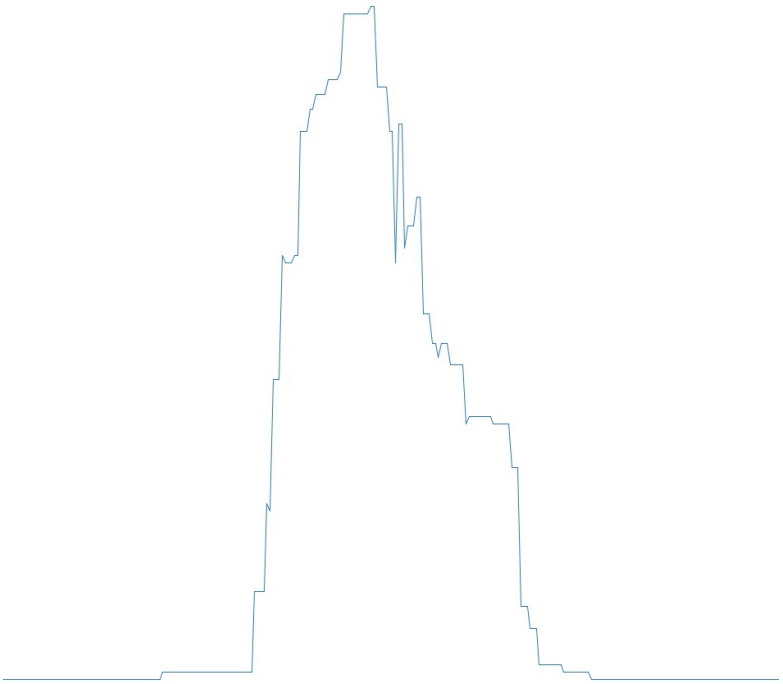


Figure 2.7: Smooth signal of a microphone with single power op-amp.

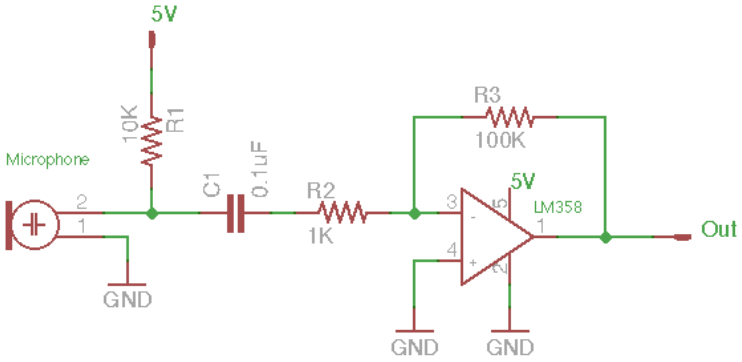


Figure 2.8: Circuit used for electret microphone sensor.

the Fayum oasis, from 1930 to 1936. At the moment, the origin of the Pan flute is not completely certain. This ancient musical instrument was stored in a box, originally made for photographic plates belonged to Gilbert Bagnani, Italian-English archaeologist who assisted Anti. The box cover reports a French sentence in the tiny handwriting of Bagnani's wife, which sets the original finding in Saqqara, in the area of the Mastaba n. XV, thus near Pepi II's tomb [8]. Further information is found in Anti's archive

Figure 2.9: The restored Pan flute (frontal and posterior views).

and in a letter written by Evaristo Breccia (Director of the Archaeological, Museum of Alexandria), where he asked Anti about this instrument which he saw during a visit in Tebtynis. This hypothesis of origin is supported by the presence in Padova of other antiquities from Bagnani's campaigns, stored in small boxes similar to the Pan flute one, and unlike other archaeological materials. Except for a few exceptions, the findings were recovered at the MSA in 1935, therefore this probably corresponds to the year of discovery of the Pan flute.

2.3.1 Pan Flute tuning estimation

It is known that the internal lengths of the pipes are reduced by carefully increasing the thickness of the closed ends through the addition of wax or propolis, in order to fine-tune fundamental frequencies [33]. Despite the restoration of the ancient pan flute, some pipes are still partially obstructed, therefore the interior of these pipes is not completely visible and not directly inspectable. In a previous work [9] a preliminary estimation of pipe lengths was obtained from external measures taken on a laser-scanned 3D model. In order to refine these measurements, computerized tomography (CT) scan was used here.

Specifically, in order to determine fundamental frequencies of the pipes, two measures were estimated: internal length and internal diameter. The three dimensional image of the interior of the instrument is obtained with a GE LightSpeed VCT 64 Slice CT scan. The scanning was then read with the open-source software Horos, a medical image viewer which also provides tools to extract reliable measures from the CT scan. In order to browse inside the three-dimensional image, and to perform precise measures, the operators alternated two different views: a 3D MultiPlanar Reconstruction (MPR) and a 2D orthogonal MPR. Figures 2.10(a), 2.10(b), and 2.10(c) show the latter view and the three orthogonal planes, defined as axial, coronal and sagittal, respectively. Since some parts are damaged or corrupted, a total of eighteen measurements for every pipe were collected, with the goal of obtaining more robust estimates.

With regard to the pipe lengths, six measures were extracted from axial and coronal planes. Pipe openings are not straight, they are slightly u-shaped at one side in order to provide an embouchure to the player (the opening shapes can be observed in Fig. 2.9, posterior view): therefore, the difference between the maximum and the minimum point of the opening was measured on both planes. Moreover, the internal shapes of the closed pipe ends are also not straight: therefore, for each plane, the maximum and minimum internal lengths were measured (Fig. 2.10(b) shows an example of measuring the maximum length of a pipe in the coronal plane).

With regard to pipe diameters, twelve measures were collected. One measure was taken in the axial plane and a second one in the coronal plane, whereas two measures were taken in the sagittal plane (one for each axis of the pipe, see Fig. 2.10(c)), because pipe sections are oval-shaped rather than circular. These four measures were repeated at three different levels: near the opening of the pipe, at the mid point and near the closed end (Fig. 2.10(a) shows a diameter measure at the mid point in the axial plane).

The measurement process highlighted several issues that required some subjective interpretations by the operators. The longest pipe, for example, is broken and curved, therefore a specific tool of 3D Curved-MPR was then used so that, given a set of reference points on the curve, would virtually straighten the pipe, providing a more usable view for the correct measurement. In other cases, the presence of obstructing materials impaired

a correct evaluation of the internal surface of the pipe. This is the main reason why redundant measurements were taken. Nonetheless, for the shortest pipes it was not possible to measure directly the position of the openings because these pipes were more heavily damaged. Consequently, opening positions were estimated from the neighboring pipes. The error that mostly impacts the measures was the CT scan resolution: every voxel (volumetric pixel) is isometric and it measures $0,625$ mm. All these difficulties affected the accuracy of the measures, however using redundant measurements provided a range for a plausible estimation of lengths and diameters.

2.3.2 Tuning

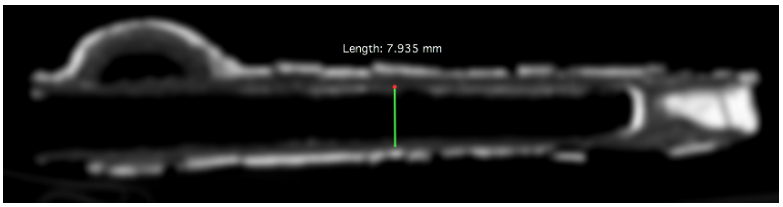
The measurements of the internal length and diameter of the pipes were used to estimate their fundamental frequencies, under the assumption of ideal open-closed cylindrical pipes:

$$f = \frac{c}{4(l_{\text{int}} + \Delta l)} \quad \text{Hz}, \tag{2.1}$$

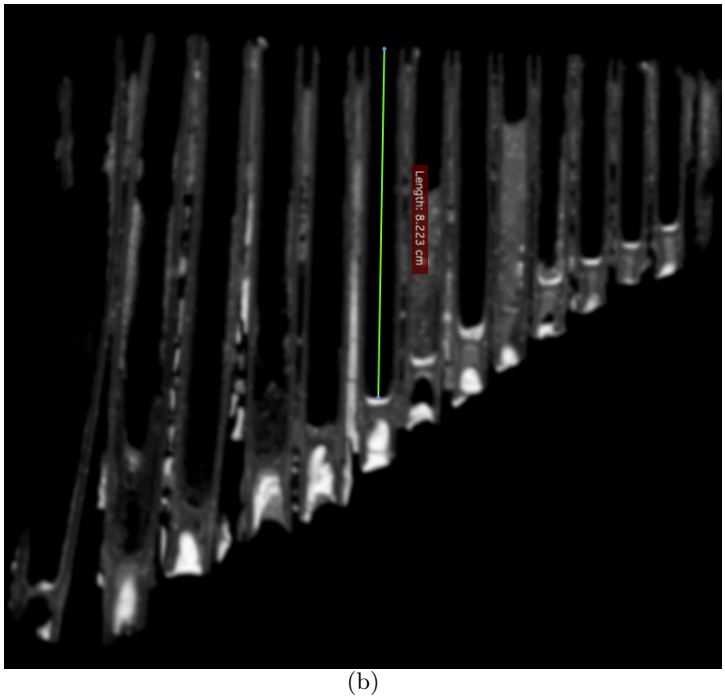
where c is the sound velocity, l_{int} is the internal pipe length, and $\Delta l \sim 0.305 d_{\text{int}}$ is the length correction at the open end, which is proportional to the internal pipe diameter d_{int} [61]. As the measurements are affected by the errors reported in the previous section, for each pipe we considered the minimum and maximum values of internal length and diameter.

Then, in order to take into account the effects of error propagation, an interval of values was calculated for each pipe as an estimate of the fundamental frequency: in particular, f_{min} was calculated from Eq.2.1 using the maximum values of length and diameter, whereas f_{max} was calculated from the corresponding minimum values (see Tab.2.1).

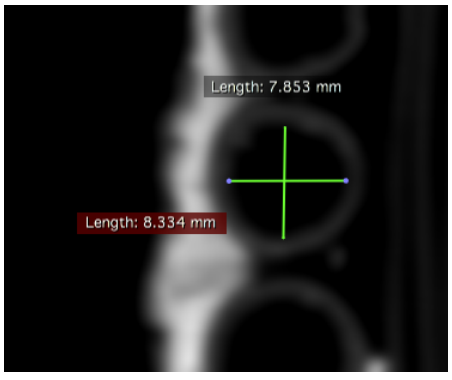
It is interesting to verify whether these frequency ranges are compatible with predictions derived from music theory. According to theorists [76], the ancient Greek music system was based on the *tetrachord*, i.e a group of four notes (often associated to the four strings of the lyre or the kithara) where the ratio between the pitches of the fourth note and the first note is equal to $4:3$, namely a perfect fourth. Fig. 2.11 shows the pitch ratios calculated as $f(n+3)/f(n)$ for $n = 1, 2, ..., 11$, where $f(n)$ is the fundamen-



(a)



(b)



(c)

Figure 2.10: Views from the CT scan; (a) example of diameter measurement on the axial plane; (b) example of length measurement on the coronal plane; (c) example of diameter measurement on the sagittal plane.

pipe	$f_{min}[Hz]$	$f_{max}[Hz]$
1	638.7	649.7
2	677.2	700.7
3	753.6	773.5
4	843.1	874.4
5	928.3	974.7
6	1010.1	1041.3
7	1142.2	1184.3
8	1283.2	1346.4
9	1389.6	1438.2
10	1538.3	1602.0
11	1721.8	1758.1
12	1901.4	1957.3
13	2128.4	2205.1
14	2292.9	2499.7

Table 2.1: Fundamental frequencies (min and max) estimated for each pipe starting from the measurements taken from the CT scan.

tal frequency of the n^{th} pipe. Due to error propagation, for each pair of pipes a range of values (reported with the vertical lines) was obtained. Comparing the ranges with the horizontal line representing the 4 : 3 ratio (dot-dashed line), it is possible to see that all the intervals are compatible with the tetrachord definition. It is known that the tetrachord is subdivided into three pitch intervals that can have various configurations. In particular, three *genera* can be distinguished: diatonic, chromatic, and enharmonic. As an example, the diatonic tetrachord is characterized by intervals that are less than or equal to half the total interval of the tetrachord. Usually, this tetrachord begins with one small interval followed by two larger intervals, corresponding approximately to a tone (9 : 8). Fig.2.12 shows the pitch ratios between adjacent pipes, i.e $f(n+1)/f(n)$: it is possible to recognize some intervals that are compatible with a tone (9 : 8) and other smaller intervals compatible with what some theorists call *diesis*, corresponding to the ratio 256 : 234. Two tetrachords can be joined following two different schemes, called *synaphē* (conjunction), when the top note of the lower tetrachord corresponds to the bottom note of the higher one, and *diazeuxis* disjunction, when there is an interval of a tone between the tetrachords. Observing the sequence of intervals of Fig.2.12, some joint tetrachords can be recognized: e.g., the pitch of the first eight pipes are compatible with two disjoint tetrachords, as represented in Fig.2.13.

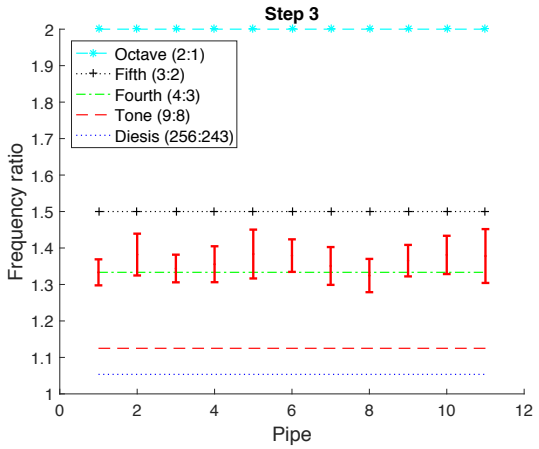


Figure 2.11: Pitch ratios calculated as $f(n+3)/f(n)$, where $f(n)$ is the fundamental frequency of the n^{th} pipe. The horizontal lines correspond to the basic theoretic intervals.

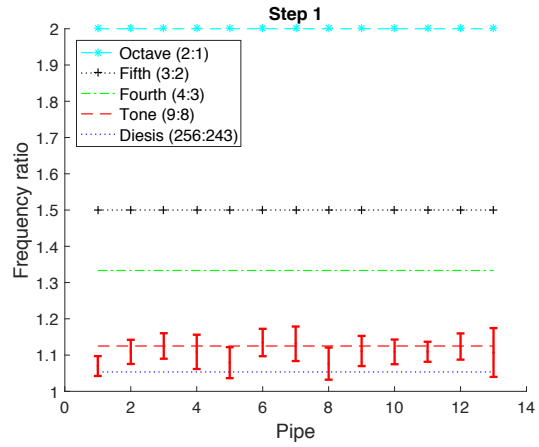


Figure 2.12: Pitch ratios calculated as $f(n+1)/f(n)$, where $f(n)$ is the fundamental frequency of the n^{th} pipe. The horizontal lines correspond to the basic theoretic intervals.

2.4 The Multimedia Installation

2.4.1 Pan Flute Audio Samples

In order to implement the virtual instrument, it was important to decide between the two main techniques of synthesis since they will drive the next steps: the model of the source (physical model) or the model of the signal (wavetable) [34]. The first one allows a very precise and accurate definition of the physics of the instrument. However, the interaction with the model needed basic knowledge on the way of playing the instrument which might not be always obvious for the user (for example the inclination of the mouth over the sensor, the minimum air pressure necessary to trigger the natural resonance of

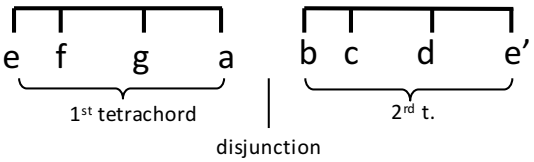


Figure 2.13: Schema of two disjoint tetrachords, compatible with the pitch of pipes 1-8. The letters used to represent the notes do not correspond to modern pitches.

a pipe, etc.). The second model ensured good reliability to the synthesis of the original sound of the flute by maintaining, at the same time, a mode of interaction accessible to all kind of users. This last option was used since the aim of this section of the installation was to allow user to naturally play with the installation. Wavetable technique implied the use of a software for reproducing the audio, which could manage some parameters to better interpret the signal coming from the sensor. The software used was Ableton Live 9, which allowed to simply interface the Blow sensor through MIDI messages.

2.4.2 Visual feedback

In the context of multimodal interaction, the installation included an element to make more interactive. A visual feedback provided by a led strip was inserted in the stylized reeds of the installation. In order to have a faithful response, the light was intensity modulated according to the intensity of the sound. In order to obtain this effect, the fifteen digital pin of Arduino Mega that was user to provide a PWM signal/modulation to power the led strips.

PWM is a signal that changes its value from $S = 1$ for part of its period, to $S = 0$ for the rest. The duty cycle D refers to the percentage of the period for which the signal is on. The duty cycle can be anywhere from 0, the signal is always off, to 1, where the signal is constantly on. A 50 D% results in a perfect square wave. The average value of voltage (and current) fed to the system is controlled by turning a switch between supply and load on and off at a fast rate. The longer the switch is on compared to the off periods, the higher is the total power supplied to the load. Using this system beside that power loss in the switching devices was very low, by modulating the signal changing the duty cycle it was possible to change the current flowing to the strip LED and to vary

its intensity [83]. The maximum current that Arduino can feed by the digital port is 40mA that was not enough for a led strip. So it was necessary to use an external power supply. To merge it with the PWM modulation it was necessary to add a BJT that would intermittently open and close the circuit that was connecting the power supply to the strip LED according to PWM frequency (duty cycle). In the circuit shown in Fig.2.14, the BJT switched from saturation to cut-off mode by closing and opening the negative terminal of the strip to ground. The polarization circuit ensured the flow of a current

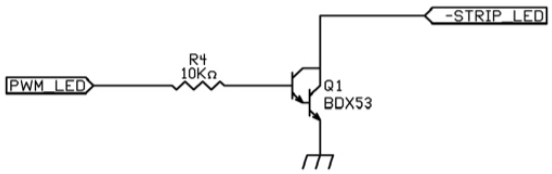


Figure 2.14: Circuit used for strip LED

in base in order to start the conduction, and it was provided by PWM modulation of a digital port ($I_B > 0$).

2.4.3 PCB

After the first part, when all the software and the test were completed, a PCB of the circuital part was developed. This was possible thanks to the collaboration with Lanfranc s.n.c. whereby Gerber file were designed within a collaboration for the project. The complete schematic is shown in Fig.2.15. All the incoming and outgoing connections are connected to plastic component provided hooks able to hold terminal parts of the cables, avoiding the needs of soldering them. The two sides of the PCB are shown in Fig.2.16. Since the PCB must be placed on the mouthpiece (diameter 3 cm) of stylized flute, it can be partially seen below it. To avoid it, the PCB back side was covered of a black film and arranged as upper side. The front, with all the components, except for the microphone, becomes the back side, thus not visible. As partially said, the microphone is the only component soldered on the black side just below the mouthpiece to be reached by the blow. The final result is shown in Fig.2.17.

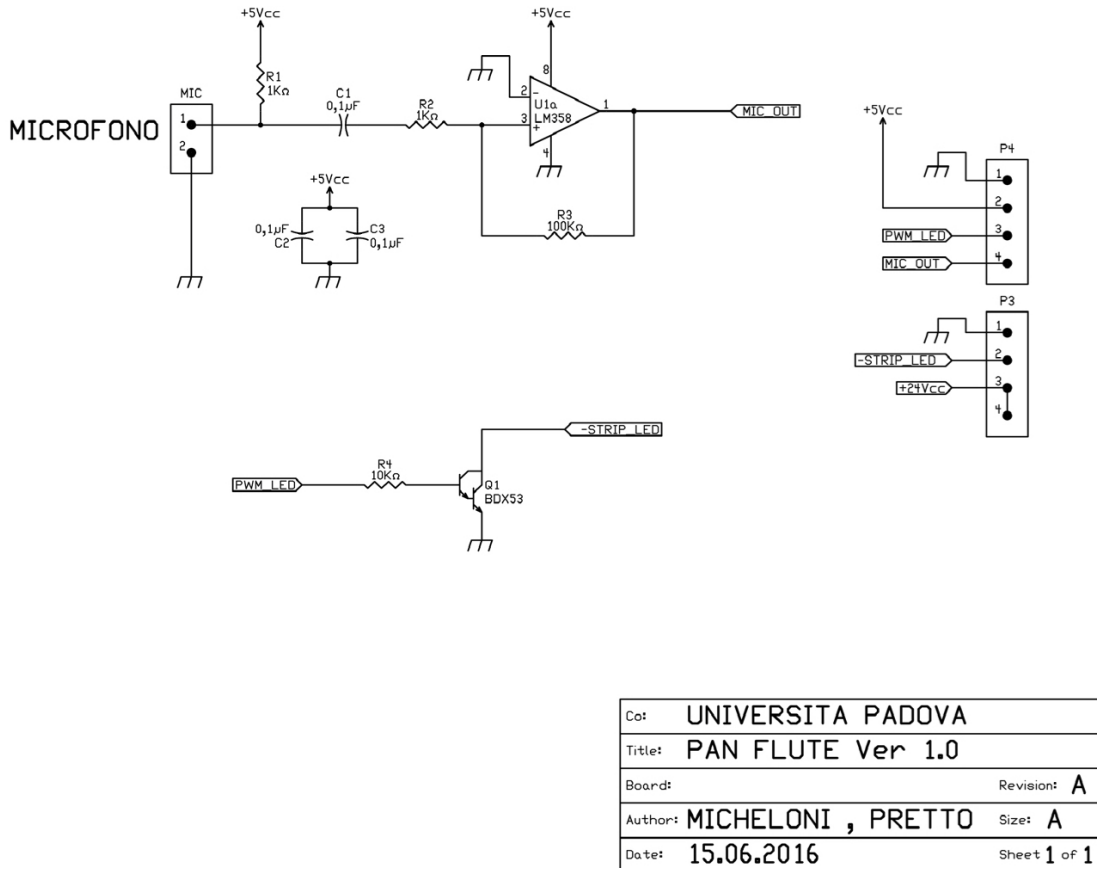
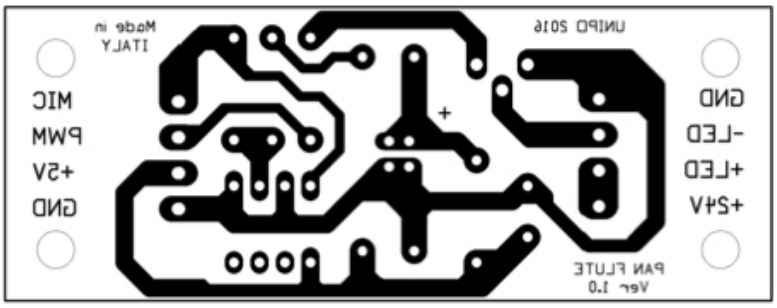
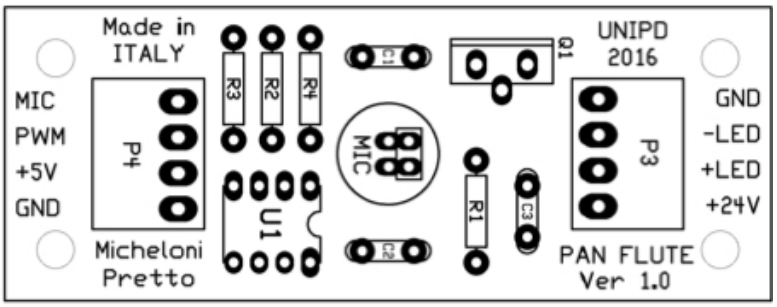


Figure 2.15: Complete schematic of the circuit.



(a) Back side of the PCB



(b) Front side of the PCB

Figure 2.16: Front and back side of the PCB.

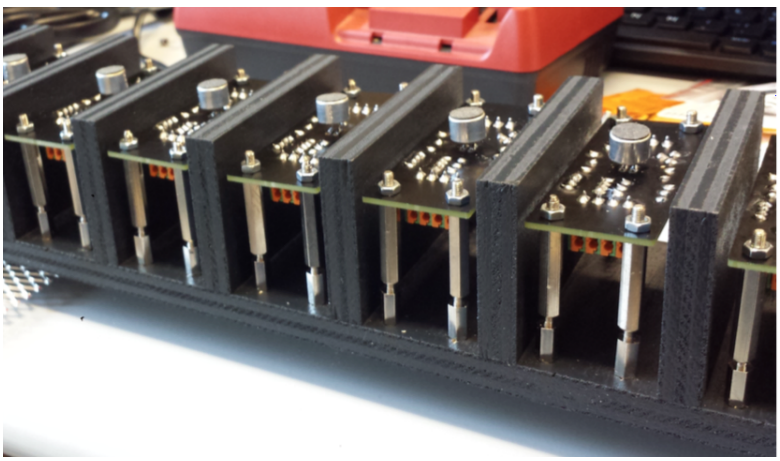


Figure 2.17: Image of the PCB and sensors location.

2.4.4 Touchscreen section

The second part of the multimedia installation consisted of a 42" touch screen, connected to a pc, embedded in the furniture. The application was developed using the framework Unity 3D[1]. The information was subdivided in five different chapters reachable through a menu: Myth, History, Sound, Flute, and 3D. All sections had in common elements, such as the font (Titillium), the chromatic range and the menu bar at the bottom of each screen, whereas some other interchangeable features were customized on the basis of contents. In the first section, several comic strips shown the mythological story of the Pan flute. Through simple swipes, the visitors could explore the excerpts of Ovid's Metamorphoses and their related illustration. The section Sound provided an alternative interaction with the virtual Pan flute. A stylized musical instrument was visualized and could be played by simply touching its pipes. The physical and virtual sound parts were developed by mutual exclusion, so it was impossible to play both at the same time. Unlike the virtual instrument that uses the blow as input, the touch version did not allow to change the sound level. The section History proposed a meaningful part of the European literary and iconographic sources collected by archaeologists involved in the project. Each source was disposed in a temporal bar that covers the range between VIII century B.C. and the X century A.D. By selecting a mark of the desired source, a pop-up window showed images or texts. Each source was also geographically located on a map of Europe. The Pan Flute section was composed by four thematic subsections:

- the archaeological information about the discovery of the Pan flut;
- the cultural and musical context in which the instrument was produced;
- the study concerning the sound of the musical instrument
- the reconstruction of a similar copy of the archaeological find applying experimental archaeology

The last section consisted in two parts concerning the virtual 3D model and the Computerized Tomography (CT). The raw model of the flute was fulfilled using a texture

[1] unity3d.com/

obtained from several photos. With well-known gestures on the touchscreen, such as swipe and pitch-to-zoom, it was possible to rotate and thoroughly examine the exterior peculiarities of the musical instrument. The second part provided a different approach to the exploration of the CT. A set of markers placed on a bar enables to discover all the peculiarities of the artifact. With a simple click on the marker, the CT browsed the three orthogonal views. As soon as the selected peculiarity was shown and highlighted, a description appeared on the left of the CT model. The resulting multimedia installation can be observed in Fig.2.18. It was located in the museum, where it is now permanently exhibited. After a while, a new re-collocation of the multimedia installation was required. With a large number of visitors, the proximity of the artifact constitutes a danger for the artifact itself. After the methodology design, the assessment of the methodology and the impact of this choice on the overall evaluation are shown in the Assessment section.

Figure 2.18: The first realization of the multimedia installation.

2.5 Methodology of Design

In the field of museum exhibitions, especially the interactive ones, the design methodology is often interchangeable and adaptable to the final scope of the exhibition itself. For example, Lord at al. [117] proposes an approach called *VX Principles*, centered on visitor experience, while Falco at al. [56] suggests an approach based on interactive storytelling. The methodology at the base of the design process (Figure 2.19) was mainly centered on *Design Thinking* (DT) [22], although it develops the phase Define in a different way, and it considers relevant the user experience as much as the promotion of the exhibition object. This approach evolved from theories such as Participatory Design [166], Co-

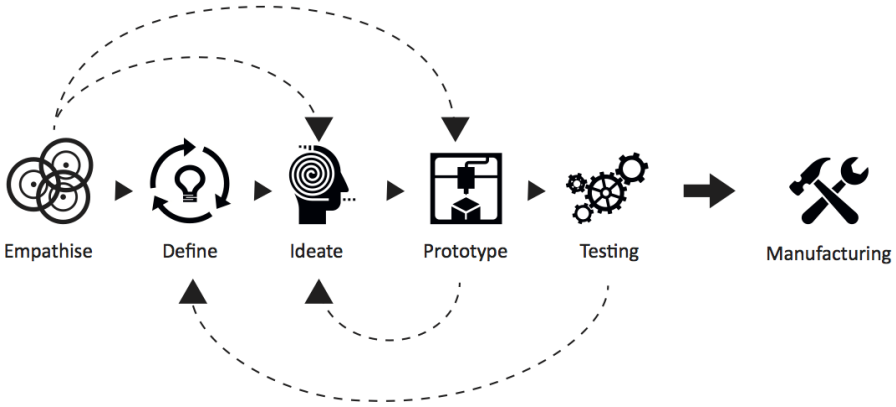


Figure 2.19: The visual representation of the Design Thinking process

Design [167], and Human Centered Design [35], and formed at the Stanford University and at the Ideo Corporation[2] during the Nineties and the beginning of the 21st Century. DT is a process that may solve complex problems and it generates innovative ideas assuming that each person can give a personal contribution related to a specific knowledge area. This multidisciplinary approach,usually used in different innovation challenges, such as Social Innovation, Product Design Innovation, Service Innovation etc., may be considered appropriate for the development of interactive multimedia installations so as to convey the importance of historical artifacts. The mutual exchange of ideas among the several professionals involved and their respective distinct viewpoints helps the DT process in achieving better results.

DT is frequently applied to a new design brief or challenge and it is commonly composed of six steps: Emphasize, Define, Ideate, Prototype, Testing and Manufacturing. This process is not necessarily linear but, according to the needs of the project, it allows the participants to move along the process and to go back to the previous phases or jump to the following ones. Such process may be repeated several times until the project is completed and ready for production.

The first step (Emphasize) consists of the analysis of the issue from several points of view, mostly involving the user. In design culture, the goal is to gain "an empathic understanding of the people you are designing for and the problem you are trying to solve" [38], and it may involve different approaches, from brainstorming to interviews

[2] An international design agency. Website: `ideo.com` (Retrieved January 25, 2019)

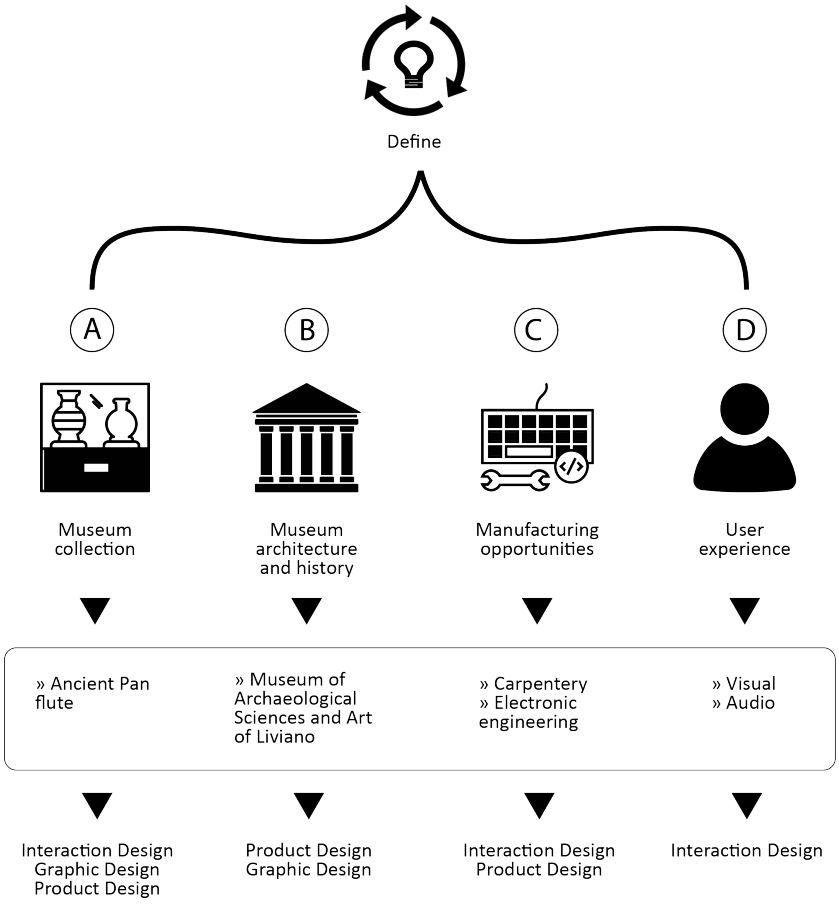


Figure 2.20: The four stages of the phase Define applied to the Pan flute project.

[22].

The second step of the DT process (Define) concerns the definition of the "problem" or "challenge" by considering different aspects of the project. This stage, in the context of interactive installations for museums, may be divided in four different sections (Figure 2.20): (A) the museum collections (or a single artifact), (B) the environment, (C) the manufacturing opportunities and (D) the user experience. In describing this general methodology, the term "museum collection" will be maintained, but the methodology can be used also for a single artifact as in the case of the Pan flute described in the following sections.

The first analysis (A) concerns the study of the museum collection to complement, in order to determine the piece of information that should be emphasized. This step involves all of the three design disciplines (ID, PD, and GD): ID to define the type of interaction, which could be, for example, tactile, visual, auditory, spatial or multisensory;

GD to define the communication tools and their stylistic characteristics; PD, involving the shape definition of the ts in the collection.

The second step (B) consists of the analysis of the environment (peculiarities, style, rules or limitations) in which the collection and the multimedia installation will be located. In this step, PD and GD are necessary tools as the installation aesthetic form should be consistent with the environment, and the definition of the right visual language should take into account the one adopted by the museum.

The third section (C) aims at identifying the true manufacturing opportunities, understanding the engineering skills of the team, finding potential partners, and determining the economical resources which may be allocated for the task. In this step, ID and PD are used in the development of the installation structure and interaction.

The last part of the process (D) consists in the user experience analysis. Designing an interaction entails imaging how it is possible to interact with an object. This way, the last phase of the analysis takes place considering the user's point of view. According to [50], "in designing artifacts we do not merely design the artifacts themselves: deliberately or not, we also design conditions for their human use". In this phase it is important to include an analysis of the museum public: visitors may vary in age, skills and knowledge, and this differences need to be take into account in order to create a reliable communication channel. The conventional visitor of a museum can be considered as an "audience" (people who like to watch and expect to be entertained). However, in this case, the goal was to convert the audience into "players" (people who want to enjoy themselves), and "participants" (people doing something, whether fun or not) [112].

The third phase of the design process (Ideate) is the definition of the idea, mostly exploiting the collaboration of the several types of stakeholders involved, and the use of ideation methods, such as brainstorming, sketching or co-creation workshops.

Once each aspect of the project is developed, the prototyping phase takes place. Soon after that, both the Testing and Manufacturing phases will simultaneously start. As a matter of fact, the DT methodology is conceived to be used in industrial design productions, service design or for improving innovative ideas, and the process may be adapted to many different purposes. In the case of multimedia installations, the final

output is a single product designed for a single museum or exhibition. Therefore, the testing and the manufacturing stages may be carried out at the same time.

2.6 Assessment

Once the installation was manufactured and in place, an assessment have been performed to measure the result of the project. More specifically, it is important to measure the quality of the user experience with the installation but, most of all, to have feedback on whether the decisions taken during the design phase, guided by the methodology previously described, had been appropriate or not. To this aim, the assessment included two self-developed questionnaires: one on the user experience, and one on the methodology and design. It is well known that several methods have been studied to evaluate the user experience, and the design as well [191], thus many studies are based on self-developed questionnaires [11], which makes comparison problematic. Even if the field was restrict to the evaluation of the interactive experience in a cultural heritage environment, different evaluation procedures [105] in the literature are available. While an effort towards standardization and generalization must undoubtedly be made, the literature includes:

- different environments (museums [10, 37, 87, 119, 202], public spaces [86], exhibitions [150], not to mention evaluations in a controlled setting [42]);
- different users (e.g., children [202], adults [10, 37, 150], or both [86, 119, 160]);
- different devices to interact with (e.g., hand-held devices [37, 87], virtual reality devices [150], multimedia installations with touch displays [10, 86, 202], 3D printings of artifacts [42].

As a consequence, a point may be made that distinct evaluation procedures would be required. As far as the assessment in the present paper is concerned, the interactive installation of this project bear similarities to the ones evaluated in [10, 202]. However, such installations do not offer any possibility of interaction by blowing an experience that definitely needed evaluation, and the assessment performed is much broader in scope: what eventually came to be mainly evaluated were the user experience but also

the soundness of the design methodology. To this aim, customized questionnaires based on the target have been realized.

View that this project focuses on methodological and technical aspects, general public was not involved in the assessment. Instead, a group of researchers and professionals with experience in the fields of engineering, music, musicology, and history have been selected. Given the fact that the project is highly multidisciplinary, people with experience in (or, at least, an interest in) more than one field was preferred. In what follows, they will be referred to as *experts*. The group of experts included twenty-three people with an average age of about fifty-two years: three of them were less than forty years, fourteen were between forty and sixty years, and six were over sixty years old. Sixteen of the experts were University professors, and seven were researchers or freelance professionals. The experts were interviewed over a period of seven weeks.

A pre-established sequence of steps guided the visit of each expert at the museum in the installation interaction, and in the filling in of the questionnaires. First, the expert was welcomed and walked to the area of the Pan flute. The expert was allowed to observe the real flute and then freely interact with the installation. Interaction time was covertly recorded. As little information as possible was given at this time. Concerning the installation, the expert was simply told that two parts were accessible from the front panel: in the former, interaction was possible by blowing; in the latter, by the use of the touch. The expert was instructed to postpone any questions at a later time, and act as if she was visiting the museum alone. When the expert finished interacting, a first paper questionnaire about User Experience was administered to them. After the first questionnaire, an open discussion session was held. The expert was explained the design methodology of the installation, with particular reference to the second step of the DT process. Pending questions (e.g., about the flute, the installation, or the methodology) were answered. When the discussion was over, the expert answered a second questionnaire about Museum Architecture and History, Museum Collection and Manufacturing Opportunities (see Figure 2.20). The questionnaires prepared were based on the Likert method: they contained a list of statements, and the expert was asked to indicate how much she agreed with each of them using a 5-level scale from 1 (strongly

Statement	Avg Score	Std Dev	Min	Max
1	4.91	0.29	4	5
2	4.78	0.42	4	5
3	4.09	0.68	0	5
4	3.61	0.92	0	5
5	2.78	1.57	0	5
6	4.52	1.36	2	5
7	4.00	0.85	2	5
8	4.74	0.45	4	5
9	4.78	0.42	4	5
10	4.43	0.59	3	5
11	4.65	0.57	3	5
12	4.87	0.34	4	5
13	4.48	0.59	3	5
14	4.43	0.66	3	5
15	4.48	1.04	1	5
16	4.57	0.73	3	5
17	4.65	0.57	3	5
18	3.26	1.10	0	5
19	3.43	0.99	2	5

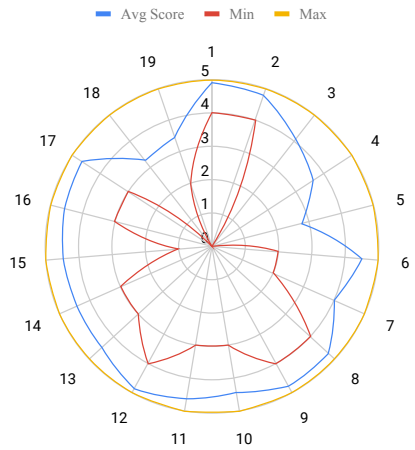
Table 2.2: Summary of results for the first assessment questionnaire (User Experience).

disagree) to 5 (strongly agree). A detailed description of the questionnaires, including a full list of the statements, is provided in Appendix. Questionnaires and notes (e.g., suggestions from the experts) collected during the assessment were anonymised.

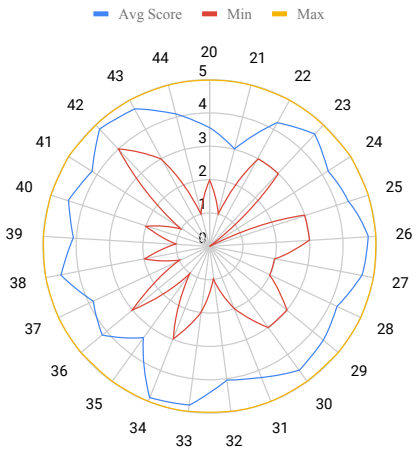
The average time recorded during the free interaction was about 16 minutes. The results of the interviews are summarized in Tables 2.2 and 2.3. Figure 2.21 gives a further summary of the results in a graphical fashion via radar charts. From the quantitative results of the assessment, it is possible to highlight that the experts liked the interaction with the installation, they appreciated the methodology used, and they judged the methodological aims to be fulfilled by the installation. Concerning the user experience, 15 out of 19 statements in the dedicated questionnaire received an average score above 4, meaning that the experts largely agreed with them. In particular, the interaction with the touch screen was definitely perceived as simple (Statement 2, avg: 4.78) and easy to understand (Statement 1, avg: 4.91). These findings corroborate evidence from another user experience evaluation study [10], where users unanimously preferred the touch screen rather than manipulating 3D objects. This is due both to the innate excellence of the touch screen as means of direct interaction and to the widespread

Statement	Avg Score	Std Dev	Min	Max
20	3.57	1.16	2	5
21	3.00	1.13	1	5
22	4.22	0.74	3	5
23	4.61	0.72	3	5
24	4.22	0.85	0	5
25	4.39	0.78	3	5
26	4.78	0.52	3	5
27	4.65	0.71	2	5
28	4.27	0.85	2	5
29	4.43	0.84	3	5
30	4.57	0.59	3	5
31	4.22	0.80	2	5
32	4.04	1.15	1	5
33	4.78	0.67	2	5
34	4.87	0.46	3	5
35	3.39	1.23	1	5
36	4.17	0.83	3	5
37	3.87	1.06	1	5
38	4.52	0.79	2	5
39	4.09	1.04	1	5
40	4.43	0.85	2	5
41	4.17	1.30	1	5
42	4.83	0.39	4	5
43	4.70	0.56	3	5
44	4.09	1.08	1	5

Table 2.3: Summary of results for the second assessment questionnaire (Museum Architecture and History, Museum Collection and Manufacturing Opportunities).



(a) First questionnaire (User Experience).



(b) Second questionnaire (Museum Architecture and History, Museum Collection and Manufacturing Opportunities).

Figure 2.21: Radar chart summarizing the results of the two assessment questionnaires.

familiarity with touch screens stemming from the ubiquitous adoption of mobile devices. It is also important to point out that the navigation structure of the installation was easy to use (Statement 12, avg: 4.87), and, most of all, that the interactive experience as a whole was consistently regarded as pleasant (Statement 9, avg: 4.78). Only 4 statements, reported below for convenience, received a score below 4. All of them are connected with the interaction with the blow:

4. BlowFlute (blow sensors section) is a convenient means of interaction (avg: 3.61).

5. BlowFlute (blow sensors section) is simpler to use than the touch screen so as to appreciate the flute sound (avg: 2.78).

18. The variation over time of the flute sound was perceptible (avg: 3.26).

19. I easily perceived the nuances related to the sound of the flute (avg: 3.43).

It must be remarked that the scores were negatively influenced by compromises made during the design phase. In order to maintain the installation sober and visually integrated into the museum, no prominent cues were added about the possibility of blowing into the holes over the furniture; hygienic consideration also contributed to this decision. However, this choice made interaction less intuitive and some experts initially missed the blow sensors functionality altogether, with an understandable impact on their ratings. Some experts also complained that blowing in the holes required them to stoop because the height of the cabinet was insufficient. Again, the height was a compromise, given the fact that the installation must be usable both by kids and adults.

In the second questionnaire, 21 out of 25 statements were uniformly agreed upon by the experts, with an average score above 4. In particular, it is possible to state that the installation effectively communicates information about the flute (Statements 26, 27 and 30, all with average scores above 4.5), and it does so better than it can be done by conventional means (Statement 33, avg: 4.78). Indeed, the virtual experience with the installation was consistently perceived as superior, from an information standpoint, to direct manipulation of the artifact itself (Statement 34, avg: 4.87), if so had it been possible. The multimedia nature of the installation, and primarily the sound, was regarded

as a plus (Statement 38, avg: 4.52). Only 4 out of 25 statements received lower scores, as summarized below for convenience.

20. The installation integrates aesthetically in the context of the room where it is located (avg: 3.57).

21. The installation aesthetically enriches the room where it is located (avg: 3.00).

35. Manipulating a virtual model of the flute is better than manipulating a physical reconstruction of the flute (avg: 3.39).

37. Blowing into a hole (blow sensors) is preferable with respect to other blowing possibilities (e.g., blowing into a straw) (avg: 3.87).

The scores for statements 20 and 21 were heavily influenced by the fact that the installation is not currently placed in the exact museum spot it was designed for. The actual installation position contains furniture that is aesthetically different from the installation. As far as the remaining two statements are concerned, in both cases, it is possible to attribute the low (albeit not disastrous) scores to design choices that were necessary to take for practical and hygienic considerations. Indeed, it was possible to provide visitors a physical reconstruction of the flute because it would have been too fragile for kids and, in any case, too difficult to clean when handled by hundreds of visitors. For the same reason, any artifact for blowing that could be receptacle for dirt was ruled out during the design phase. The opinions of the experts show that such design decisions, albeit sensible, had a measurable impact on the appeal of some aspects of the installation.

Chapter 3

Position Tracking and Painting Installation

In the case of impulsive sounds, a good recognition system should take care of the highly non-stationary properties of the signals, and the developed methods should be designed considering temporal dynamics with great attention. Some examples of impulsive sounds are door slams, explosions, footsteps, gunshots. In this project a special focus was devoted to footsteps recognition and tracking over a wooden runway. In literature several researches focused on tracking people indoor's movements using expensive sensors [6, 52] as geophones [93, 163] or video-cameras [110, 159]. Starting from the basic features of a single footstep [88], one of the most common techniques of indoor localization is the multilateration through Time Difference Of Arrival (TDOA) approach [115]. This technique gives good results starting from the basic principle of constant speed of propagation along the medium. Indeed, the position of the user is detected through the recording of the difference of arrival time of the sound source to at least three sensors receiving the signal. According to the definition of speed, the distance from a source is proportional to the product of speed of propagation and the time to reach the sensor. After the triangulation of the data is possible to detect the position of the user. However, what happens if the medium in which the wave travels is not homogeneous? This is the case of a wooden runway, where a slightly different approach is needed. The project has

been developed with the collaboration of Microtec[1] a company specialized in diagnosis of wooden board and interested in the study of the time propagation of sound waves in wood. In this project, the algorithm developed could be executed on embedded systems and this allows users to interact with natural user interfaces in a multimodal interaction architecture. Indeed, the implementation of the algorithm has been applied to multimedia installations for the valorization of contemporary art. The position of the user represents one of the layers of colors of a painting over the runway. By walking on it, the user explores the different layers that are explored thorough images projection over the painting, lights inside the runway and soundscapes.

3.1 Multimodal interaction and art

In the late '60s and '70s, installations became one of the favorite form for artists to work against the notion of the permanent, and therefore collectable, art object. Interactive installations are a sub-category of art installations and, for instance, the technology paradigm involved can be: i) mobile-based [39, 122], where the screen and the sensors of a phone are exploited to increase engagement during the visit; ii) web-based [108, 154], that allow to re-experience the contest and contents of the museum even after the visit; iii) electronic-based, where the principles of gestural and tangible interfaces are applied [23]. Several projects have been carried out in order to technology augment arts in the museums with interactive technologies. For instance, one example of these projects is [184] that introduced the notion of 3D sound in headphones for an art museum, providing the user with a contextual and spatial audio guide. Furthermore, O. Bimber et al. [16] introduced the idea of using computer graphics and augmented reality techniques in order to provide projected overlays on backgrounds with arbitrary color and reflectance. Kortbek et al., in [106], proposed three spatial multimedia techniques for communication of art in the physical museum space avoiding disturbance to other visitors. Furthermore, Myron Krueger is considered to be one of the first-generation virtual reality and augmented reality researchers. He conceived art of interactivity as opposed to interactive art: there was more interest in interactivity design than in the art itself. One of his

[1] https://microtec.eu/it/

Figure 3.1: The painting of Hartwig Thaler subject of the installation developed.

most representative works is *VIDEOPLACE* [109], one of the first systems that combines a participant's live video image with a computer graphic world. Finally, [72, 73] described innovative projects where interactive floors are used as interfaces to interact with the user's body. In spite of the technology used, interactive art frequently involves the audience acting on the work of art, increasing their involvement and engagement [51, 85].

3.1.1 The Painting Sonification

The goal of the installation of this project was to provide the user with the immersive experience of a soundwalk simulating an imaginary navigation inside the painting. The painting belongs to the South Tyroler artist Hartwig Thaler[2]. The environment was formed by a rich sound background. Such outcome were realized by the painting sonification and the rendering of the users' footsteps over a bed of dry leaves, such as the one represented in the painting used (shown in Fig. 3.1). Both the soundscape and the steps sound changed according to the user's distance from the painting, which resulted in a three-dimensional exploration of the original two-dimensional artifact. The painting was chosen among a series of nine, where various leaves textures were depicted. Although the textures vary for density, leaf dimension and color, the paintings were united by

[2]Further information about the painter Hartwig Thaler can be found at his website `https://www.hartwigthaler.de/`

Figure 3.2: Example of image processing from the original painting to a single color mask.

the presence of a single subject (a carpet of dry leaves) occupying uniformly their entire surface. Moreover, the paintings were the result of the superimposition of different layers representing leaves of similar color, as confirmed by the creative process declared by the artist himself. This suggested the possibility of decomposing the image in different color matrices and to consider the painting a three-dimensional object of different superimposed layers (i.e. see Fig. 3.1.1).

This idea envisioned that various layers should not simply have been seen as one above the other on the surface of the painting but that they could have been disposed in the space in front of it. Thus, the painting is stretched from the wall into the space, allowing a user to walk through its various layers. The resulting abstract model of interaction space is depicted in Fig. 3.3 where the user walks on the wooden runway positioned in front of the painting. The runway is divided into three sections, each corresponding to the imaginary projection of the color matrices which can act as triggering points of some changes in the audio output. Hence, the installation does not only qualify as a mere image sonification project, but it becomes a truly interactive environment based on the spatial projection of the painting. The runway represents the natural user interface through which is possible to interacts with the installation, without the need to understand how it works since it represents a common structure that is easily interpreted by the user.

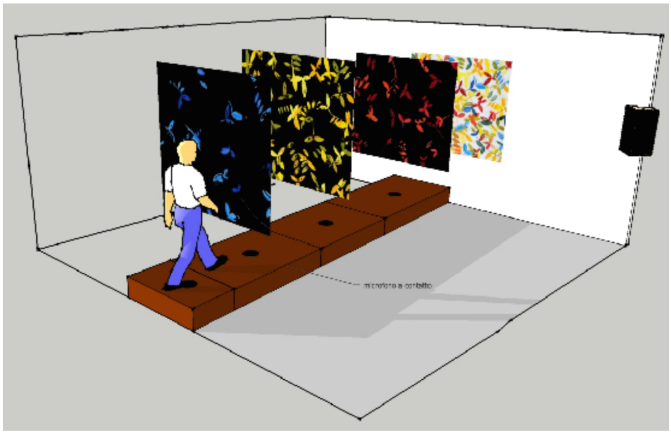


Figure 3.3: The painting sonification setup with the imaginary projection of three color matrices in the space in front of the painting.

3.2 System architecture

The installation was characterized by two main elements that were the input and output sources: the firsts were represented by i) the sensors on the wooden runway, that led the visitor toward the projection of the painting positioned at one and of it and ii) static image features extracted from the painting; the outputs were i) an audio reproduction system, fed from a sound background derived by the painting and users' step sonification and ii) a projection of images that could highlights some details of the painting and led strip integrated in the runway representing a 3D extension of the painting. Fig.3.4 shows the scheme of the installation, outlining the interaction components.

The soundscape of the installation changed according to the distance of the visitors from the image of the painting. In order to detect the position on the wooden runway, a network of piezoelectric sensors was used (Sensing). The signals acquired were then processed by means of a localization algorithm developed to work with a non homogeneous medium as wood. For the basis of this algorithm, the multilateration through TDOA approach (Localization) have been used. Other localization approaches were considered. For instance, computer vision algorithm using a kinect sensors would have worked perfectly in this context. However, this approach was not considered since the main goal was to study an approach based on waves propagating in the medium. Furthermore, computer vision algorithms would not have been able to detect the steps of the user.

The information acquired is used to control the soundscape and the color projec-

tions. The sound is produced by a bank of oscillators, fed by pink noise. Its outputs (Soundscape reproduction) have been differently weighted according to an algorithm of image processing that extracts the needed features (Image feature extraction) from the three color layers of the painting (color matrix extraction). Furthermore, the runway is provided with a led "stream" able to change its color according to the color matrix of the painting (Color 3D Projection). This provided a visual feedback so as to outline different portions of the runway, each of which corresponded to a different color layer. Finally, the color matrix extracted from the painting have been then projected over the painting to highlight the respective element of the painting. Further details on the elements illustrated in Fig.3.4 will be described in the following sections.

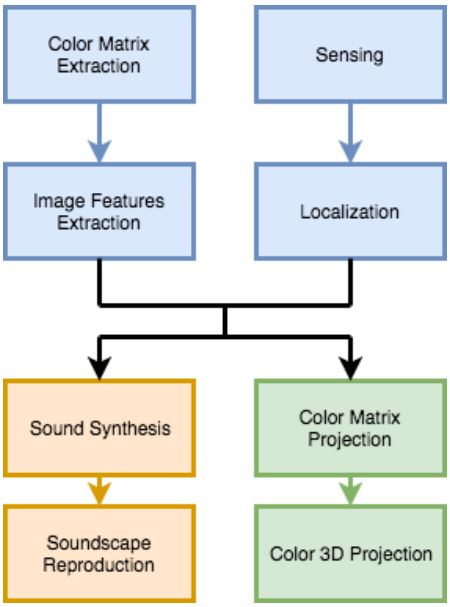


Figure 3.4: Schema of the system architecture: painting analysis, position detection and sonification.

3.3 From the Painting to the Sound

The image sonification project have considered the leaves as the elements of a mask, in which the colored spots represented open holes on a solid background. Transposing this interpretation in musical terms, the background was a complex wide range spectrum of frequencies and the holes are the points where the single frequencies could be heard. In this perspective, the function of the image was similar to that of an orchestral score,

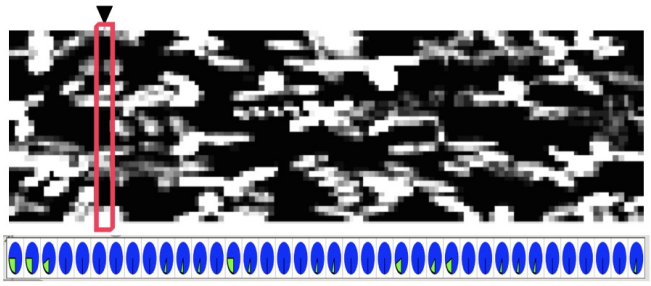


Figure 3.5: The 38x127 lightness matrix with the mixer controls of the separated filters output

where each line corresponds to a musical instrument or orchestral section. The black holes (single note or group of notes) activated the section, while the white lines (rests) muted them. Hence, the spectrum played the role of the whole orchestra and the image the role of the musical score. The image has been analyzed using the hue-saturation-value color model, where hue referred to the pure color, saturation to the quantity of white and value to the lightness [55]: this allowed to extract the different layers of colors. Choosing an appropriate hue range, it was possible to obtain matrices representing only the leaves of the selected color range. The three main colors extracted were red blue and yellow. The result of this process is depicted in Fig. 3.1.1, where a mask of yellow leaves extracted from the original image is shown.[3] A bank of 38 band pass resonant filters[4] fed with pink noise and with separate controlled outputs was employed to provide a rich background texture for the image sonification. Hereafter, in order to superimpose the mask to the background texture, the yellow matrix has been scaled to a 38 rows and 127 column matrix, where only the lightness values were reported (see Fig. 3.5). This matrix represented the real sonification score because it provided columns of lightness values scanned at a regular speed from left to right and vice versa. This timing mechanism allowed the necessary link between the visual representation and the sequential nature of music. The time-controlled columns of lightness values have been converted into numeric controls for a mixer matrix, thus making it possible to weight the filter's output according

[3] Of course many other options are available to extract color matrices from an image, starting with the number of matrices to be extracted and with their color range. In this case the choice of extracting three matrices depends not only on the artist's suggestion but also on the actual available space for the interactive sonification on the sensor-equipped runaway

[4] The sonification is implemented in a Max/MSP patch employing the fffb object. See `https://docs.cycling74.com/max5/refpages/msp-ref/fffb~.html` for reference.

to the lightness of every single pixel. This process resulted in different harmonies coming from the various weights and assigned to the spectral components. In the case represented in Fig. 3.5, the marked column has higher values in the lower part of the painting and very small values in the higher part. This opens the lowest frequencies of the spectrum much more than the highest, allowing a higher weight of the lowest components. The design was similar to a 38 voices chorale[5] where the dynamics of the various voices depended on the pixel lightness values.

3.3.1 The Installation Soundscape

The characteristic of the sonification engine described above had to be very close to the image features and to be easily adaptable to produce a painting-related soundscape. For this project, subtractive synthesis has been used [45] due to its flexibility and richness of expressive possibilities. The bank of filters could be fed with different input sounds and easily tuned according to various spectral models, which could have been changed in relationship to the color of the matrix employed as musical score. Moreover, the resonant filters output allowed a smearing effect smoothing the harmonic transitions which characterized our image sonification approach, giving a great variety to the audio rendering in a seamless way. A second bank of filters with the same characteristics of the first is fed by a file reproducing the sound of a step on a carpet of leaves, which is triggered by the user's real step on the wooden runway. Thus, not only this produced step sounds completely different from real ones, but they could be tuned according to a spectrum complementary to that used for the background. In the project, four spectra have been experimented: H (a harmonic spectrum of 38 frequencies starting from 100Hz), B (38 frequencies extracted from the full spectrum of a bell sound), hB (38 frequencies extracted from the highest band of the same bell spectrum) and lB (the same as above in the lowest band). Table 3.1 shows the full range of the possibilities in order to combine a background spectrum with its complementary with a wide number of soundscape generation possibilities, which can be chosen according to the user's position

[5]The chorale is a vocal polyphonic composition, usually a religious hymn, typical of the sacred literature of the German Protestant Church under Martin Luther (1523). The chorale is characterized by a regular homo-rhythmic proceeding of the voices, exactly in the same way of the scanned columns of the image matrix.

		STEPS			
	SPECTRA	H	B	hB	lB
BACKGD.	H	■			
	B		■		
	hB			■	
	lB				■

Table 3.1: Combinations of background and step sonification employing 4 different spectra: H (harmonic components), B (bell components), hB (highest range of bell components) lB (lowest range of bell components)

on the sensor-equipped runway.

3.4 The sensor-equipped runway

The installation is composed of a wooden runway, with dimensions of 350x100x7 centimeters, which guides the visitors towards the painting projection. The wooden board used for the runway has been individually chosen, in order to guarantee a speed of propagation of vibrations as uniform as possible (see Tab.3.2). This feature was necessary in order to reduce differences in arrival time to the sensors. The boards used for the runway have been manually chosen after measuring the intrinsic speed of propagation of sound waves along the boards before the assembly. These measurements have been done using the tool Viscan [6] produced by the company Microtec, involved in the project. Viscan Strength Grader is one of the best laser interferometer scanner for determining the Modulus Of Elasticity (MOE) of lumber. Among other things, it can measure the board's resonance frequency thanks to a high-performance laser vibrometer that works independently from environmental interferences such as noise. In table 3.2 are reported the measure of length, width of the board and the speed of propagation of the waves selected for the project.

The goal of sensing section is to acquire the footstep-induced structural vibration through sensors mounted on the runway. The sensing module consists of data acquisition, using a network of piezoelectric sensors, signal amplification and digitization by means of a digital audio interface. The localization section is mainly divided in: Short

[6] Description of this tool can be found at `http://microtec.eu/en/catalogue/products/viscan/`

# board	length (cm)	width (cm)	speed (m/s)
41	495	14,5	544,5
4	550	14,5	544,5
999	490	31	529,2
41	494	30	513,8
111	550	7,5	511,5
118	523	7,5	493,7

Table 3.2: Wooden boards description: the identifier number of the board, the length, the width and the speed of propagation of sound.

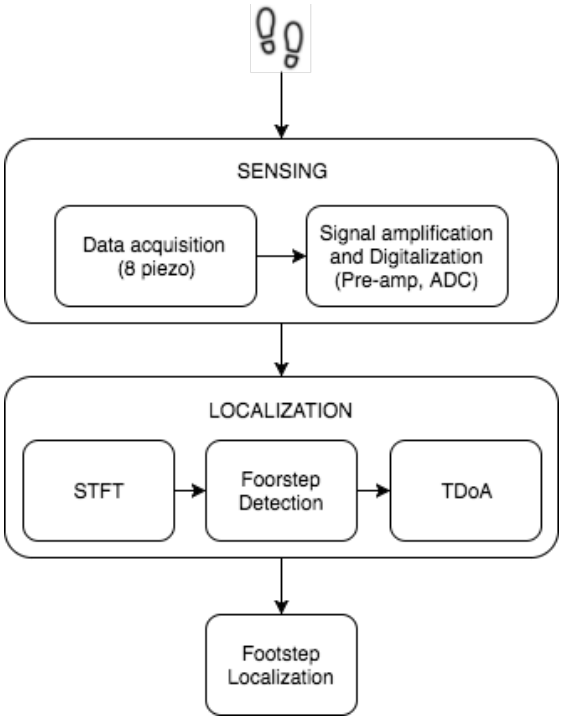


Figure 3.6: Schema of the position detection system: Sensing is responsible of the acquiring of the signals from the runway, Localization detects the steps and the position.

Time Fourier Transform (STFT), footstep detection and Multilateration through TDOA estimation (see Fig. 3.6).

3.4.1 Footsteps characteristic

Human steps generate vibrations that propagate away from the source as seismic waves. For a single impact on an elastic half space, Miller and Purssey (1955) have shown that 70% of the energy of the impact is distributed in the Rayleigh wave. The remaining 30% of the energy is transmitted into the earth via body waves (transverse and longitudinal), while diminishing in amplitude as r^{-2} [74, 181]. The signal has also frequency dependent

attenuation characteristics [5]. Furthermore, thanks to clinical gait analysis, like in

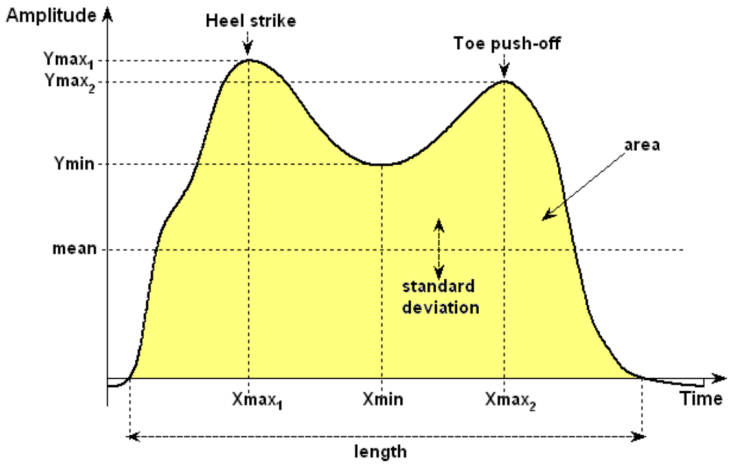


Figure 3.7: Representation of a usual Ground Reaction Force of footsteps on an Amplitude/Time graph.

[189, 198, 182, 90], it is possible to outline the Ground Reaction Floow (GRF) of a footstep. According to Newton's Law of Reaction, GRF is the force equal in magnitude but opposite in direction produced by the ground as reaction to the force the body exerts on the ground. The ground reaction force is used as propulsion to initiate and control the movement, and it is normally measured by force sensor plates. Fig.3.7 describes the main behavior on amplitude/time graph, where the first peak is attributable to the heel strike and the second to the toe push-off as the body is propelled forward. In the light of these analysis, a "delay" time after the first peak, before next detection, was necessary in order to identify the position of the whole step at the heel position and to avoid continuous changes in short periods of time. With this approach, the vibrations made by the step can be considered as produced by an impulsive hit.

3.4.2 Sensing

Usual applications of floor vibrations detection use sophisticated sensors like the geophone sm-24. It can detect very small vibrations with high precision, however, geophones can be costly (around 50$ per sensor). In order to develop a scalable and usable system, another type of sensor would be a more suitable solution. Due to the small size of the runway and the resonance of steps, a good Signal to Noise Ratio (SNR) can be ensured

with piezoelectric sensors of 2cm diameter. The sensors are equally distributed on the runway in order to divide it in homogeneous areas, each of which is identified by the four nearest sensors that the localization system will use. As previously described, the painting is divided in three color layers, each of which is assigned to a section of the runway. An example of distribution of sensors is described in Fig 3.8. Afterwards, in order to acquire the signal, the TASCAM US-2000 (8 Microphonic input) Digital converter was used. For this application, the audio interface works at 44.100kHz with a resolution of 24bit. The cable and connectors used to attach the sensors to the audio interface are respectively Klots and Neutrik so as to minimize possible filtering caused by unwanted parasitic capacitance.

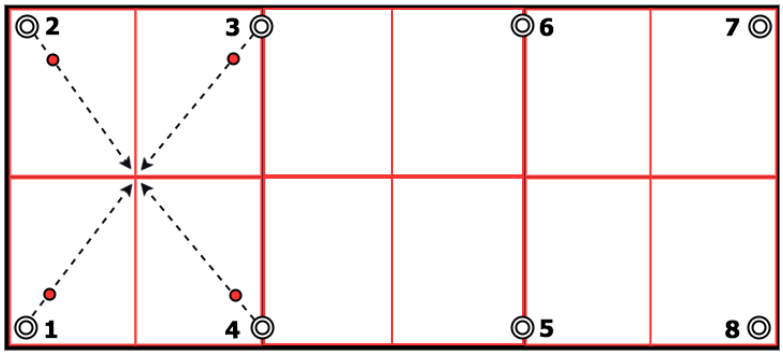


Figure 3.8: Preliminary disposition of the sensors.

3.4.3 Localization

In literature, the localization of sound events is often based on the amplitude of time signals and the cross-correlation. [165, 47]. In signal processing, cross-correlation is a measure of similarity of two series as a function of the displacement of one relative to the other (lag between them). After several tests, this approach turned out to be unable to extract the necessary information. In particular, the main issue was due to the nature of the runway: wood is a flexible and non-homogeneous material which leads to the appearance of echoes and creaks that modify both the sound signal and the speed. On the contrary, by means of STFT it is possible to clearly detect the onset of the footstep and so to highlight the TDOA. In order to detect the footsteps, a moving average method filter was used: with two windows of size N and M, with $N >> M$ and N multiple of M,

the average energy of the signal in the long period and in the short period are computed respectively. Every N samples the long period average is set, then every M samples a value of energy is computed and compared with the previous average energy (see Fig. 3.9). The algorithm was first simulated and evaluated using MATLAB with data sets of measures taken from the prototype runway equipped with four sensors. In formula:

$$\frac{1}{M}\sum_{m=0}^{M}|X(jN+kM+m)| > \alpha\frac{1}{N}\sum_{n=0}^{N}|X((j-1)N+n)|$$

where $X(f)$ is the Fourier transform of input signal $x(t)$, j the number of the N-samples windows, k the number of the M-sample windows and α a tuning factor. This double window approach is necessary in order to increase algorithm robustness to environmental noises and other sources of noises due to crackle and other wood movements. Furthermore, the α factor is a tuning parameter that allows to vary the ratio between the two windows according to the specific contest. Finally, in order to increse the SNR of the

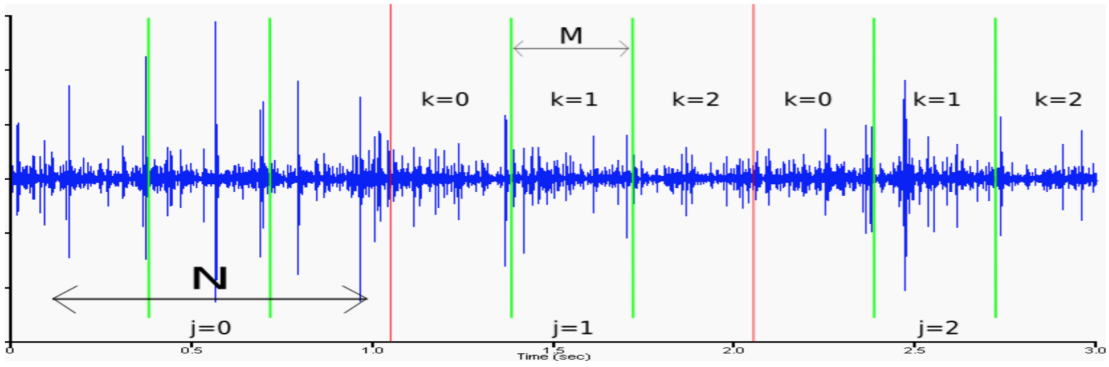


Figure 3.9: Graphical representation of the moving average algorithm implemented for the localization algorithm.

signal, a cubic factor is applied to signal values. This ensured a higher ratio between lower noise signal with respect to the footsteps.

Multilateration is one of the best approaches for source localization which uses the TDOA approach. Mathematically, the localization procedure can be described as

$$||x-p_2||_2 = v(t_2-t_1) + ||x-p_1|| \tag{3.1}$$

where p_1 is the location of the first sensor and p_2 is the location of the second sensor, x is the location of the excitation (i.e. footstep) and v is the propagation speed. Considering that non homogeneous material does not ensure constant speed of propagation of waves [190], this approach needed further elements that could take into account possible delays of the signal. This is the reason why the runway is divided in twelve regions, so as that the maximum localization precision is proportional to the dimensions of each area. The information about TDOA of each sensor is finally processed, using a decision algorithm, to identify the region in which the event occurred. This algorithm is described in the next sections in light of the results obtained from the evaluation test on the runway.

3.4.4 Real-time implementation

The localization system works in real-time thanks to the implementation of a software dedicated to the acquisition from Tascam buffer and the processing of the data received. The framework is implemented in C++ in order to maximize portability on different systems. A well-known and widely used library for managing Audio Interface buffers is used: Port-Audio. It provides a real-time working API streaming audio using a callback function or a blocking read/write interface. Finally, a visual interface using Qt[7] allowed to visualize the sensing and localization system decisions. Qt is used for development multi-platform applications and graphical user interfaces (GUIs). This interface, during test phase, allows to visualize the real-time position of the user on the runway and change algorithm parameters so as to test its correctness.

3.5 Evaluation

In order to test the algorithm of step detection and to evaluate the results, a prototype of the runway has been used in a preliminary analysis. A wooden pallet 120x80x15cm with a 2 cm thick wooden board with base measuring the same as height was used. The results obtained by the step detection algorithm on the prototype, shown in Fig.3.10, are: after 10 campaigns of data collection, each of which composed of 16 footsteps, the accuracy measured was 91%. After this preliminary part, the following test was performed on

[7] https://www.qt.io/

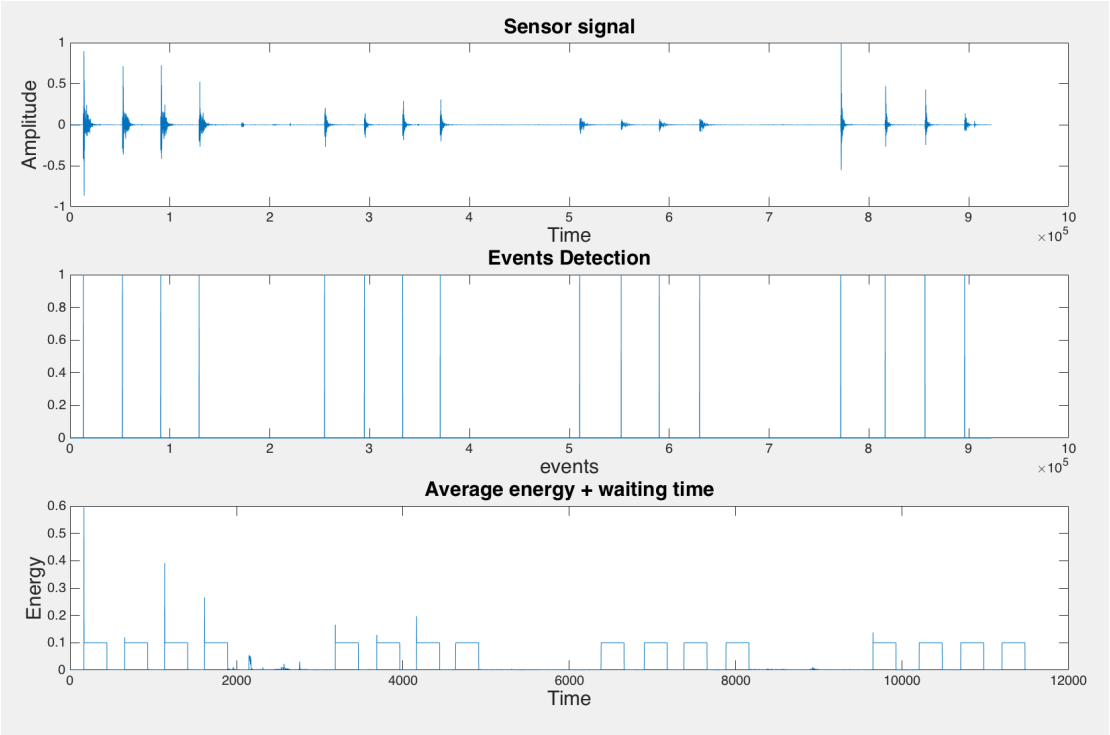


Figure 3.10: Event detection: the first image shows the signal of sixteen steps, the second the detection of the steps and the third the average energy variation (the straight line describes the waiting time after an event detection).

the wooden runway. First of all, to analyze its response, the sensors have been placed in the first four positions of Fig.3.8 as for the test on the prototype. Afterwards, an experimental set-up was designed to make the analysis easily repeatable. In order to realize an homogeneous excitation from which extract the TDOA data, the experimental set-up was mainly composed of a 1 meter height reference and a mass of 50g. Along the diagonal, at 20 cm from each sensor toward the center of the relative area (red dots in Fig.3.8), 10 campaigns of series of 10 impulse excitations have been carried out using the mass dropped from the height reference. These two elements of the experimental set-up are respectively an adjustable stand for speakers, with a transverse metal bar, and a tennis ball. The choice of using a bouncing mass was prompted by the necessity to generate an impulsive excitation and therefore to avoid further bounces. This was possible by catching the mass before the second bounce (see 3.11). The data collected from the sensors were then processed and the TDOA (measured in number of windows of size M) evaluated. Focusing on the impulses generated in the areas of sensor 1 and 2:

Figure 3.11: Experimental set-up: an adjustable stand for speakers, with a transverse metal bar, and a tennis ball.

- The first sensor able to detect the hit is the nearest;
- The second sensor is the nearest to the first (i.e. sensor 2 if impulse is generated near sensor 1 and vice versa);
- The third and the fourth are respectively the one on the longitudinal direction and the one on the diagonal one.

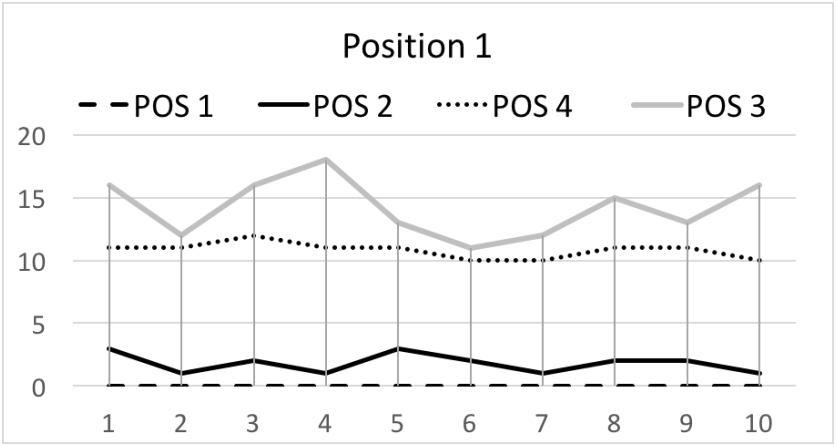
An example of data is represented in Fig.3.12(a), where a graph of the trend of TDOA between different hits in the same campaign is shown. Therefore, these results are in contrast with those collected from sensors 3 and 4:

- The first sensor able to detect the hit is the nearest, as in the previous test;
- The second sensor is no longer the nearest to the first but the one on the longitudinal direction.

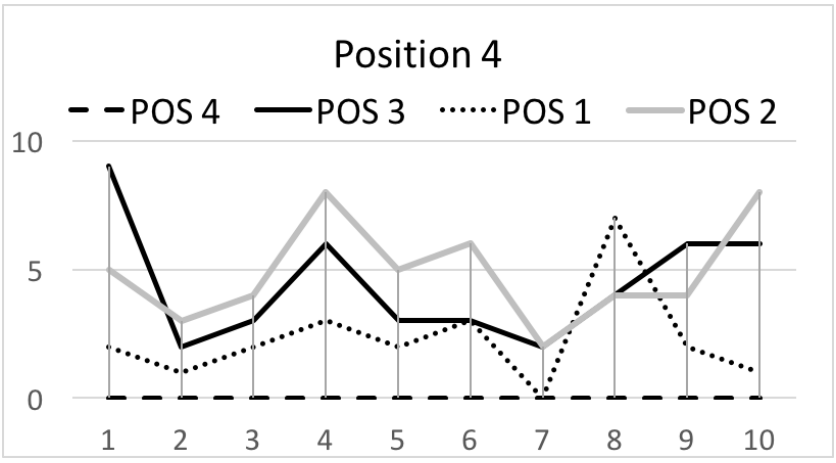
These results are shown in Fig.3.12(b) and they can be explained by the construction design of the runway: as explained in [190], the propagation of waves along the wood

grain is faster than the transversal one. In the case of position 1, the two sensors are near the edge of the runway, realized with a single piece of wood, the grains of which are horizontal with respect to long edge of the runway. Therefore, in this board the waves travel faster than the usual transversal propagation and reach the sensor in position 2 before the one in position 3. On the contrary, sensors 3 and 4 are separated by wooden boards whose grains are transversal w.r.t. the abstracted joining line of the two sensors and this contrasts the fast direct propagation of the waves.

The mean and the standard error of the data collected for all the four sensors of the example campaign just described are summarized in Tab.3.3.



(a)



(b)

Figure 3.12: Graphs of the TDOA trend (measured in windows of size N as in Fig.8) between signal sensors for excitations respectively in position 1 and 4: POS# describe the area of the sensor, the POS order in the legend is the expected order of arrival of waves at the sensors (geometric considerations); y axis describe the TDOA in number of windows and finally x axis describe the number of hit of a single campaign.

Mean±S.E.	Sensor1	Sensor2	Sensor3	Sensor4
POS1	0	1.8±0.2	10.8±0.1	14.2±0.4
POS2	2.2±0.3	0	7.1±0.1	18.5±0.3
POS3	6.5±0.1	4.2±0.3	0	6.8±0.1
POS4	2.3±0.1	4.9±0.1	4.4±1.0	0

Table 3.3: Mean and standard error of the TDOA of a campaign of measure.

Due to the results of this preliminary analysis, a slightly different approach was necessary in order to ensure that the knots of the wood and the construction features of the runway would affect the results of the localization. The main observation is that the positioning of the sensors is crucial and it must be carefully considered. The most reliable information is the first sensor reached by the step vibrations, so it is important to design the sensors network in order to guarantee: i) the proximity of the sensors to the user path and ii) the mutual position between the sensors. The proximity of the signal receivers to the steps guarantees that the first sensor reached by the vibrations is the closest representation of the user position. The more sensors used in the network, the more precise the localization accuracy. The use of piezoelectric sensors allows to have dozens of sensors without effecting scalability of the model of localization since this type of sensors are very chap with respect to geophones and more sophisticated sensors. The mutual position is important to ensure that ambiguity between different sensors is minimal. In the case of the wooden runway, the sensors have been positioned one after the other following a straight line in the center of the runway, parallel to its longest side (see Fig.3.13).

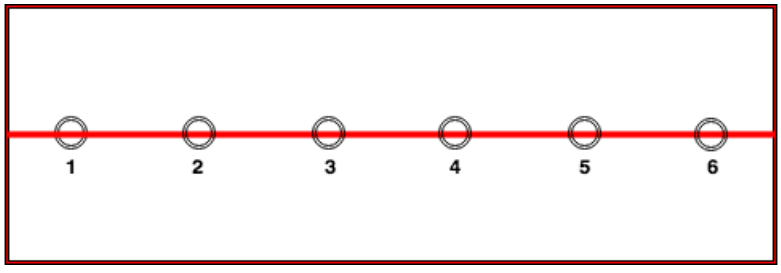


Figure 3.13: Displacement of the sensors network used for the installation.

Figure 3.14: The installation shown at the "European Researchers' Night 2018".

3.6 Assessment

The resulting installation has been assessed during the "European Researchers' Night 2018" (Fig.3.14) in Padova, during which the installation was free to be used by visitors. The installation has been assessed collecting two different data: i) the positions of users during the interaction with the runway and ii) the answers to a survey proposed to the users after interacting with the installation.

The positions collected are taken from 20 people visiting the event. They were free to interact with the installation and no suggestions were given. The positions collected are shown in the heat map of Fig.3.15 and highlight that users prefer to walk along the middle of the runway. The users tended to explored the runway walking up and down listening and looking at the changes of the soundscape and the light projections. However, after some steps they understood that even the steps were involved in the interaction and started to focus more on the sonification of them staying in place. Indeed, in Fig.3.15 there is a specific point were purple color highlights the higher frequency of steps. That position was the one used most of the time by the users. This trend of positions in the middle of the runway is highlighted in the histogram of Fig.3.16.

Ten of the 20 people (over 18 years old) answered to the questions of a survey prepared to assess their experience in using the installation. The average age was 38,7 with a

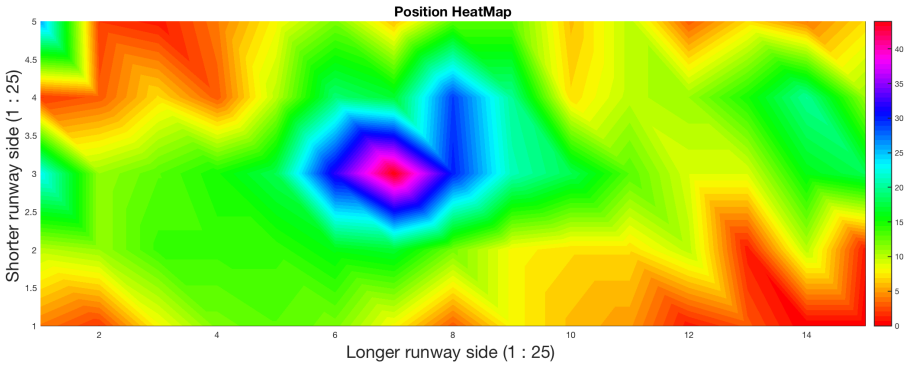


Figure 3.15: Heat map of users positions over the runway. The painting is positioned at the left side of this image.

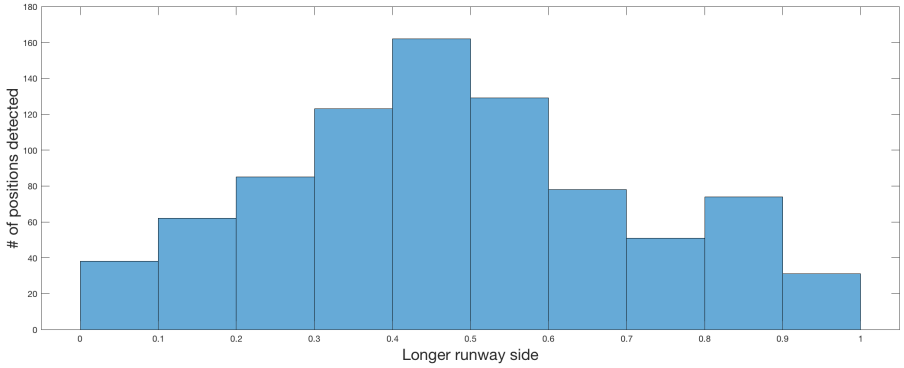


Figure 3.16: Histogram of users positions over the runway. The painting is positioned at the left side of this image.

Statements	1	2	3	4	5	6	7	8	9	10	11
User 1	4	5	5	4	4	4	4	5	5	5	5
User 2	3	4	3	2	4	4	3	4	3	4	4
User 3	4	5	5	4	5	5	5	3	4	4	3
User 4	4	5	4	4	4	5	5	4	4	3	5
User 5	4	4	5	5	4	5	3	5	4	5	5
User 6	4	5	4	4	4	4	4	3	3	4	5
User 7	5	5	5	5	5	5	5	4	5	5	5
User 8	5	4	5	5	5	5	5	4	4	5	5
User 9	5	4	3	4	4	4	4	4	4	3	3
User 10	5	4	4	4	3	4	4	5	4	3	3
Mean	4.3	4.5	4.3	4.1	4.2	4.5	4.2	4.1	4	4.1	4.3
Dev.St.	0.7	0.5	0.8	0.9	0.6	0.5	0.8	0.7	0.7	0.9	0.9

Table 3.4: Survey results during "European Researchers' Night 2018".

deviation standard of 17,19 years old. Three of them were female, 7 were male and 8 out of 10 had little or no experience with multimedia installations, while 2 out of 10 had an average experience. The users could autonomously answer the survey, written on paper and composed of 11 statements, about the installation and their experience in interacting with it, using a Likert scale between 1 and 5 (from strongly disagree to fully agree). The survey was presented in Italian. Detailed statements are shown in the Appendix. Some examples are: "Understanding how to interact with the installation was easy", "The interactive experience was pleasant" and "The sound of footstep led me to walk more". The results of the survey are shown in Tab.3.4. Users stated that interact with the runway was easy and led them to walk more. Furthermore, with an average score greater than 4 (statement 4), most of the users agreed that the installation valorized the painting allowing them to discover peculiarities of the painting they would have overlooked otherwise.

Chapter 4

Multi-pitch Detection and Piano Teaching Game

As described in the introduction of this thesis, studies on HCI have always paid close attention to the design and evaluation of interfaces that simplify and make the interaction more intuitive with digital contents. In the majority of the applications considered in these studies, the purpose is to access digital contents in order to obtain useful information, modifying or sharing them with other users. In different contexts, for example video games, the user interface involves specific hardware control devices, in order to achieve a goal or to perform a task in a more or less realistic virtual scenario. In these cases, the user interface is the means that should ease the access or carry out the task. However, what if the interface was the aim and not the means? What if the task to accomplish in the virtual scenario was just the means to learn how to use an interface or a controller? Under this assumption, the objective would not be to design an intuitive and easy-to-use interface, but rather to design contents and interaction modalities that stimulate the employment of not-intuitive and difficult-to-use interfaces. In this project, the goal is to present and discuss a case-study wherein the control interface of a video game is an acoustic piano. In this case, the video game, called *Musa*, is the means through which the user is induced to learn how to use a complex and not-intuitive control interface like the keyboard of a piano (compared to the usual controller for video-games like the joystick). The interaction paradigm is based on the idea, common to the Tangible In-

terfaces, of employing everyday object, or not designed for video games, as control user interfaces. The innovative part consists of using a particularly complex interface as a piano keyboard to control a real video game (e.g. a first-person shooter), that does not have explicit musical purposes. *Musa* can be considered a *serious game*, it means that is designed for a primary purpose other than pure entertainment. The "serious" adjective refer to video games thought for different purpose as education, scientific exploration, health care, etc [128]. Only one example of this kind exists (*WildChords*), though it is a commercial product whose interaction-associated aspects have not been analysed and evaluated. Over the past few decades, video game controllers have undergone continuous changes, starting from the classical input device such as joystick, mouse and keyboard, up to more contemporary and unusual trends. As described by Caroux et al. [30], new devices have been developed and used in commercial games. These input devices are based on motion (exergames [142]), touch, tangibles, gaze, or brain control. As far as motion is concerned, the review highlighted that this type of interaction allows to have a level of motion interactivity that provides for positively perceived realism, ease of use, spatial presence and enjoyment. Touch interfaces are perceived as easy to use, engaging and more flexible and precise than the motion-based interfaces. However, nowadays there is more interest towards games on mobile devices. Some authors have also designed and tested types of game control that have not been used in mainstream commercial games yet, such as gaze control, brain control and tangible interfaces. These input devices are designed and implemented with the aim of simplifying the usage of interfaces to play videogames. The result of these reviews is that the most important aspects are the ease of use and then the engagement. The wordplay of the title of this specific project is Playing to play. It is based on the double meaning of the verb "to play", and it can be seen with two different meanings: as playing to play the piano, which highlights the purpose of the game to teach how to play the piano, or as playing the piano to play, which highlights the piano as an innovative control interface to play with a video game, where the desire to win becomes an incentive to learn. Nowadays, a growing trend to bring more physical movement and social interaction into games is spreading in the gaming industry [120], while it strives to keep the benefits of computing and graphical systems.

Well-known examples about this assumption are the *Xbox Kinect* and the *Nintendo Wii*. These consoles invite the player to use its body to solve the game quests and to detect his movements through some sensors inserted in the controllers. Therefore, though they give a rough idea of the actual action (for instance, throwing a bowling ball), they still represent a facilitation because of the lack of some crucial real aspects (like the weight of the ball). Beside this tendency, it has been witnessed a push towards new ways of interfacing the user with musical games. Some successful examples from the past decade are *Guitar Hero*, *Singstar* and *Rockband*, which proposed controllers similar to the real but very simplified in their working principles. What these approaches have in common is that they all have a recreational goal, they do not pay any attention to learning. For example, the *Guitar Hero* controller does not produce any sound and does not give the mechanic of the real instrument. Therefore, it is difficult to highlight the logical succession between playing the game and playing the guitar. Other projects are trying to overcome the interface problem, by creating a replica of the instrument on a touch screen. Some examples are: *Magic Piano* by Smule, *Piano Notes* by Visions Encoded, *Pianist HD* by Rubycell, *Piano Magic Tiles 2018* by Piano Music House and project on the reconstruction of a virtual Pan flute [8, 7]. These last examples highlight the problem of preservation of sound art because of their short life expectancy, which is significantly shorter than other cultural material [21]. Along this tendency, some projects realise skeuomorphic interfaces creating a replica of tape recorders to access tape music documents using the original devices [28, 57]. On the contrary, recently there have been different approaches that tried to create more innovative controllers, for example by transforming any kind of surface (*Mogees*) or the human body itself (*Music Glove*) into a musical instrument. It is worth mentioning them since they represent an alternative solution to the problem. As they define brand new human-computer interaction modalities, they overcome the logical gap between interface and real object by making the latter exactly the same as the former. What these examples have in common is that they exploit audio as intermediary in the communication process between human and machine. Typically, music and sound, has been used by interface designers to convey information to users. The theoretical principles behind this choice were analysed

by William W. Gaver [65] and they can be summarised under the concept of auditory icons. The opposite process, which is controlling computers via musical messages, has been instead somewhat neglected. The main reason is probably the complicated intrinsic structure of music, which is for the most part associated to emotions and therefore it is difficult for a machine to handle it. Some valuable reason to continue accurate research in this field are described by James L. Alty [4]: music is pervasive in daily life, it is memorable and durable, it contains a large quantity of highly structured information and it involves the simultaneous transmission of a set of complex ideas related over time within an established semantic framework. However, up to this point, this kind of interactions have been basically limited to MIDI devices and audio interfaces [193]. Summarising, the former could be similar in the appearance of existing acoustic instruments (for instance, keyboards) or it could embody a new interaction concept. In both cases, their main characteristic is that they do not emit any sound unless they are connected to a digital platform which synthesises it. They are both used in music and video games industries. On the contrary, the latter are instead hardware tools that can bring the audio elaborations outside the CPU and allow recording in real time elaboration of acoustic instrument or electric ones. Some considerations about these different approaches can be extracted from the work by Gather W. Young et al. [201]. The authors focus on Digital Musical Instruments (DMI) devices and they recommend to evaluate the interaction in terms of functionality, usability and also user's experience. They bring out the example of playing music on a basic MIDI keyboard: the usability of this interface is poor with respect to playing on a grand piano, nonetheless the experience may still be believable to the performer. However, for Nijs Luc et al. [135] the importance of the musical instrument as a natural extension of the musician is highlighted. Along this trend, some other projects put the real instrument as the core interface to interact with the game: *Simply Piano* or *Piano Dust Buster* by JoyTunes, *Yousician*, *Flowkey* in cooperation with Yamaha or closed project as *Guitarbots* and *WildChords*. These are thought to teach to play instruments using simple graphics and a gamification approach. The interfaces could be compared to the one of *Guitar Hero*, as far as guitar games are concerned, since notes to be played are highlighted on a guitar neck, or related to the keyboard and music staff for

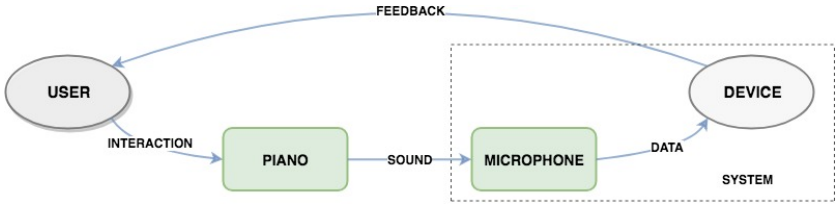


Figure 4.1: Interaction model scheme.

piano. These are not thought to be primarily a video game, since there are no settings or specific storytelling that can help students to get involved. Their approach is based on gamification [41], where the exercises to learn to play an instrument are integrated in basic principles, such as challenges and scores. As far as this study is concerned, *Musa* is first of all a videogame.

4.1 System architecture

The general principles of the interaction model presented in this case-study are depicted in Fig.4.1. The characteristics of the two communication channels are different. Messages from the device are sent through one or more direct media. This is meant to avoid an unpleasant latency which could impoverish the experience but also to give the user an instant feedback about the action he has just performed, in order to keep him focused on the actual target that is the keyboard. Responses are given exploiting both video games characteristics, such as storylines and settings, and a visual representation of the keyboard, which is easily comparable to the real one to help the user when it is necessary. On the contrary, since the goal is to make the user interact only with the keyboard, messages arrive to the system through an indirect channel. Therefore, the device must be able to detect the audio signals coming from the piano through its default microphone and to acquire from them the necessary information to give feedbacks to the user.

As far as the interface is concerned, a piano keyboard is made up of 88 black and white keys, each one associated to a specific pitch, increasing in frequency from left to right. White keys represent natural pitches (C, D, E, F, G, A, B) and are linearly arranged in a lower level with respect to the black keys, which correspond to altered notes (# or ♭). The related position between white and black keys follows a constant

pattern that repeats every octave (12 keys) and it represents the chromatic scale: the 7 white keys are always separated by a black key (5 in total), except for the B and C keys and E and F, that are not separated. The keys' shape and related distance are designed to allow a comfortable and quick sliding of the hands, while their weight permits the user to control the sound dynamics. In addition, learning gestural movements and pattern visualisation on a piano keyboard through combinations of notes could represent a basic technique appliable in other contexts, such as stenotype.

4.2 Gameplay

The gameplay is structured with increasing difficulty levels, thus children progressively improve their capabilities as musicians by playing. Besides, the game is thought with required movements and free exploration is not allowed. Primary targets are children between 6 to 9 years old. Children eventually are positioned at the beginning of the exercise. Thereafter, an avatar (see Fig.4.2) acts as the guide during all the gameplay, assisting you with a tutorial, which also allows kids to get familiar with the mechanics they will need to use. Then, the actual exercise begins and children need to realise on their own which keys must be played to progress. Elements throughout the game are

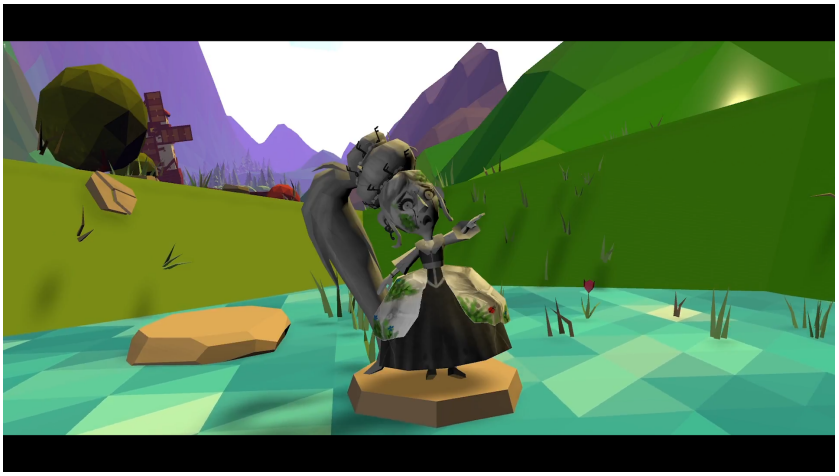


Figure 4.2: Avatar acting as guide in the game.

disposed according to associative rules, in order to help the children: for example, to perform a decreasing scale F-E-D-C, the avatar has to jump on flying stones organised one below the other such as the decreasing pitch of the notes (see Fig.4.3). After a predefined

number of seconds of inactivity or several mistakes, at the bottom of the screen an image of a keyboard will appear, giving hints on which is the correct key to play. This is not going to happen always though: once an ability is considered consolidated, hints will not be displayed anymore in order to help children to be independent, without waiting for suggestions. Each level has a limited number of attempts, which imposes to restart from the tutorial once exceeded. All these facilitations are not strictly related to the structure of the game, since it could be played with a different input device and the associations with the actions will not be lost. At the beginning of the game, it is assumed that children have never played a musical instrument before and do not know anything about music. Therefore, what is initially requested is just to play random notes, so they get acquainted with the keyboard as a whole. Then, the real learning phase begins, in which users are requested to play a single specific note (precisely note C in exercises number 1, 3, 5, 6, and F in exercises 4, 9). Kids have to look for the correct key, even by making mistakes: the purpose is that, once found, it will be linked to a specific "power" inside the game. This way, every time children encounter an obstacle which needs that particular power to be overcome, they will be able to autonomously make the association. This process is meant to be repeated for every single note, because it has been proven that for children learning through visuals and links it is a simpler and more effective method than the traditional mnemonic way [161, 169]. In more complicated exercises, notes are combined, making necessary to play them one after the other (ex. 2), and the concepts of musical scales (ex. 7, 8), repeated notes and notes variation (ex. 10) are then introduced. The performed test refers to these particular levels. Up to this moment, notes have been freely played, without any limitation: however, from the second level, it is necessary to introduce the notion of rhythm through some exercises in which children have to follow a certain tempo. Finally, at the end of this level, the first basic short piece is provided: through the melody of a famous song, children defeat their first enemy in what is a brief examination of the abilities they have learned.

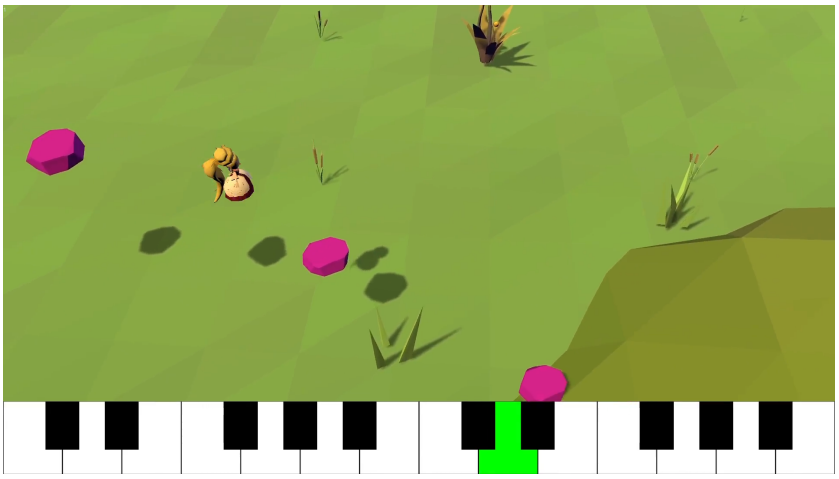


Figure 4.3: Exercise of the game where avatar jump between flying stones.

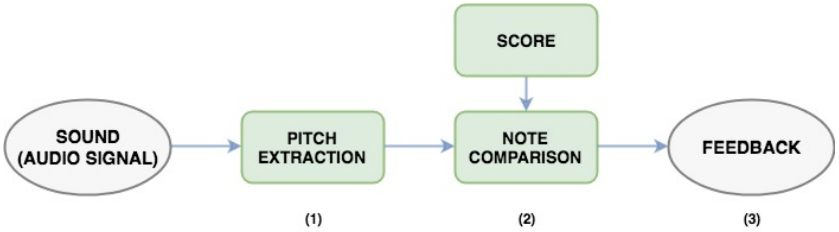


Figure 4.4: Score following schema solution for *Musa*.

4.3 Notes detection

In this particular case-study, the critical design step is the development of a note recognition engine. This problem falls into the *score following* research, defined by Orio et al. [143] as the synchronisation of a computer with a performer playing a known musical score. The authors also give a model solution which provides three phases: extraction of features of interest from the signal, acquisition of the response from the learning/matching model and finally execution of the related action. This structure is shown in Fig.4.4 in reference to this particular situation.

The requirements to be met in the project were high effectiveness, cross-platform portability and low delay. These prerogatives have brought to establish three gradual algorithmic development steps, each one corresponding to a different sub-problem: voice activity detection, single pitch detection and multi-pitch detection. For the first two, signal processing approaches give good results at the current state-of-the-art, while for the last one modern trends tend to prefer machine learning approach (see e.g. [175]).

However, this is still an open problem because the percentage of errors is high even with best solutions and this is the reason why the system was strengthened with a more robust algorithm for the single pitch detection. Nevertheless, performances can be improved in both cases if some *a priori* knowledge is available. In this particular context, this information is represented by the expected notes present in the game quests. Considering also the difficulty to collect a valuable dataset to train a machine learning model, it has been decided to adopt a signal processing approach for the multi-pitch phase too.

All the algorithms were implemented in C++ language inside the Unity framework[1], which allows the native code execution through a dynamic library compiled for the destination platform, thus solving the portability issue. Besides, efficiency has been granted through the use of *Eigen*[2], a template library in C++ for linear algebra, and low latency audio APIs of each operating system.

4.3.1 Voice Activity detection

The algorithm is based on the works presented in [53, 66, 177]. Its function is to detect the presence of useful signal, distinguishing it from the background noise, so that one of the pitch detection algorithm can be activated. Without going into details and referring to Fig.4.6, the system is active in a frame if both the mean spectral energy is over a minimum threshold (defined as the estimation of initial noise mean) and the estimated probability of useful signal presence given the observed signal is high. Further customization of the algorithm, property of the company involved in the project, is devoted to the recognition of harmonic components to discern between the sound of the piano, with respect to other environmental noises, as shown in Fig.4.5, where the red color highlight the piano sound and the green component correctly highlight pitch and further sounds.

4.3.2 Single Pitch Detection

The single pitch detection algorithm is based on [40], where the played note is recognised after the estimation of the signal period, which is the inverse of the fundamental frequency associated to the pitch. As shown in Fig.4.7, this processing relies on the cumulative

[1] https://unity3d.com/
[2] https://eigen.tuxfamily.org/dox/

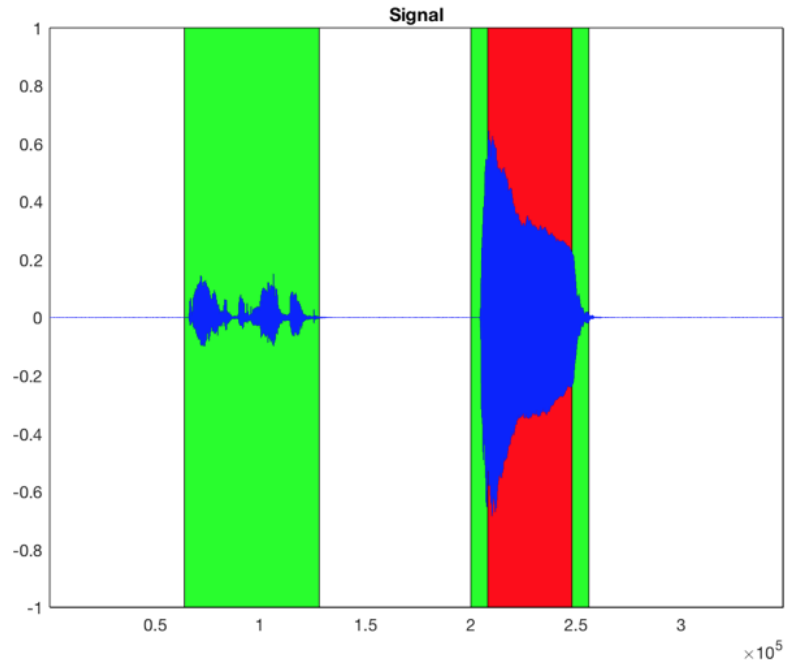


Figure 4.5: Piano sound detection. Red color highlights piano sounds. Green color highlights sounds as human voice or noises.

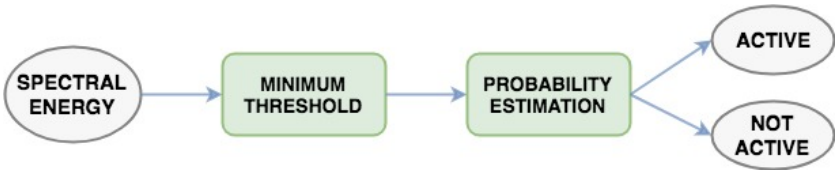


Figure 4.6: Voice activity detection algorithm scheme.

mean normalised difference function, that is a way to correlate a signal frame with its varying time-shifted version. The algorithm aims at approximating the lag value for which this function is null, as it is the signal period.

Figure 4.7: Single pitch detection algorithm scheme.

4.3.3 Multi-pitch detection

Due to the complex harmonic structure of the musical signal spectrum, multi-pitch detection is the most challenging problem. The chosen solution constitutes a trade-off between *a priori* detection and accuracy requirements. Indeed, as depicted in Fig.4.8,

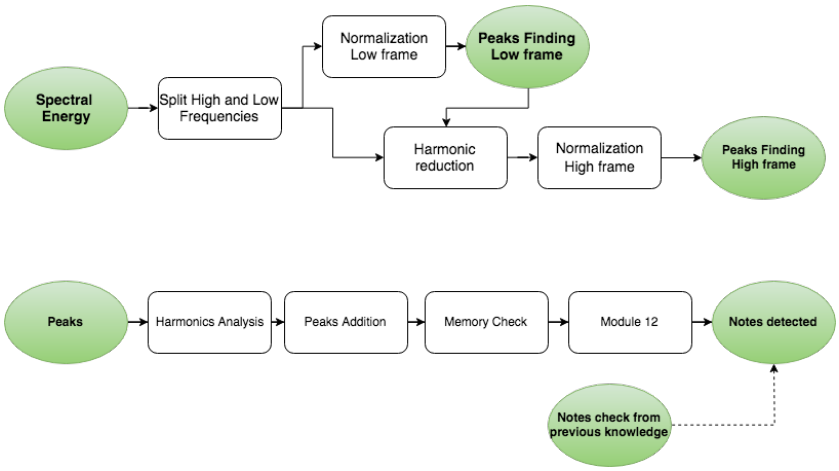


Figure 4.8: Multi-pitch detection algorithm scheme.

what the algorithm does is recognising notes along the whole keyboard through the analysis of the energy peaks and the harmonic ratios between them and then narrowing the research into a neighbourhood of the expected notes. When the user makes a mistake, the wrong note is expected to be close to the one that should be played. This constrain allows to increase the accuracy of the algorithm.

The Multi-pitch algorithm is based on a modular approach, where each part is devoted to the analysis of a specific behaviour of the signal. The Fig.4.8 is divided in two parts, corresponding to note detection and analysis and correction of the results. The modules are: i) spectral peaks detection, ii) peaks analysis, iii) expected notes comparison. The development of the algorithm started from the considerations that can be deduced from the harmonic spectrum of the signal in Fig.4.9, where the ideal normalized spectrum of a note (A3) played by a piano is shown. Its nominal fundamental frequency $f_0 = 220$ Hz corresponds to the highest peak on the left of the image. The value can be obtained using the formula:

$$f_0 = F \cdot 2^{\frac{\Delta_S}{12}} \tag{4.1}$$

Generalizing these observations, when the key of a piano is pressed, the spectrum of the audio signal ideally contains peaks, called harmonics, located to the right side of the fundamental and in harmonic relation with it, whose amplitude decreases with increasing frequency. Therefore, in a polyphonic context, the presence of the peaks associated with all the fundamentals and all the harmonics of the notes played is shown. However, the

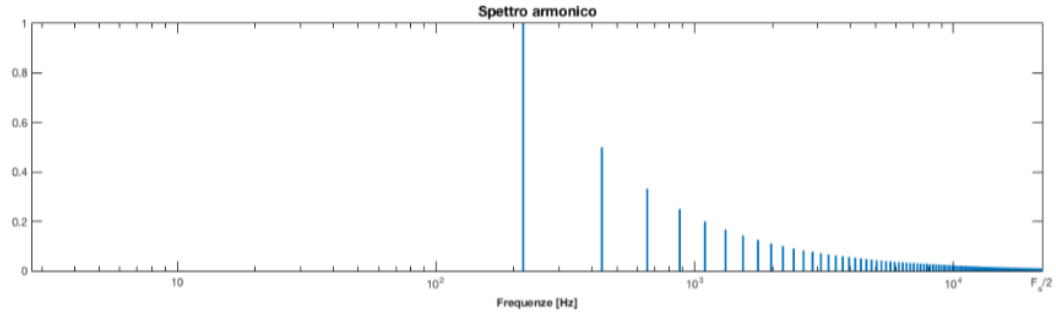


Figure 4.9: Ideal Spectrum of note A3 played by a piano.

characteristics of the real spectrum differ from those just described due to the variability of the signal. The peaks can be decentralized with respect to nominal frequencies, the amplitudes can take on values sometimes even random due to resonances or dynamics, or the signal can be disturbed by noise or pulses, for example due to the pressure of the key. Noting also exponential expansion of the Eq.4.1, it is necessary to consider that in the area of low frequencies (0, 100] Hz) there are many notes close to each other, a factor that can seriously affect the precision and therefore the performance of a detection algorithm. A further important observation, although more connected to the technologies used in *Musa* with respect to the nature of the problem, concerns the instrument used for the acquisition of the signal, and the standard microphones installed in PCs, tablets and smartphones. In fact, despite being of ever better quality, they have little sensitivity in the lower frequency band because they are thought to work in the range of human voice.

Peaks detection

The first module of the pitch recognition is devoted to the peak detection. First of all, the spectral energy X_n of the input frame is divided into two parts, $X_l(n)$ and $X_r(n)$, left and right respectively positions in analogy to the piano keyboard. The cutoff frequency f_c obtained by averaging the fundamental values corresponding to B3 (246.94 Hz) and C4 (261.63 Hz), and associated with the number 60, so that the first note is included in the left part and the second in the right one. Subsequently, the algorithm shown in Fig.4.10 is applied to the normalized left side, to detect the low frequency peaks. This procedure consists of scanning the vector containing the left part of the spectrum and

Input: $\mathbf{X}_l, f_c, \mathbf{F}$.
1: // n_w: length of FFT; Fs = Sampling Frequency
2: $i \leftarrow \text{round}(f_c \cdot n_w / F_s) - 1$.
3: $flag \leftarrow$ **false.**
4: **while** $(i \geq 0)$ **do**
5: **if** $(\mathbf{X}_l[i-1] < \mathbf{X}_l[i] \quad \&\& \quad \mathbf{X}_l[i] > \mu_l)$ **then**
6: $flag \leftarrow$ **true.**
7: $f \leftarrow i \cdot F_s / n_w$.
8: /* Search in F of the closest fundamental frequency f saved in left spectrum /*
9: **while** $(\mathbf{X}_l[i-1] < \mathbf{X}_l[i])$ **do**
10: $i \leftarrow i - 1$.
11: **if** $(flag ==$ **false**$)$ **then**
12: $i \leftarrow i - 1$.
13: $flag \leftarrow$ **false.**
14: **return** $\mathbf{P}_l$.

Figure 4.10: Algorithm for low frequencies peaks detection.

returns the indexes corresponding to the values greater than a dynamic threshold μ, which are later correlated to the fundamental of the nearest note by means of a lookup table F containing all matches between frequencies and piano notes. The peaks are found following the method of the gradient descent, solving the inverse problem, looking for the maximum values [19]. After that, notches filter, with central frequencies chosen from the ones detected in the previous step, are applied to the right side of the spectrum. This processing is necessary in order to reduce the problem introduced by the harmonics of these notes, creating further peaks that could be recognised as played notes instead of harmonics. The notes detected are stored in vector P_l.

After the normalization of vector $x_r(n)$ the algorithm in Fig.4.11 is applied to detect the peaks greater than 130 Hz. The algorithm uses banks of notch filters centered according to the fundamental frequencies of piano notes starting from C4 (131 Hz). When a given filter eliminates a high frequency peak from the spectrum, the difference between the average energy of the frequency components before filtering and the average of the same components after filtering will increase. Therefore, it can be concluded that the note associated with the specific notch filter constitutes a peak, proceeding to insert it in P_r. The threshold μ_r to establish minimum difference in energy is calculated dynamically mediating between the static value of the background noise threshold, computed at the beginning of the interaction, and the average between the differences obtained with the

Input: $\mathbf{X}_r, \mathbf{N}$.
1: $m_{pre} \leftarrow \text{mean}(\mathbf{X}_r)$.
2: **for** $(i \leftarrow 60$ **to** $108)$ **do**
3: $\quad \mathbf{m}_{post}[i] \leftarrow \text{mean}(\mathbf{X}_r \cdot \mathbf{N}[i])$.
4: $\boldsymbol{\Delta} \leftarrow m_{pre} - \mathbf{m}_{post}$.
5: $\mu_r \leftarrow \text{mean}(\sigma, \text{mean}(\boldsymbol{\Delta}))$.
6: $flag \leftarrow 1$.
7: **while** $(flag > \mu_r)$ **do**
8: $\quad j \leftarrow k : \boldsymbol{\Delta}[k] = \max\{\boldsymbol{\Delta} \setminus \mathbf{P}_r\}$.
9: $\quad$ /* Save note j in P_r /*
10: $\quad flag \leftarrow \boldsymbol{\Delta}[j]$.
11: /* Remove note j from P_r /*
12: **return** $\mathbf{P}_r$.

Figure 4.11: Algorithm for high frequencies peaks detection.

filters.

Peaks ranking

After that, the relations between peaks are analysed, in order to promote the fundamental frequencies. For this purpose, it has been developed algorithm of Fig.4.12, whose goal is to perform a ranking of the different peaks. In fact, it scans the vector containing all

Input: $\mathbf{P}_l, \mathbf{P}_r$.
1: $\mathbf{P} \leftarrow \mathbf{P}_l \cup \mathbf{P}_r$.
2: **for** $(i \leftarrow 0$ **to** $\text{length}(\mathbf{P}) - 1)$ **do**
3: $\quad a \leftarrow \mathbf{P}[i]$.
4: $\quad$ **if** $(a - 12 \in \mathbf{P})$ **then**
5: $\quad\quad \mathbf{C}[a - 12] \leftarrow \mathbf{C}[a - 12] + 1$.
6: $\quad$ **if** $(a - 19 \in \mathbf{P})$ **then**
7: $\quad\quad \mathbf{C}[a - 19] \leftarrow \mathbf{C}[a - 19] + 1$.
8: $\quad$ **if** $(a - 24 \in \mathbf{P})$ **then**
9: $\quad\quad \mathbf{C}[a - 24] \leftarrow \mathbf{C}[a - 24] + 1$.
10: $\quad\quad$ /*Save notes k -> $\mathbf{C}[k] > 0$ in **notes**. */
11: **return notes.**

Figure 4.12: Harmonic analysis and ranking.

the detected notes (both on left and on the right side of the spectrum) and it increments a counter corresponding to the specific fundamental, if a first, second or third order harmonic of a specific note are detected. The choice of stopping the analysis to the third harmonic derives from the observation that occurred during the development phase regarding the typical shape of the elaborate spectra, often disturbed by the noise in the

high frequency area, where the highest order harmonics are found. It is important to point out that, if the peak of a fundamental exists but its value is too low and it is undetected, the corresponding note is not stored in the output vector, since it is assumed that if a note was played then its peaks must also be detected. In order to avoid that a peak of a fundamental frequency is found but no harmonics are detected, a further analysis allows to introduce the notes corresponding to these peaks if the energy detected is higher than the average energy of the detected notes, regardless the number of harmonics found in the signal.

It is necessary to distinguish again between peaks of the left side of the spectrum and peaks of the right side. In the first case, the notes are included regardless of the value assumed since we accept the hypothesis that low frequencies have higher energy than high frequencies. Moreover, since the static threshold μ is set on the basis of the average energy of the entire spectrum, it is also assumed that if they have been detected then they will certainly be added. As far as the right side is concerned, the notes in P_r were saved, in the previous step, in descending order with respect to the value assumed by the corresponding peak, therefore they are saved until a more strict condition than the one of the surveyed one is respected.

Peaks Memory

Unfortunately, the audio signal is often subject to variations of an extemporaneous and transitory nature, like in-harmonic peaks due to the resonance of the piano, therefore the system has been strengthened providing it with a memory function. This method recognises notes that appear for at least two frames of signal analysis and is kept in the notes detected even if they do not appear again for maximum one frame. For clarification, each frame keeps track of the previous status of the notes detected, i.e. the number of frames passed in which every note appears, with respect to the frame immediately preceding the current one. The current status is updated based on the previous status and the notes are taken. If a note has been detected, then its counter in the current state is incremented, despite limiting its maximum value to 2, so as not to refer to a distant past more than 2 frames. If, on the other hand, a note has not been detected,

then its counter is decremented. Subsequently, a check is carried out on the states: the notes undetected in the current state are maintained equally for another frame if in the previous state they appeared from 2 frames. If in the next frame these notes are not detected for another time, then they would be erased from the memory, otherwise they would be retained.

Matching between detected and expected notes

The note detection phase, previously described, can applied to different contests of use, since its principles are based on a more general analysis of harmonic signals. However, due to the context of application, where note detection is guided by a gameplay, it is necessary to compare the results with the vector containing the expected notes for the current frame. In this function, it will be clear how *a priori* knowledge of the expected result allows to improve performance of the algorithm. First of all, before analyzing the notes individually, a collective check is performed on the algorithm output. This step can help to eliminate octave errors due to the highly harmonic nature of the spectrum: it is not unusual, in fact, that the first harmonic is detected instead of the fundamental note played. Furthermore, it is plausible and fully acceptable, in the context of using this application, for a child to perform the score correctly but mistaking the octave in which to set the hand. For these reasons, an additional indication is provided on what was found, comparing expected notes and notes taken: if all the expected ones are present in the vector of the notes taken, regardless of the octave associated with them, then a positive marker that indicates the correctness of the execution it is returned to output.

It has been noticed that, when piano is played with both hands on the keyboard, it is not rare that the harmonics of the notes of the left hand are detected as notes. This is caused by increased energy addressed to low frequency notes. Furthermore, further detection errors are made due to impurities in the signal, such as notes that are distant from the fundamentals and without any harmonic relationship with the others detected (in-harmonicity due to resonance of the piano). Therefore, all the notes that are found outside the specified intervals starting from the expected notes are deleted. The latter are divided between notes to be played with the left hand and with the right hand,

where it is assumed that a hand can cover at most one full octave, which is certainly acceptable for children. Subsequently, intervals are established on the basis of the lowest and highest notes expected to be played with two hands, allowing to distance themselves from these by a specific maximum number of semitones. In this application, the distance allowed was four semitones. This control is possible accepting the hypothesis that if an error occurs, this will happen close to the expected notes.

4.4 Performance evaluation

4.4.1 Dataset

In order to evaluate the performance of the algorithm, a dataset of piano notes has been created. In literature and in several databases online there are different datasets available, recorded in a controlled environment. However, this algorithm was created to work in real time and analysing stream audio recorded in houses or room thought for the music. This opened the need for an acoustic recording of the piano labeled and that could be subjected to reverberation of the room. Furthermore, in order to remove uncertainty due to the performance of the player during the recording, a *Yamaha Disklavier piano* has been used. This instrument is a special piano which has a system of actuator that can be controlled by MIDI messages. This allows to record several notes with a specific velocity more precisely. Therefore, the dataset has been prepared using a Matlab script to record MIDI messages. It consists of 3348 notes of 1.5 seconds length and interspersed with 0.5 seconds of silence, allowing the tail of the notes to stop resonating in the room. The entire reproduction of the dataset takes almost 56 minutes. It consists of:

- scale of bichords one tone far apart (one hand);
- bichords maximum 5 tones far apart (one hand);
- bichors played with two hands (one note each);
- two same notes but played in different octave (two hands);
- scale of C chords with three notes (one hand);

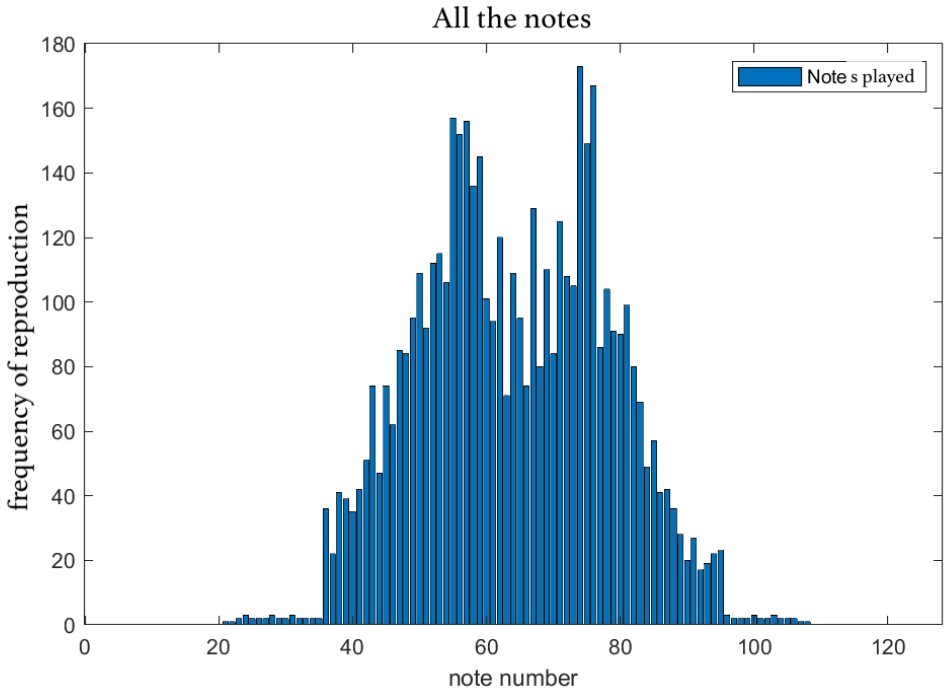


Figure 4.13: Distribution of dataset notes along the piano and their frequency of reproductions.

- bichord with left hand and one note with right hand;
- bichord with right hand and one note with left hand;
- three notes chord with left hand and one note with right hand;
- three notes chord with right hand and one note with left hand;
- bichords with both hands (two notes each);
- chords with four notes (one hand).

Apart from the scale of bichords one tone far apart, all the other notes have been recorded avoiding the two extreme octaves both in low and high frequencies, since the notes in those octaves of the piano are non harmonic due to rigidity and heaviness of the string, in low frequency range, and very short and without any resonance, in the high frequencies. In Fig.4.13 the distribution of the notes of the dataset over the keyboard and the frequency of reproductions is shown. In order to record it, a computer has been connected to the piano to reproduce the notes. The sound has been recorded using the

microphone AKG C414 in omni-directional configuration allowing to store also the room response to the sound of the piano and the *RME Fireface Uc* sound card[3]. In a frame of 4000 samples, corresponding to about 90.7 ms of audio signal with sampling frequency Fs = 44100 Hz, audio is processed on average in an interval of 3.49 ms, thus occupying the 3.85% of the duration of the window. Furthermore, since the first application of the algorithm will be the analysis of audio recorded from embedded systems, a smartphone *Huawei Mate 10 Pro*[4] has been used to record with 44100Hz as sampling frequency. Finally, the MIDI notes have been imported in the freeware DAW *Reaper*[5] and, using the virtual piano instrument provided by the software, they have been synthesised to also have a simulated reproduction of the dataset in a controlled environment as well.

4.4.2 Results

The three audio files (recording of an acoustic piano, using a smartphone and a microphone with related sound card, and piano synthesis through piano virtual instrument of *Reaper*) have been analysed using the algorithm previously described implemented in *Matlab*. The results have been analysed and compared with the correct labels of each note, obtaining the corresponding results of accuracy, precision and recall of the algorithm. Furthermore, the results have been mainly divided into number of notes played simultaneously. It is possible to divide the dataset in groups of 2, 3 and 4 notes.

Professional microphone results

The results obtained for the recordings with the microphone are shown Tab.4.1. Fig.4.14 shows the percentage results compared with the note position on the keyboard, where it is possible to highlight that in the extreme parts of the keyboard the algorithm does not recognise the notes played (as expected). The worst results are obtained with two notes, however these results are biased by the notes in the extremes of the keyboard. As previously described, only the bichord were performed all over the keyboard, where the

[3]The compact Fireface UC has been uncompromisingly optimized for highest performance under Windows and Mac OS. 24-Bit/192kHz

[4]The Huawei Mate 10 Pro is an Android smartphone designed and marketed by Huawei as part of the Huawei Mate series. It was first released on 16th October 2017.

[5]https://www.reaper.fm/

Group of notes	Accuracy	Precision	Recall
All notes	0.99	0.87	0.81
Two notes	0.99	0.86	0.78
Three notes	0.99	0.88	0.86
Four notes	0.98	0.87	0.80
Left chord Right note	0.98	0.89	0.77
Left note right chord	0.98	0.86	0.85

Table 4.1: Results of the algorithm with recordings obtained from piano Disklavier using professional microphone.

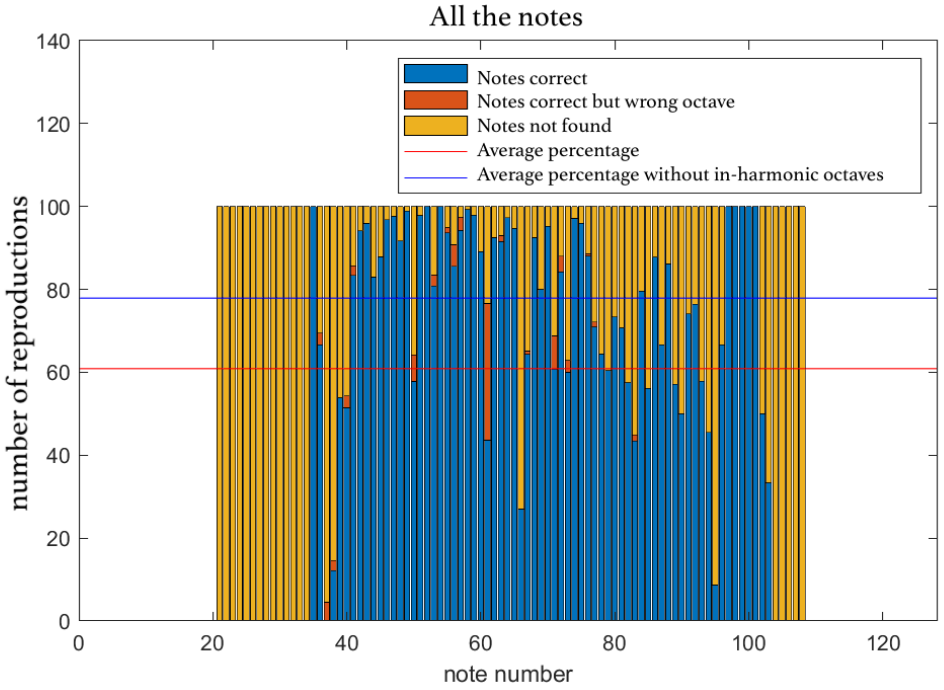


Figure 4.14: Microphone recording. Percentage of notes recognised considering the entire keyboard. Blu is the percentage of correct notes, orange is the percentage of correct notes but in a wrong octave and yellow is the percentage of notes not recognised. The blue line highlights the percentage of correct notes considering only the central portion of the keyboard. The orange line highlights the percentage of the results considering the entire keyboard.

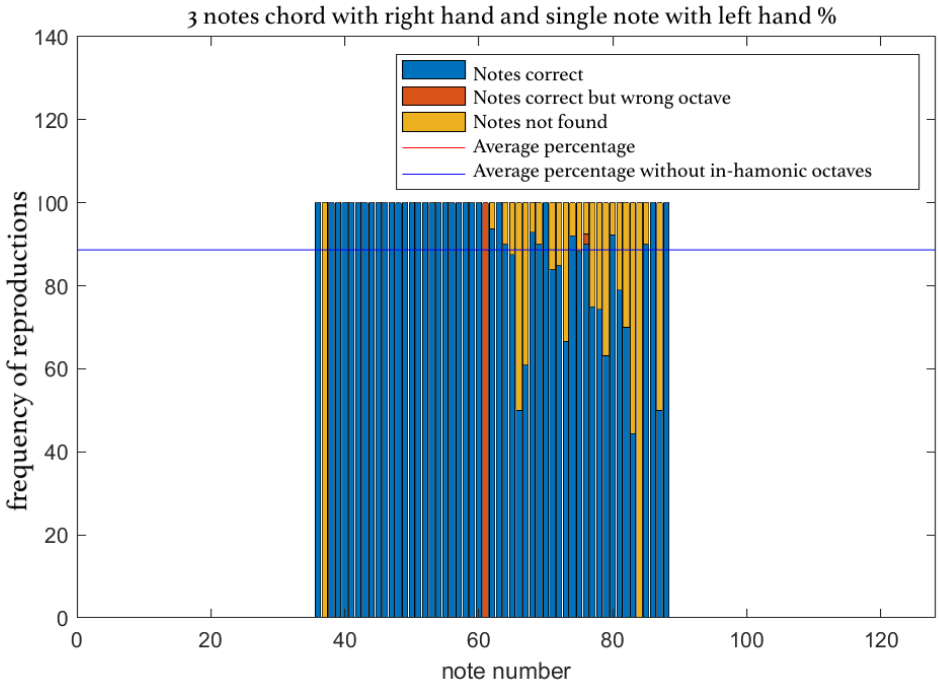


Figure 4.15: Microphone recording. Percentage of notes recognised considering chord of three notes played with the right hand and a single note played with the left hand. Blu is the percentage of correct notes, orange is the percentage of correct notes but in a wrong octave and yellow is the percentage of notes not recognised. The blue line highlights the percentage of correct notes considering only the central portion of the keyboard. The orange line highlights the percentage of the results considering the entire keyboard.

algorithm was not able to recognise the notes. Generally, the best result is obtained when three or four notes are played but there is only one note in the left side of the keyboard and the others are in the right side. This can be seen in the last row of Tab.4.1 and in Fig.4.15 where the results of three notes played with the right hand and one with the left hand are shown. To summarise, the results obtained were promising and highlighted little variation even with increasing number of notes played simultaneously.

Smartphone results

The results obtained for the recordings with the microphone are shown in Tab.4.2. It is possible to highlight that the algorithm has an overall precision higher than 84%, allowing the possibility to use the algorithm in the game *Musa* and complete the technology transfer of this project. In Fig.4.16 the percentage results compared with the notes

Group of notes	Accuracy	Precision	Recall
All notes	0.99	0.84	0.76
Two notes	0.99	0.84	0.73
Three notes	0.99	0.87	0.83
Four notes	0.98	0.82	0.74
Left chord Right note	0.98	0.84	0.73
Left note right cord	0.98	0.81	0.73

Table 4.2: Results of the algorithm with recordings obtained using a smartphone.

position on the keyboard are shown. Compared to the recording with a professional microphone, the resulting percentages are generally lower but they still allow the use of the algorithm in a contest where the technology involved is not performing and the audio is flawed by environmental noises. First of all, the decrease of the performance of the algorithm is due to the frequency response of the microphone and the high-pass filter applied to the signal present in every commercial smartphone. The reduction of precision in the low frequency range ensures a reduction of precision in the higher frequency due to the absence of harmonic reduction of the low note detected in the left side of the keyboard (as previously described). The result has a higher error rate in the high frequency notes detection, more than in the lower range.

4.5 Assessment

In order to assess some aspects of *Musa* engagement, usability and teaching ability, two different experiments have been conducted with children between the age of 6 to 11 years old. The procedure is inspired to the guidelines for usability testing with children [77, 123]. Experiment 1 evaluates some aspects of a) engagement, b) usability and c) the performance of the note detection algorithm. In Experiment 2, the effectiveness of the game to teach basic knowledge of the piano keyboard has been evaluated.

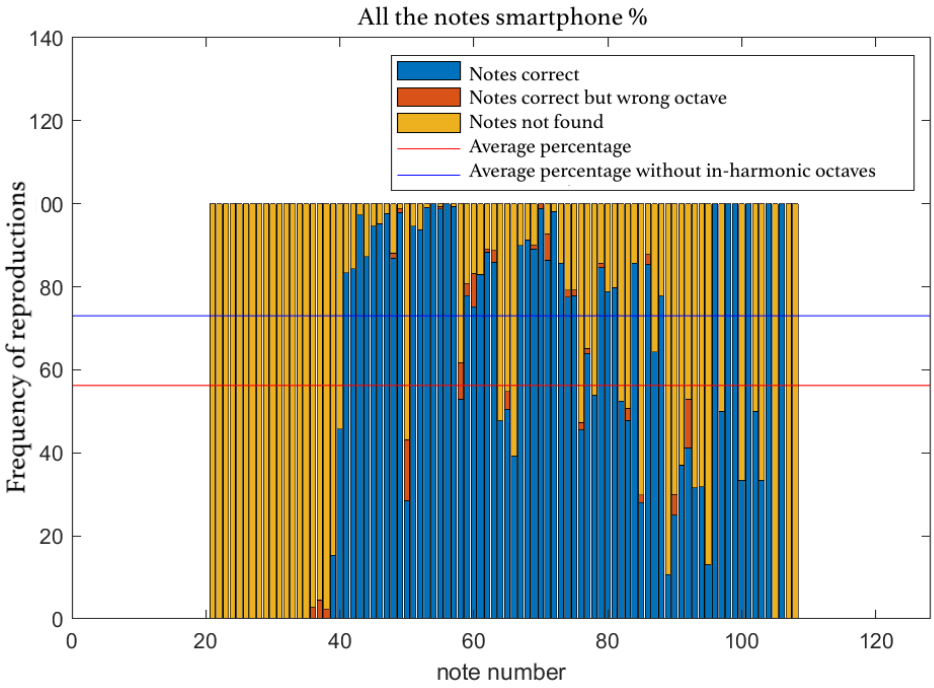


Figure 4.16: Smartphone recording. Percentage of notes recognised considering chord of three notes played with the right hand and a single note played with the left hand. Blu is the percentage of correct notes, orange is the percentage of correct notes but in a wrong octave and yellow is the percentage of notes not recognised. The blue line highlights the percentage of correct notes considering only the central portion of the keyboard. The orange line highlights the percentage of the results considering the entire keyboard.

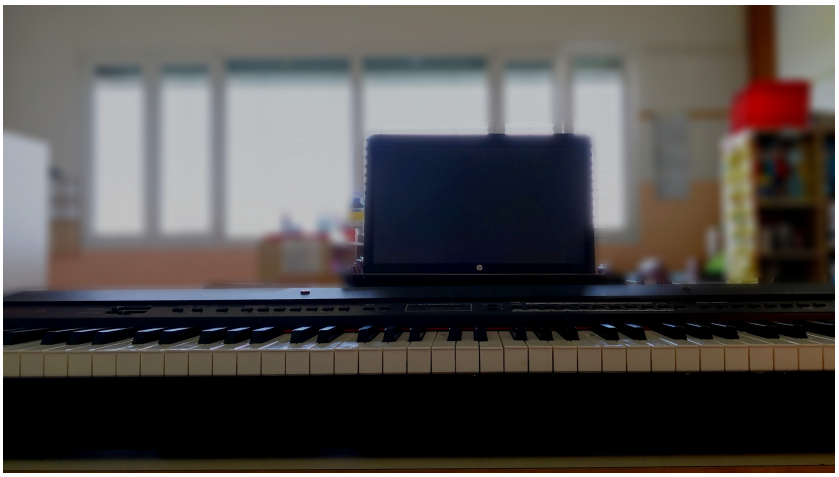


Figure 4.17: Experiment environment setting.

4.5.1 Experiment 1: engagement, usability and algorithm performance

Subjects

This test has involved 11 primary school students, 5 were males and 6 females, between the age of 6 to 9 years old (mean = 8.27, sd = 1.27). In this group, 9 children had no previous music training, while one student has been studying flute for 2 years and one guitar for 1 year. All the participants had previous experiences with videogames.

Materials

An Orla digital piano, with incorporated speakers placed on a desk in front of a 15-inch. laptop, was used to play the game (see Fig.4.17). In addition, a second laptop, with turned off screen (in order not to distract the students), was placed next to the first one to record their faces. In order to collect experimental data, the following have been used: a) pre-test questionnaire in order to collect personal information of the student and emotional state before starting the experiment; the last aspect has been collected using a 3 grade smiley faces scale; b) a post-test questionnaire in order to collect emotional state after playing the game and some questions about their appreciation of the game, reported in Tab: 4.4; c) an evaluation grid to help an observer to document relevant events and note detection algorithm errors (see Appendix for the list of observed items).

Glence Direction	Screen [s]	Keyboard [s]	Elsewhere Beginning [s]	Elsewhere Half Time [s]	Elsewhere End [s]
Mean (St.Dev)	583(113)	184(59)	6(9)	5(7)	5(7)

Table 4.3: Mean time (and standard deviation) of glances direction annotation. The elsewhere gazes are divided in the three sections of the exercise: at the beginning (first 5 min), in the middle (from 5 to 10 min) and at the end (last 5 min).

Procedure

The test for engagement, usability and algorithm performance have been conducted in the music classroom at school, in order to make the children feel comfortable and in a friendly environment. First of all, a relationship with children has been established by engaging them in some small talk to find out more about them and collect some personal information. After that, the game has been simply described (e.g. "*you can play the game by pressing the keys of the piano*") and then they have been encouraged to have fun playing the game. The test lasted 15 min for each student, at the end of which they were interrupted. The time was enough to complete the first two exercises of the game. The participants played one by one, besides them a facilitator and an observer were involved during the test [77]. At the end of the game, the post-test questionnaire has been submitted. The video of each trial has been analysed by two of the authors (observers) that manually annotated the direction of glances, using the software Elan[6].

Results

What a person is looking at is assumed to indicate the thought "on top of the stack" of cognitive processes. This eyes-mind hypothesis means that eyes movement recordings can provide a dynamic trace of where a person's attention is being directed in relation to a visual display [151]. In this publication, this method has been used to evaluate some aspects of engagement and usability of the interface, collecting information on glances direction (screen, keyboard, elsewhere) and trying to understand if a keyboard, may be a distraction or being too difficult to be used.

The eyes tracking results are shown in 4.3, where the average duration for each focus

[6]https://tla.mpi.nl/tools/tla-tools/elan/

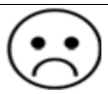
Question	How are you? [Pre-Test]	Did you like playing?	Would you like to play again?	How are you? [Post-Test]
	10	11	9	9
	1	0	2	2
	0	0	0	0

Table 4.4: Results from surveys to assess students' engagement.

direction (keeping only the annotation in common between the two observers) is reported. To analyse engagement, the level of distraction has been estimated proportionally to the duration of elsewhere glances, considered as a sign of boredom. The results highlight the children's reduced distraction and, dividing the third column of the table according to occurrence period (first five minutes, from 5 to 10 min and last five minutes) as shown in 4.3, it can be noticed that the absolute duration of the glances is not increasing with the progress in the game. The result of a one-way ANOVA test with $\alpha = 0.05$ highlights that there is not significant statistical difference. Therefore, the average time can be considered constant. The results of the questionnaire are shown in 4.4 where the percentage of students that liked the game is 100% and the percentage of interest in playing again is the 82%. The result of a t-test made on these two variables is $p = 0.34$, highlighting no significant statistical variations. This means that the percentages are not a correct representation of the distance between the two values and they are closer than what shown with these numbers.

As far as usability concerns, the analysis of some of these aspects is based on the observations collected from the facilitators involved in the test. One of the main results is the difference between first- and fourth-year primary school children: the former needed more help by the facilitators to understand the dynamic to correctly interact with the game with respect to the latter, who were more independent from external helps. Another observation is about the level of confusion that is induced by the 88 keys of the keyboard: during the initial exercises, students were confused where to focus to find the right notes, wandering around the whole keyboard. However, after this initial phase, they

started focusing on the right position for the rest of the game. The elements which helped the children were the repetitive pattern of white and black keys and keeping the hands in the same position, after understanding what was the right one for that specific exercise. Regarding the algorithms evaluation, it emerged that the voice activity detection algorithm works perfectly fine with respect to the background noise, but it is sensible to speech and creates false positives, as we expected. The result for the single pitch detection algorithm is that the ratio correct/wrong notes is 583/3. Instead, the multi-pitch detection algorithm has not been tested during this phase since playing chords is a more advanced skill and students will need a longer training on basic piano techniques.

4.5.2 Experiment 2: game teaching effectiveness

Subjects

The experiment has involved 40 primary school students from 6 to 11 years old (mean = 8.23, sd = 1.29), 22 were male and 18 were female. In this group, 22 children had no previous music training, while 11 have been studying piano for at least 1 year and 7 play another instrument. All the participants had previous experiences with videogames.

Materials

A MIDI keyboard M-audio keystation was placed on a desk in front of a 15-in. laptop, connected to it with cables and used to play the game. The sound was reproduced through the headphones Sony Mdr-zx110, since the setup has been replicated for 10 different posts. In order to collect experimental data, the following have been used: a) pre-test questionnaire in order to collect emotional state before starting the experiment and, before a second session of exercises performed the next week, to collect memory about first session; the emotional state has been collected using a 3 grade smiley faces scale; b) a post-test questionnaire in order to collect emotional state after playing the game and some questions about their appreciation of the game, as the one of the previous experiment. Other data related to the game were automatically collected in a log file.

Exercise	Age 6-7 (1st session)[s]	Age 6-7 (2nd session)[s]	Age 9-11
1	145 (60)	78 (11)	125 (9)
2	564 (202.59)	195.93 (60.82)	382.04 (172.51)
3	52.7 (28)	22.2 (8.9)	30.27 (15.09)
4	55.5 (11.78)	62.11 (29.4)	47.80 (5.18)
5	29.47 (21.95)	14.09 (9.39)	18.98 (6.78)
6	20.08 (17.2)	16.08 (4.41)	11.85 (4.13)
7	275.32 (73.08)	368.36 (220.77)	216.93 (86.59)
8	118.51 (63.41)	124.17 (51)	130.3 (46.93)
9	51.57 (20.97)	34.05 (9.76)	33.2 (6.17)
10	175.33 (105.96)	146.31 (13.47)	154.76 (73.22)

Table 4.5: Mean execution time (and standard deviation) for each exercise of experiment 2.

Procedure

The procedure show some differences with respect to the previous experiment: a) the participants could play until the end of the first level (exercises from 1 to 10 described in the previous section) with mean duration = 31 min and sd = 6.1; b) multiple tests were performed in parallel using the 10 different posts because of the reduced availability of classrooms and the high number of students involved; c) no information has been collected by the observer; d) the 11 six-year-old students took part to a second session of exercises, with the same setup and procedure, 1 week away.

Results

In the light of the usability differences highlighted in the previous experiment, data have been analysed separately between first- and fourth-primary school classes (8 students in the first group and 32 in the second). Hereafter, three different analyses are presented:

- time necessary to complete each exercise, during a game session;
- time necessary to complete each exercise, during two game sessions a week away from each other;
- time necessary to complete each exercise, separating the results of the 18 students that can play an instrument from the 22 that cannot.

Exercises	Age 6-7 (1st session)	Age 6-7 (2nd session)	Age 9-11
1-3	p <0.0125	p <0.0125	p <0.0125
1-5	p <0.0125	p <0.0125	p <0.0125
1-6	p <0.0125	p <0.0125	p <0.0125
3-5	p = 0.0368	p = 0.0465	p <0.0125
3-6	p <0.0125	p = 0.0291	p <0.0125
5-6	p <0.0125	p = 0.4034	p <0.0125

Table 4.6: Post-hoc t-test with Bonferroni correction results for exercises comparison from experiment 2.

Results presented in Tab.4.5 show to the average time that each class needed to complete the exercises. As described in the previous section, exercises 1, 3, 5, 6 are the same and are designed to teach the students to recognise a specific note (C3) on the keyboard. Analysing the time needed to complete each of them, it is possible to see a decreasing average time for both groups. As far as the younger group concerns, this trend is analysed with a one-way ANOVA test made on time data collected, which result is $F(3.28) = 20.073$ with $p < 0.005$, highlighting that the null hypothesis can be rejected between at least two groups. In order to determine the differences between the couples of exercises, a post-hoc t-test with Bonferroni correction is performed. The results are shown in Tab.4.6, highlighting that the null hypothesis is rejected for each couple of groups except between second and third exercise, which p value is 0.036, that is greater than $\alpha = 0.0125$. The same analysis is performed on the second group, highlighting that the null hypothesis is rejected for all the pairs of groups, having $F(3,128) = 984.732$ with $p < 0.05$ and having post-hoc t-test with Bonferroni correction with $p < 0.0125$ for every couple. This analysis highlight the statistically significant decreasing time necessary to conclude the exercises where is asked to play a C3. Furthermore, the resulting average time for each exercise session repeated the next week, involving the first-year students, is shown in Tab.4.5, where the same decreasing trend of time to complete the exercise of the first section is highlighted. The data collected are analysed as in the first session and the result is $F(3,28) = 148.59$ with $p < 0.05$. Furthermore, a post-hoc t-test with Bonferroni correction (see Tab.4.6) highlighted that this trend is not stable as the first session of the game: excluding the pair of the first two exercises, the others do not have significant statistical differences. It means that the execution time is no more decreasing.

Exercises	Age 6-7 (1st session)	Age 9-11
1	p <0.05	p=0.2582
3	p<0.05	p=0.2631
5	p = 0.0901	p = 0.5831
6	p = 0.4837	p = 0.9011

Table 4.7: t-test results of session comparison for the same exercises and t-test for comparison between musicians and non-musicians.

This could be explained by the results of t-test comparing the same exercise taken from the two sessions (see Tab. 4.7): since the first two exercises confirm the null hypothesis and the others reject it, it is possible to assume that the time necessary to complete the single exercise has reached its minimum in the latters. In addition, before beginning this second session of game, students were asked to point out the key to perform the "power" related to the exercises 1, 3, 5, 6. This could underline the teaching ability of the game for learning the C3 key. The result is that three students correctly recognised C3 (one of them had basic knowledge of playing piano) and one recognised C4, while another student pointed out the D4 and the last three could not answer (one of them had basic knowledge of playing piano). Ultimately, a final analysis has been done on the data collected from the fourth elementary school classes, splitting the students between the ones that can play the piano and the ones that cannot. The two groups are respectively of 8 and 24 students and the results of the t-test performed for each exercise is shown in Tab.4.7. The interesting result is that there is no statistically significant difference between the two groups, underlining that the use of the keyboard for playing the game is not easier if you have basic knowledge of the instrument.

4.5.3 Discussion on experiments results

As far as engagement and usability concern, the results obtained are encouraging and highlight that the keyboard is not a distracting element in the game design, not leading to an increasing distraction or boredom by the users. However, the results obtained with the survey highlight that the game was generally appreciated but some of the students did not want to play again the game. This could be caused by one of the exercise which involves more than one key to play. In particular, during this test, this exercise is the

one for the exploration of the keyboard, that, in the light of this test, is probably too difficult for beginners and should be either inserted in a second phase of the game or differently organised. The results obtained from these analyses underline an increment of the knowledge of the students about specific simple concepts related to the keyboard. One example is the data collected from the exercise involving the note C3, which highlight that the students learned to associate to the right key the specific magical power, both in the experiment done in two different phases and during the same session of the game. Another interesting result is that there are no significant differences between the results obtained from students who are able to play the piano and the students who are not. This result can be seen form two different viewpoints: i) the interaction with a game using a keyboard is simple for both the students' groups, ii) the difficulty comes from the game and time is more related to the exercise than to the interaction. Usability assessment presented in this paper, where no students with knowledge of the keyboard were involved, highlighted the initial difficulty to find the right position between the different octaves. Whereupon, the wandering of the gaze was no more recognisable from the eyes-tracking analysis. However, it was not possible to see a so evident learning trend for the other exercises of the game. The problem is that they were repeated fewer times during the session of the game, while the C3 exercise was repeated four times. Even in the two game sessions, no interesting increment in the performance was recognised. This probably highlights the necessity of a multiple session to appreciate an increasing trend of the performances or, as previously discussed, could be necessary to differently design these exercises.

Chapter 5

Discussion and Conclusion

This thesis aimed to open new research lines in Human Computer Interaction exploring the potentialities of sound signals, going beyond the usual verbal communication.

The main goal of the case studies presented was to solve three main issues in the context of sound-based input in natural user interfaces. First of all, the interpretation of sound signals behaviours, giving relevance to the interaction. Secondly, the modelling of real-time interaction design with pervasive systems technology and embedded systems to ensure a scalable implementation. Finally, the development of methods and assessments of Natural User Interfaces (NUIs) in the context of sound and cultural heritage, focusing on the *problem solving* capability of the system developed. This thesis tried to go beyond the simple development of new technologies/systems; indeed, it strived to ensure a complete NUI development assessing them through real applications. This process allowed to highlight the importance of multidisciplinary competence in order to merge the results and experiences of different fields to achieve the ideal NUIs development. The results of this thesis were promising and opened to a productive technology transfer between university and companies/institutions involved.

As far as broadband noise are concerned, the blow sensor and the installation for the valorization of a Pan flute from ancient Egypt is now exposed at the Museum of Archaeological Science and Arts of Padova. The user localization over a wooden runway through footsteps sound detection and the installation to valorize of a contemporary painting have been exposed during the "European Researcher's Night 2018" and it will be shown

withing the context of other exhibition such as the *Biennale di Venezia*. Furthermore, the results and the data obtained from the analysis of wave propagation along wood have been shared with the company Microtec s.r.l., which has used this information for their internal researches. Finally, the algorithm for the recognition of multi-pitched sounds, as the chords of a piano, and the data collected during the assessment of the Musa game, in which the algorithm was implemented, have been shared with the homonymous company.

In a context of Human Computer Interaction and Natural User Interfaces for a Multimodal interaction, the field of information engineering research can help to delineate methodologies and methods to study, develop and experience the user interaction. This thesis fits in this research field and it proposes new methodologies for interacting in the cultural heritage context through non verbal sounds. The methodologies applied, the models, the experiments, the results and outcomes for each case study are presented in the next sections.

5.1 Blow Sensor and Pan flute Installation

The multidisciplinary work concerning the valorization of the ancient Pan flute has brought important outcomes. The first one is the development of user blow recognition system, scalable and easy to interact with, which can be applied to different contexts regarding interaction through user's breath. A second achievement has been the development/adaptation of a methodology to develop a multimedia installation that communicates and valorizes archaeological musical instruments by considering the cultural context and the natural user interaction. The multimedia installations can be a valid means to provide an interaction with an artifact that is usually inaccessible in museums. Furthermore, the blow sensor allows to *play* the flute in the most *natural* way, i.e. blowing. The interaction model used to provide access to the general public leverages on a multimodal (visual, auditory, tactile) interplay. The blow recognition system is based on a microphone sensor and the algorithm for the interpretation of the blow behaviours, aiming at recognising the attack, sustain and release of the signal. The methodology for the installation development includes an adaptation of Design Thinking to the context of

interactive museum installations. A complete description of the system development and design process is proposed by highlighting the importance of each phase. The methodology developed as well as the practical findings can be fruitfully applied to other museum artifacts and in a varied range of cultural contexts. A group of experts evaluated this installation and the methodology through a questionnaire. The assessment results have confirmed that the methodology has led to the user engaging. Further improvements are however possible in the interaction through the sound, considering in any case the constraints of the museum. In fact, the interaction with the sound of the instrument using the blow sensor can be unnoticed from the user due to design choice. The multimodal interaction model used has solved the problem allowing the interaction with the flute sound through the touch screen. Furthermore, the sound and visual feedback allowed to create a more immersive and engaging interaction. This could open to further studies from the design point of view, in order to highlight elements of the installation without resulting inappropriate in the context of the museum. Moreover, other studies on tuning can be performed merging the musicological theory in order to provide a better approximation of the original sound. Otherwise, a new virtual instrument based on its physical model can improve the present one. A key point is how to work in a multidisciplinary team: the paradigm is not that of working separately and merging the results, but rather the idea of merging the working methods and achieving new results together.

5.2 Position Tracking and Painting Installation

The outcomes of the project concerning the valorization of contemporary paintings has brought important results. The first one is the development of a user position tracking system over a non homogeneous material as wood, scalable and that can be applied to different context where cameras are not allowed (i.e. for privacy issues) and the footsteps of the users can be the added value to the interaction. Secondly, a methodology to develop a multimedia installation that communicates and valorizes paintings by considering a 3D paradigm of exploration, instead of the usual 2D. Furthermore, a *natural* interaction for users is developed considering that the presence of a wooden runway and sound of steps easily draw the attention of the user to the paradigm of interaction. The

interaction model developed allows users interaction leveraging on a multimodal (visual, auditory, body-based) interplay. The footstep detection system is based on a network of piezoelectric sensors and the algorithm for the position detection is based on a TDOA approach, modified to avoid wrong tracking due to resonance and echoes of the wood. The methodology developed for the installation development is the result of a strong collaboration with designers, music composers and artisans highlighting the importance of a multidisciplinary collaboration. A complete description of the system development and design process is proposed by highlighting the importance of each phase. The methodology developed can be applied to other paintings and in a varied range of 3D exploration from a classical 2D paradigm. The assessment results have confirmed that the methodology has led to the user engaging. Furthermore, they stated that interacting with the runway was easy and led them to walk more. With an average score greater than 4 out of 5, most of the users agreed that the installation valorized the painting allowing them to discover peculiarities of the painting they would have overlooked otherwise. One possible further step along this project could be the use of machine learning techniques in order to map the position on the runway by means of a single sensor and studying the different features that can describe user position. Some of these techniques have already been used for indoor localization and space mapping using microphones. With this application, this type of study could be a further step to overcome the non uniform speed of propagation of the wood. In the near future, in order to save the installation at the end of the exhibition, a preservation strategy has to be implemented, creating a digital copy and transposing the concepts of active preservation used for audio documents in the field of interactive installations [29, 20, 58].

5.3 Multi-pitch Detection and Piano Teaching Game

In the related chapter, it is presented and discussed a study conducted on a videogame, as the means through which the user is induced to learn to use/play a complex and not-intuitive control interface like the keyboard of a piano (with respect to the usual controllers such as the joystick). The case study is the video game called "Musa", in which players are led into an imaginary world where music is magic. Alongside this

project, a multi-pitch algorithm for the recognition of notes played simultaneously has been developed. The results obtained, using a *score following* approach, have an average precision of recognition greater than 84%. The algorithm has been tested with a dateset of piano chords having up to 4 notes simultaneously recorded in a reverberant room. The algorithm struggles to recognise notes played in the extreme octave of the keyboard due to the in-harmonicity of the related strings. As far as the NUI assessment is concerned, the results of the two evaluation experiments showed that students were able to learn during the game some basic knowledge to play the piano after the first couple of sessions. The main acknowledgment that can be accounted for is the ability to recognise a specific key of the keyboard after 1 week from the first session of play. In addition, it has been shown that there were no differences between students that were already studying piano and the others. Some aspects of the engagement have been analysed, highlighting that there were no lack of attention during the progress of the game. In light of these results, it is possible to argue that children were learning by playing and that the playing to play paradigm could be deeper studied in the next assessments. The results of this project were so promising that the knowledge developed has been transfered to the company involved in the project responsible for the development of the game Musa (the company is homonymous to the game). In the future, a longer period of training could be performed to see the improvements of students on more complicated tasks. Furthermore, a deeper analysis with this paradigm will be performed on engagement and emotions shown while playing, in order to integrate the game with an intelligent adaptation according to the level of arousal that can be inferred by the keys' pressure. Finally, more recordings of acoustic piano performances with related labeling of the notes could be gathered, in order to try, in the future, a machine learning approach for the multi-pitch detection. The results obtained from these experiments highlight the real possibility to use this application with different instruments such as the guitar, the flute or the voice. In the near future it will be taken into account the extension to other application areas, where it is necessary to use complex devices which, like musical instruments, require face-motor coordination, such as a frame or an abacus.

5.4 Research Challenges

Natural User Interfaces is a young research field, come to light with the purpose of enabling users to use an interface with very little training, while drawing instead on experiences lived through other activities or interfaces. Nowadays, speech recognition is starting to be one of the most common interfacing technology. However, when the natural user interface paradigm started to appear in literature, gestural and tangible interfaces were more studied and considered in HCI research (almost twenty years ago). In the context of this Ph.D., the exploration of different methods for user interaction has been possible, and this highlighted how speech recognition is one of the most "natural" interaction system, even though it can be quite limiting. There are situations where speech might prove difficult to use, i.e. for children or for people with disabilities including speech-related difficulties. In these cases, giving specific commands can be harder than pressing a button or performing a specific gesture. On the other hand, noises or non verbal sounds recognition can be simpler and open to faster communication, representing action with a single sound. Indeed, complex commands are difficult to express through verbal communication. Furthermore, in the case of artistic performances and cultural heritage, different communication channels can open new ways of expression and comprehension to both performer and visitor. These are only few examples of applications of non verbal communication, where there is room for research in order to conceive new interaction methods that the user could still perceive as “natural” as possible. The results of this thesis and further projects developed during the Ph.D. highlighted that today’s companies, such as Google, are acquiring the control in the development of speech recognition systems using bigger and bigger datasets collected form smartphone users. This might provide hints for academic research to focus on new approaches that are not driven by commercial interests.

As far as NUIs are concerned, this thesis helped to understand that a multidisciplinary approach is the only way to ensure the development of new NUIs. HCI is a multidisciplinary field where engineers, designers and psychologists should work together and develop interfaces that are the result of a strong collaboration, merging the different competences to reach the final goal. From the engineering point of view, it is possible to

develop most advanced technologies; and yet, if the wrong design was applied to it the user might not understand it. *Vice versa*, a very simple and intuitive design is useless without a technology that can correctly interpret the user's will. Indeed, it often happens than natural user interfaces are considered "not natural" since their functioning needs to be learned in advance, or because much effort is asked to the user in order to understand the interaction procedure. This thesis wants to emphasize the limits of the single research field, encouraging a stronger interdisciplinary collaboration through more and more "non-disciplinary" HCI field.

Chapter 6

Appendix

6.1 Pan Flute Installation Questionnaire

The two questionnaires designed for the assessment are based on the Likert method and consist of a list of statements. For each statement, the expert was asked to indicate how much she agreed with it using a 5-level scale:

1. strongly disagree,
2. disagree,
3. neither agree nor disagree (undecided),
4. agree,
5. strongly agree.

Since it was projected that the experts would all have been Italian, the questions were drafted in Italian. An English translation of the 44 statements is provided.*Blowflute* is the name of the section related to the interaction with the installation through the blow sensors.

USER EXPERIENCE – Interaction

1. It was simple to understand how to interact with the touch screen.
2. The touch screen is a convenient means of interaction.

3. It was simple to understand how to interact with BlowFlute.
4. BlowFlute is a convenient means of interaction.
5. BlowFlute is simpler to use than the touch screen to appreciate the flute sound.
6. During the interactive experience, the sound is not a disturbing factor ("noise").
7. The interactive experience was different, in a positive sense, from that of other installations I have experienced.
8. The interactive experience was easy to understand.
9. The interactive experience was pleasant.
10. The interactive experience was fun.

USER EXPERIENCE – Communication

11. Information is presented clearly.
12. The navigation structure is easy to use.
13. The navigation bar in the "Sources" section is a convenient tool.
14. Captions and icons are clear and unambiguous.
15. Chosen colors do not create difficulties in identifying the navigation structures.
16. Visual effects are consistent with the action performed.
17. The interface is visually pleasing.
18. The variation over time of the flute sound was perceptible.
19. I easily perceived the nuances related to the sound of the flute.

MUSEUM ARCHITECTURE AND HISTORY – Product

20. The installation integrates aesthetically in the context of the room where it is located.

21. The installation aesthetically enriches the room where it is located.

22. The installation is suited to the context of the museum from an aesthetically.

23. The sound is suitable for the context of the museum (it is not a disturbing element).

MUSEUM COLLECTION – Interaction

24. The interaction through BlowFlute renders the nature of the flute as a musical instrument.

25. Touch screen interaction with the flute model renders the nature of the flute as a product.

MUSEUM COLLECTION – Product

26. The installation allows to adequately know the history of the flute.

27. The installation allows to adequately know the construction details of the flute ("the way it was built").

28. The installation allows to properly understand the analysis performed on the flute.

29. The graphical user interface effectively communicates the history of the flute.

30. The graphical user interface effectively communicates the construction details of the flute ("the way it was built").

31. The graphical user interface effectively communicates the analysis performed on the flute.

MUSEUM COLLECTION – Communication

32. The sound interface conveys additional information about the flute compared to the graphical interface.

33. The installation communicates more information than a traditional museum exhibition (artefacts in a display case with information panel).

34. The installation communicates information that would not be otherwise accessible even if the artifact could be handled and played without limitations.

MANUFACTURING OPPORTUNITIES – Interaction

35. Manipulating a virtual model of the flute is better than manipulating a physical reconstruction of the flute.

36. Manipulating the virtual model of the flute by touch screen better than other means of interaction (e.g., via hand gestures).

37. Blowing into a hole (BlowFlute) is preferable with respect to other blowing possibilities (e.g., blowing into a straw).

38. Having two different interfaces (BlowFlute and touch screen) to play the sound of the flute adds something to the experience.

39. The interactive experience was different, in a positive sense, from the one I had with other computer tools I experienced.

40. The interactive experience presents innovative aspects.

MANUFACTURING OPPORTUNITIES– Product

41. The BlowFlute functionality is well integrated into the installation.

42. The touch screen functionality is well integrated into the installation.

43. The contributions of the various professionals (designer, engineer, carpenter, ...) who worked on the installation are well integrated with one another.

44. The information on the flute (e.g., the shape of the flute) impressed me more than the technological aspects of the installation (e.g., the fact that I can rotate the model of the flute with a finger).

6.2 Painting 3D Exploration Assessment

The questionnaire designed for the assessment is based on the Likert method and consist of a list of statements. For each statement, the users were asked to indicate how much they agreed with it using a 5-level scale:

1. strongly disagree,
2. disagree,
3. neither agree nor disagree (undecided),
4. agree,
5. strongly agree.

Since it was projected that the experts would all have been Italian, the questions were drafted in Italian. An English translation of the statements is provided.

1. Understanding how to interact with the platform was easy.
2. The runway is a simple means of interaction.
3. During the interactive experience, the sound has enriched my experience.
4. Thanks to the installation, I have better appreciated the peculiarities of the picture.
5. The interactive experience was positively different from the others I had.
6. The interactive experience was pleasant.
7. The interactive experience was fun.
8. Visual effects are consistent with the picture.
9. Installation is visually pleasing.
10. The presence of the sound of footsteps made me more immersed in the experience.
11. The sound of footsteps pushed me to move more on the platform.

6.3 Multi-pitch algorithm Assessment

This information has been evaluated with a Likert scale with score between 1 and 5. List of observed items during first experiment:

1. Number of algorithm errors;
2. Level of initial clarification questions;
3. Level of understanding of the actions performed;
4. Level of interaction not required;
5. Level of attention during the game;
6. Level of time reaction when required.

Bibliography

[1] Diana Africano, Sara Berg, Kent Lindbergh, Peter Lundholm, Fredrik Nilbrink, and Anna Persson. Designing tangible interfaces for children's collaboration. In *CHI'04 extended abstracts on Human factors in computing systems*, pages 853–868. ACM, 2004.

[2] Sama'a Al Hashimi. Vocal telekinesis: towards the development of voice-physical installations. *Universal Access in the Information Society*, 8:65–75, 2009.

[3] Kazi Masudul Alam, Abu Saleh Md Mahfujur Rahman, and Abdulmotaleb El Saddik. Mobile haptic e-book system to support 3d immersive reading in ubiquitous environments. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 9(4):27, 2013.

[4] James L Alty. Can we use music in computer-human communication? In *BCS HCI*, pages 409–423, 1995.

[5] H. Amick. A frequency-dependent soil propagation model. In *in Proc. of SPIE*, 1999.

[6] Pradeep K Atrey, Namunu C Maddage, and Mohan S Kankanhalli. Audio based event detection for multimedia surveillance. In *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, volume 5, pages V–V. IEEE, 2006.

[7] Federico Avanzini, Sergio Canazza, Giovanni De Poli, Carlo Fantozzi, Edoardo Micheloni, Niccolo Pretto, Antonio Roda, Silvia Gasparotto, and Giuseppe Salemi. Virtual reconstruction of an ancient greek pan flute. In *Proceedings of the 13th*

International Conference on Sound and Music Computing Conference (SMC-2016), Hamburg, Germany, August, volume 31, 2016.

[8] Federico Avanzini, Sergio Canazza, Giovanni De Poli, Carlo Fantozzi, Niccolo Pretto, Antonio Roda, Ivana Angelini, Cinzia Bettineschi, Giulia Deotto, Emanuela Faresin, et al. Archaeology and virtual acoustics-a pan flute from ancient egypt. In *Proceedings of the 12th International Conference on Sound and Music Computing*, pages 31–36, 2015.

[9] Federico Avanzini, Sergio Canazza, Giovanni De Poli, Carlo Fantozzi, Niccolò Pretto, Antonio Rodà, Ivana Angelini, Cinzia Bettineschi, Giulia Deotto, Emanuela Faresin, Alessandra Menegazzi, Gianmario Molin, Giuseppe Salemi, and Paola Zanovello. Archaeology and virtual acoustics. a pan flute from ancient Egypt. In *Proc. Int. Conf. Sound and Music Computing (SMC2015)*, pages 31–36, Maynooth, July 2015.

[10] Loris Barbieri, Fabio Bruno, and Maurizio Muzzupappa. Virtual museum system evaluation through user studies. *Journal of Cultural Heritage*, 26:101–108, 2017.

[11] Javier A. Bargas-Avila and Kasper Hornbæk. Old wine in new bottles or novel challenges: A critical analysis of empirical studies of user experience. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 2689–2698, New York, NY, USA, 2011. ACM.

[12] Marian Stewart Bartlett, Gwen Littlewort, Ian Fasel, and Javier R Movellan. Real time face detection and facial expression recognition: Development and applications to human computer interaction. In *2003 Conference on computer vision and pattern recognition workshop*, volume 5, pages 53–53. IEEE, 2003.

[13] Emmanouil Benetos, Sebastian Ewert, and Tillman Weyde. Automatic transcription of pitched and unpitched sounds from polyphonic music. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3107–3111. IEEE, 2014.

[14] Yoshua Bengio et al. Learning deep architectures for ai. *Foundations and trends® in Machine Learning*, 2(1):1–127, 2009.

[15] Jeff A Bilmes, Xiao Li, Jonathan Malkin, Kelley Kilanski, Richard Wright, Katrin Kirchhoff, Amarnag Subramanya, Susumu Harada, James A Landay, Patricia Dowden, et al. The vocal joystick: A voice-based human-computer interface for individuals with motor impairments. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, pages 995–1002. Association for Computational Linguistics, 2005.

[16] O. Bimber, F. Coriand, A. Kleppe, E. Bruns, S. Zollmann, and T. Langlotz. Superimposing pictorial artwork with projected imagery. In *In ACM SIGGRAPH 2006 Courses*, SIGGRAPH '06. ACM, New York, NY, 2006.

[17] Michael Blommer, Norman Otto, Gregory Wakefield, Ben John Feng, and Cerita Jones. Calculating the loudness of impulsive sounds. Technical report, SAE Technical Paper, 1995.

[18] Richard A Bolt. *"Put-that-there": Voice and gesture at the graphics interface*, volume 14. ACM, 1980.

[19] Léon Bottou. Stochastic gradient descent tricks. In *Neural networks: Tricks of the trade*, pages 421–436. Springer, 2012.

[20] F. Bressan and S. Canazza. The challenge of preserving interactive sound art: a multilevel approach. *Int. J. of Arts and Technology*, 7(4):294 – 315, 2014.

[21] Federica Bressan and Sergio Canazza. The challenge of preserving interactive sound art: a multi-level approach. *International Journal of Arts and Technology*, 7(4):294–315, 2014.

[22] Tim Brown. *Change by design. How Design Thinking Transforms Organizations and Inspires Innovation.* Harper Collins Business, 2009.

[23] L. Bullivant. *Responsive Environments: Architecture, Art and Design.* V & A Publications, London, 2006.

[24] Harry Bunt, Robbert-Jan Beun, and Tijn Borghuis. *Multimodal human-computer communication: systems, techniques, and experiments*, volume 1374. Springer Science & Business Media, 1998.

[25] Christopher JC Burges, John C Platt, and Soumya Jana. Extracting noise-robust features from audio data. In *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages I–1021. IEEE, 2002.

[26] Carlos Busso, Panayiotis G Georgiou, and Shrikanth S Narayanan. Real-time monitoring of participants' interaction in a meeting using audio-visual sensors. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, volume 2, pages II–685. IEEE, 2007.

[27] Arturo Camacho. Detection of pitched/unpitched sound using pitch strength clustering. In *ISMIR*, pages 533–537, 2008.

[28] Sergio Canazza, Carlo Fantozzi, and Niccolò Pretto. Accessing tape music documents on mobile devices. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 12(1s):20, 2015.

[29] Sergio Canazza, Carlo Fantozzi, and Niccolò Pretto. Accessing tape music documents on mobile devices. *ACM Trans. Multimedia Comput. Commun. Appl.*, 12(1s):20:1–20:20, October 2015.

[30] Loïc Caroux, Katherine Isbister, Ludovic Le Bigot, and Nicolas Vibert. Player–video game interaction: A systematic review of current concepts. *Computers in Human Behavior*, 48:366–381, 2015.

[31] Fang Chen, Natalie Ruiz, Eric Choi, Julien Epps, M Asif Khawaja, Ronnie Taib, Bo Yin, and Yang Wang. Multimodal behavior and interaction as indicators of cognitive load. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 2(4):22, 2012.

[32] Selina Chu, Shrikanth Narayanan, and C-C Jay Kuo. Environmental sound recognition with time–frequency audio features. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(6):1142–1158, 2009.

[33] Edgardo Civallero. *Introducción a las flautas de pan.* Madrid, first edition, 2013. Creative Commons.

[34] Perry R Cook. *Real sound synthesis for interactive applications.* AK Peters/CRC Press, 2002.

[35] Mike Cooley. Human-centered design. *Information design*, pages 59–81, 2000.

[36] Alan Cooper, Robert Reimann, and David Cronin. *About face 3: the essentials of interaction design.* John Wiley & Sons, 2007.

[37] Dan Cosley, Jonathan Baxter, Soyoung Lee, Brian Alson, Saeko Nomura, Phil Adams, Chethan Sarabu, and Geri Gay. A tag in the hand: supporting semantic, social, and spatial navigation in museums. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1953–1962, New York, NY, USA, 2009. ACM, ACM.

[38] Rikke Dam and Teo Siang. Design thinking: Getting started with empathy. Technical report, Interaction Design Foundation, 2018.

[39] Areti Damala, Pierre Cubaud, Anne Bationo, Pascal Houlier, and Isabelle Marchal. Bridging the gap between the digital and the physical: design and evaluation of a mobile augmented reality guide for the museum visit. In *Proceedings of the 3rd international conference on Digital Interactive Media in Entertainment and Arts*, pages 120–127. ACM, 2008.

[40] Alain De Cheveigné and Hideki Kawahara. Yin, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America*, 111(4):1917–1930, 2002.

[41] Sebastien Deterding, Dan Dixon, Rilla Khaled, and Lennart Nacke. Du game design au gamefulness: définir la gamification. *Sciences du jeu*, (2), 2014.

[42] Paola Di Giuseppantonio Di Franco, Carlo Camporesi, Fabrizio Galeazzi, and Marcelo Kallmann. 3d printing and immersive visualization for improved per-

ception of ancient artifacts. *Presence: Teleoperators and Virtual Environments*, 24(3):243–264, 2015.

[43] Agostino Di Scipio. 'sound is the interface': from interactive to ecosystemic signal processing. *Organised Sound*, 8(3):269–277, 2003.

[44] Alan Dix. *Human-computer interaction*. Springer, 2009.

[45] Charles Dodge and Thomas A Jerse. *Computer music: synthesis, composition and performance*. Macmillan Library Reference, 1997.

[46] Paul Dourish. *Where the action is: the foundations of embodied interaction*. MIT press, 2004.

[47] D.Salvati and S. Canazza. Adaptive time delay estimation using filter length constraints for source localization in reverberant acoustic environments. *IEEE Signal Processing Letters*, 20(5):507–510, 2013.

[48] Alain Dufaux, Laurent Besacier, Michael Ansorge, and Fausto Pellandini. Automatic sound detection and recognition for noisy environment. In *2000 10th European Signal Processing Conference*, pages 1–4. IEEE, 2000.

[49] Ernest Edmonds. The art of interaction: what hci can learn from interactive art. *Synthesis Lectures on Human-Centered Informatics*, 11(1):i–73, 2017.

[50] Pelle Ehn. *Work-oriented design of computer artifacts*. PhD thesis, Arbetslivscentrum, 1988.

[51] M. Eisenberg, N. Elumeze, G. Blauvelt L. Buechley, S. Hendrix, and A. Eisenberg. The homespun museum: computers, fabrication, and the design of personalized exhibits. In *In Proceedings of the 5th Conference on Creativity & Cognition*, pages 13–21, ACM, New York, NY, 2005.

[52] Alexander Ekimov and James M Sabatier. Vibration and sound signatures of human footsteps in buildings. *The Journal of the Acoustical Society of America*, 118(3):2021–768, 2006.

[53] Yariv Ephraim and David Malah. Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. *IEEE Transactions on acoustics, speech, and signal processing*, 32(6):1109–1121, 1984.

[54] Florian Eyben, Martin Wöllmer, and Björn Schuller. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 1459–1462. ACM, 2010.

[55] Mark D Fairchild. *Color appearance models*. John Wiley & Sons, 2013.

[56] Federica Dal Falco and Stavros Vassos. Museum experience design: A modern storytelling methodology. *The Design Journal*, 20(sup1):S3975–S3983, 2017.

[57] Carlo Fantozzi, Federica Bressan, Niccolò Pretto, and Sergio Canazza. Tape music archives: from preservation to access. *International Journal on Digital Libraries*, 18(3):233–249, 2017.

[58] Carlo Fantozzi, Federica Bressan, Niccolò Pretto, and Sergio Canazza. Tape music archives: from preservation to access. *International Journal on Digital Libraries*, pages 1–17, 2017.

[59] Kenneth P Fishkin, Anuj Gujar, Beverly L Harrison, Thomas P Moran, and Roy Want. Embodied user interfaces for really direct manipulation. *Communications of the ACM*, 43(9):74–80, 2000.

[60] George W Fitzmaurice, Hiroshi Ishii, and William Buxton. Bricks: laying the foundations for graspable user interfaces. In *CHI*, volume 95, pages 442–449. Citeseer, 1995.

[61] Neville H. Fletcher and Thomas D. Rossing. *The physics of musical instruments*. Springer-Verlag, New York, 1991.

[62] Rita Francese, Ignazio Passero, and Genoveffa Tortora. Wiimote and kinect: gestural user interfaces add a natural third dimension to hci. In *Proceedings of the International Working Conference on Advanced Visual Interfaces*, pages 116–123. ACM, 2012.

[63] Richard J Fridrich. Percentile frequency method for evaluating impulsive sounds. Technical report, SAE Technical Paper, 1999.

[64] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*. Number 10. Springer series in statistics New York, 2001.

[65] William W Gaver. Auditory icons: Using sound in computer interfaces. *Human-computer interaction*, 2(2):167–177, 1986.

[66] Timo Gerkmann and Richard C Hendriks. Unbiased mmse-based noise power estimation with low complexity and low tracking delay. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(4):1383–1393, 2011.

[67] Alexandre Gillet, Michel Sanner, Daniel Stoffler, and Arthur Olson. Tangible interfaces for structural molecular biology. *Structure*, 13(3):483–491, 2005.

[68] Rolf Inge Godøy and Marc Leman. *Musical gestures: Sound, movement, and meaning*. Routledge, 2010.

[69] Kristen Grauman, Margrit Betke, James Gips, and Gary R Bradski. Communication via eye blinks-detection and duration analysis in real time. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, volume 1, pages I–I. IEEE, 2001.

[70] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 6645–6649. IEEE, 2013.

[71] Saul Greenberg and Chester Fitchett. Phidgets: easy development of physical interfaces through physical widgets. In *Proceedings of the 14th annual ACM symposium on User interface software and technology*, pages 209–218. ACM, 2001.

[72] K. Gronbaek, O.S. Iversen, K.J. Kortbek, K.R. Nielsen, and L. Aagaard. igame-floor - a platform for co-located collaborative games. In *In Proceedings of the International Conference on Advances in Computer Entertainment ACE*, pages 64–71, Salzburg, Austria, 2007.

[73] K. Gronbaek, O.S. Iversen, K.J. Kortbek, K.R. Nielsen, and L. Aagard. Interactive floor support for kinesthetic interaction in children learning environments. In *In Proc. of INTERACT*, Rio de Janeiro, Brazil, 2007.

[74] G.Succi, G. Prado, R. Gampert, T. Pedersen, and H. Dhaliwa. Problems in seismic detection and tracking. In *in Proc. of SPIE*, volume 404, page 165, 2000.

[75] Isabelle Guyon, Steve Gunn, Masoud Nikravesh, and Lofti A Zadeh. *Feature extraction: foundations and applications*, volume 207. Springer, 2008.

[76] Stefan Hagel. *Ancient Greek music: a new technical history*. Cambridge University Press, 2009.

[77] Libby Hanna, Kirsten Risden, and Kirsten Alexander. Guidelines for usability testing with children. *interactions*, 4(5):9–14, 1997.

[78] Chris Harrison, Hrvoje Benko, and Andrew D Wilson. Omnitouch: wearable multitouch interaction everywhere. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*, pages 441–450. ACM, 2011.

[79] Renate Häuslschmid, Benjamin Menrad, and Andreas Butz. Freehand vs. micro gestures in the car: Driving performance and user experience. In *2015 IEEE Symposium on 3D User Interfaces (3DUI)*, pages 159–160. IEEE, 2015.

[80] Geoffrey Hinton, Li Deng, Dong Yu, George Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Brian Kingsbury, et al. Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal processing magazine*, 29, 2012.

[81] Simon Holland, Andrew P McPherson, Wendy E Mackay, Marcelo M Wanderley, Michael D Gurevich, Tom W Mudd, Sile O'Modhrain, Katie L Wilkie, Joseph W Malloch, Jeremie Garcia, et al. Music and hci. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, pages 3339–3346. ACM, 2016.

[82] Simon Holland, Katie Wilkie, Paul Mulholland, and Allan Seago. Music interaction: understanding music and human-computer interaction. In *Music and human-computer interaction*, pages 1–28. Springer, 2013.

[83] D. Grahame holmes and Thomas A. Lipo. *Pulse with modulation for Power converters*. IEEE Series on power engineering.

[84] Lars Erik Holmquist, Oren Zuckerman, Rafael Ballagas, Hiroshi Ishii, Kimiko Ryokai, and Haiyan Zhang. The future of tangible user interfaces. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, page panel02. ACM, 2019.

[85] E. Hornecker and M. Stifter. Learning from interactive museum installations about interaction design for public settings. In *In Proc. of the 20th Conference of the Computer-Human interaction*, pages 135–142, OZCHI '06, vol. 206. ACM, New York, NY, 2005.

[86] Eva Hornecker and Matthias Stifter. Learning from interactive museum installations about interaction design for public settings. In *Proceedings of the 18th Australia conference on Computer-Human Interaction: Design: Activities, Artefacts and Environments*, pages 135–142, New York, NY, USA, 2006. ACM, ACM.

[87] Sherry Hsi. A study of user experiences mediated by nomadic web content in a museum. *Journal of Computer Assisted Learning*, 19(3):308–319, 2003.

[88] LJ Hu, R Desjardins, and YH Chui. Nature of vibrations induced by footsteps in lightweight and heavyweight floors. In *9th World Conference of Timber Engineering. Portland, USA*, 2006.

[89] Takeo Igarashi and John F Hughes. Voice as sound: using non-verbal voice input for interactive control. In *Proceedings of the 14th annual ACM symposium on User interface software and technology*, pages 155–156. ACM, 2001.

[90] V.T. Inman, H.J. Ralston, and F. Todd. In *Human walking*, Baltimore, London, 1981. Williams & Wilkins.

[91] Hiroshi Ishii. Tangible bits: beyond pixels. In *Proceedings of the 2nd international conference on Tangible and embedded interaction*, pages xv–xxv. ACM, 2008.

[92] Hiroshi Ishii and Brygg Ullmer. Tangible bits: towards seamless interfaces between people, bits and atoms. In *Proceedings of the ACM SIGCHI Conference on Human factors in computing systems*, pages 234–241. ACM, 1997.

[93] Satish G Iyengar, Pramod K Varshney, and Thyagaraju Damarla. On the detection of footsteps based on acoustic and seismic sensing. In *2007 Conference Record of the Forty-First Asilomar Conference on Signals, Systems and Computers*, pages 2248–2252. IEEE, 2007.

[94] Shahram Izadi, David Kim, Otmar Hilliges, David Molyneaux, Richard Newcombe, Pushmeet Kohli, Jamie Shotton, Steve Hodges, Dustin Freeman, Andrew Davison, et al. Kinectfusion: real-time 3d reconstruction and interaction using a moving depth camera. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*, pages 559–568. ACM, 2011.

[95] Robert JK Jacob, Audrey Girouard, Leanne M Hirshfield, Michael S Horn, Orit Shaer, Erin Treacy Solovey, and Jamie Zigelbaum. Reality-based interaction: a framework for post-wimp interfaces. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 201–210. ACM, 2008.

[96] Alejandro Jaimes and Nicu Sebe. Multimodal human–computer interaction: A survey. *Computer vision and image understanding*, 108(1-2):116–134, 2007.

[97] Jhilmil Jain, Arnold Lund, and Dennis Wixon. The future of natural user interfaces. In *CHI'11 Extended Abstracts on Human Factors in Computing Systems*, pages 211–214. ACM, 2011.

[98] C. Richard Johnson, William A. Sethares, and Andrew G. Klein. *The envelope of a function outlines its extremes in a smooth manner.* Software Receiver Design: Build Your Own Digital Communication System in Five Easy Steps. Cambridge University Press, 2011.

[99] Sergi Jordà, Günter Geiger, Marcos Alonso, and Martin Kaltenbrunner. The reactable: exploring the synergy between live music performance and tabletop tangible interfaces. In *Proceedings of the 1st international conference on Tangible and embedded interaction*, pages 139–146. ACM, 2007.

[100] Emil Jovanov, Dejan Raskovic, and Rick Hormigo. Thermistor-based breathing sensor for circadian rhythm evaluation. *Biomedical sciences instrumentation*, 37:493–498, 2001.

[101] Dr Manju Kaushik and Rashmi Jain. Gesture based interaction nui: an overview. *arXiv preprint arXiv:1404.2364*, 2014.

[102] Joseph'Jofish' Kaye. Some statistical analyses of chi. In *CHI'09 Extended Abstracts on Human Factors in Computing Systems*, pages 2585–2594. ACM, 2009.

[103] Damián Keller, Cláudio Gomes, and Luzilei Aliel. The handy metaphor: Bimanual, touchless interaction for the internet of musical things. *Journal of New Music Research*, pages 1–12, 2019.

[104] Annie Kelly, R Benjamin Shapiro, Jonathan de Halleux, and Thomas Ball. Arcadia: A rapid prototyping platform for real-time tangible interfaces. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, page 409. ACM, 2018.

[105] Markos Konstantakis, Konstantinos Michalakis, John Aliprantis, Eirini Kalatha, and George Caridakis. Formalising and evaluating cultural user experience. In *2017 12th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP)*, pages 90–94. IEEE, IEEE, 2017.

[106] Karen Johanne Kortbek and Kaj Grønbæk. Interactive spatial multimedia for communication of art in the physical museum space. In *Proceedings of the 16th ACM international conference on Multimedia*, pages 609–618. ACM, 2008.

[107] Ravi Kothari and Vivek Jain. Learning from labeled and unlabeled data. In *Proceedings of the 2002 International Joint Conference on Neural Networks. IJCNN'02 (Cat. No. 02CH37290)*, volume 3, pages 2803–2808. IEEE, 2002.

[108] Victoria Kravchyna and Samantha K Hastings. Informational value of museum web sites. *First Monday*, 7(2), 2002.

[109] M. W. Krueger, T. Gionfriddo, and K. Hinrichsen. Videoplace — an artificial reality. In *in Proceeding of the SIGCHI Conference on Human Factors in Computing Systems*, pages 35–40, 1985.

[110] John Krumm, Steve Harris, Brian Meyers, Barry Brumitt, Michael Hale, and Steve Shafer. Multi-camera multi-person tracking for easyliving. In *Proceedings Third IEEE International Workshop on Visual Surveillance*, pages 3–10. IEEE, 2000.

[111] Olivier Lartillot and Petri Toiviainen. A matlab toolbox for musical feature extraction from audio. In *International conference on digital audio effects*, pages 237–244. Bordeaux, 2007.

[112] Brenda Laurel. *Utopian entrepreneur*. MIT Press, Cambridge: MA, 2001.

[113] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436, 2015.

[114] Gun A Lee, Claudia Nelles, Mark Billinghurst, Mark Billinghurst, and Gerard Jounghyun Kim. Immersive authoring of tangible augmented reality applications. In *Proceedings of the 3rd IEEE/ACM international Symposium on Mixed and Augmented Reality*, pages 172–181. IEEE Computer Society, 2004.

[115] Mauro Leonardi, Adolf Mathias, and Gaspare Galati. Two efficient localization algorithms for multilateration. *International Journal of Microwave and Wireless Technologies*, 1(3):223–229, 2009.

[116] Alexander Lerch. *An introduction to audio content analysis: Applications in signal processing and music informatics*. Wiley-IEEE Press, 2012.

[117] Gail Dexter Lord, Barry Lord, Ali Hossaini, and Ngaire Blankenberg. *Manual of Digital Museum Planning*. Rowman & Littlefield Publishers, Inc., 2017.

[118] Bruno Loureiro and Rui Rodrigues. Multi-touch as a natural user interface for elders: A survey. In *6th Iberian Conference on Information Systems and Technologies (CISTI 2011)*, pages 1–6. IEEE, 2011.

[119] Marianne Lykke and Christian Jantzen. User experience dimensions: A systematic approach to experiential qualities for evaluating information interaction in museums. In *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval*, CHIIR '16, pages 81–90, New York, NY, USA, 2016. ACM.

[120] Carsten Magerkurth, Adrian David Cheok, Regan L Mandryk, and Trond Nilsen. Pervasive games: bringing computer entertainment back to the real world. *Computers in Entertainment (CIE)*, 3(3):4–4, 2005.

[121] MM Mainsbridge and Kirsty Beilharz. Body as instrument–performing with gestural interfaces. In *Proceedings of the international conference on new interfaces for musical expression*, 2014.

[122] Jani Mantyjarvi, Fabio Paternò, Zigor Salvador, and Carmen Santoro. Scan and tilt: towards natural interaction for mobile museum guides. In *Proceedings of the 8th conference on Human-computer interaction with mobile devices and services*, pages 191–194. ACM, 2006.

[123] Panos Markopoulos and Mathilde Bekker. On the assessment of usability testing methods for children. *Interacting with computers*, 15(2):227–243, 2003.

[124] Jinesh Mathew, Yuliya Semenova, and Gerald Farrell. A miniature optical breathing sensor. *Biomedical optics express*, 3(12):3325–3331, 2012.

[125] Daniel C McFarlane. Comparison of four primary methods for coordinating the interruption of people in human-computer interaction. *Human-Computer Interaction*, 17(1):63–139, 2002.

[126] Dushyant Mehta, Srinath Sridhar, Oleksandr Sotnychenko, Helge Rhodin, Mohammad Shafiei, Hans-Peter Seidel, Weipeng Xu, Dan Casas, and Christian Theobalt.

Vnect: Real-time 3d human pose estimation with a single rgb camera. *ACM Transactions on Graphics (TOG)*, 36(4):44, 2017.

[127] David Merrill, Emily Sun, and Jeevan Kalanithi. Sifteo cubes. In *CHI'12 Extended Abstracts on Human Factors in Computing Systems*, pages 1015–1018. ACM, 2012.

[128] David R Michael and Sandra L Chen. *Serious games: Games that educate, train, and inform.* Muska & Lipman/Premier-Trade, 2005.

[129] Pranav Mistry and Pattie Maes. Sixthsense: a wearable gestural interface. In *ACM SIGGRAPH ASIA 2009 Sketches*, page 11. ACM, 2009.

[130] Sushmita Mitra and Tinku Acharya. Gesture recognition: A survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 37(3):311–324, 2007.

[131] Maria Montessori and Henry Wyman Holmes. *The Montessori Method: Scientific Pedagogy as Applied to Child Education in" The Children's Houses".* Frederick A. Stokes Company, 1912.

[132] Bonnie A Nardi. *Context and consciousness: activity theory and human-computer interaction.* mit Press, 1996.

[133] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. Multimodal deep learning. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 689–696, 2011.

[134] Laurence Nigay and Joëlle Coutaz. A design space for multimodal systems: concurrent processing and data fusion. In *Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems*, pages 172–178. ACM, 1993.

[135] Luc Nijs, Micheline Lesaffre, and Marc Leman. The musical instrument as a natural extension of the musician. In *the 5th Conference of Interdisciplinary Musicology*, pages 132–133. LAM-Institut jean Le Rond d'Alembert, 2009.

[136] Kuniaki Noda, Yuki Yamaguchi, Kazuhiro Nakadai, Hiroshi G Okuno, and Tetsuya Ogata. Audio-visual speech recognition using deep learning. *Applied Intelligence*, 42(4):722–737, 2015.

[137] Donald A Norman. The next ui breakthrough, part 2: physicality. *Interactions*, 14(4):46–47, 2007.

[138] Donald A Norman. Natural user interfaces are not natural. *interactions*, 17(3):6–10, 2010.

[139] Donald A Norman and Jakob Nielsen. Gestural interfaces: a step backward in usability. *interactions*, 17(5):46–49, 2010.

[140] Jan Noyes. The qwerty keyboard: A review. *International Journal of Man-Machine Studies*, 18(3):265–281, 1983.

[141] Zeljko Obrenovic and Dusan Starcevic. Modeling multimodal human-computer interaction. *Computer*, 37(9):65–72, 2004.

[142] Yoonsin Oh and Stephen Yang. Defining exergames & exergaming. *Proceedings of Meaningful Play*, pages 1–17, 2010.

[143] Nicola Orio, Serge Lemouton, and Diemo Schwarz. Score following: State of the art and new developments. In *Proceedings of the 2003 conference on New interfaces for musical expression*, pages 36–41. National University of Singapore, 2003.

[144] Ilya V Osipov and Evgeny Nikulchev. Wowcube puzzle: A transreality object of mixed reality. In *Proceedings of the Future Technologies Conference*, pages 22–33. Springer, 2018.

[145] Antti Oulasvirta and Kasper Hornbæk. Hci research as problem-solving. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 4956–4967. ACM, 2016.

[146] Sharon Oviatt. Ten myths of multimodal interaction. *Communications of the ACM*, 42(11):74–81, 1999.

[147] Sharon Oviatt. Advances in robust multimodal interface design. *IEEE computer graphics and applications*, (5):62–68, 2003.

[148] Sharon Oviatt, Rachel Coulston, and Rebecca Lunsford. When do we interact multimodally?: cognitive load and multimodal communication patterns. In *Proceedings of the 6th international conference on Multimodal interfaces*, pages 129–136. ACM, 2004.

[149] Sungmee Park and Sundaresan Jayaraman. Enhancing the quality of life through wearable technology. *IEEE Engineering in medicine and biology magazine*, 22(3):41–48, 2003.

[150] Eva Pietroni, Alfonsina Pagano, and Claudio Rufa. The etruscanning project: Gesture-based interaction and user experience in the virtual reconstruction of the regolini-galassi tomb. In *2013 Digital Heritage International Congress (DigitalHeritage)*, volume 2, pages 653–660. IEEE, Oct 2013.

[151] Alex Poole and Linden J Ball. Eye tracking in hci and usability research. In *Encyclopedia of human computer interaction*, pages 211–219. IGI Global, 2006.

[152] Pilar Manchón Portillo, Guillermo Pérez García, and Gabriel Amores Carredano. Multimodal fusion: a new hybrid strategy for dialogue systems. In *Proceedings of the 8th international conference on Multimodal interfaces*, pages 357–363. ACM, 2006.

[153] Jenny Preece, Yvonne Rogers, Helen Sharp, David Benyon, Simon Holland, and Tom Carey. *Human-computer interaction*. Addison-Wesley Longman Ltd., 1994.

[154] Mathilde Pulh and Rémi Mencarelli. Web 2.0: Is the museum-visitor relationship being redefined? *International Journal of Arts Management*, 18(1):43–51, 2015.

[155] Francis Quek, David McNeill, Robert Bryll, Susan Duncan, Xin-Feng Ma, Cemil Kirbas, Karl E McCullough, and Rashid Ansari. Multimodal human discourse: gesture and speech. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 9(3):171–193, 2002.

[156] Siddharth S Rautaray and Anupam Agrawal. Vision based hand gesture recognition for human computer interaction: a survey. *Artificial intelligence review*, 43(1):1–54, 2015.

[157] Mitchel Resnick, Fred Martin, Robert Berg, Rick Borovoy, Vanessa Colella, Kwin Kramer, and Brian Silverman. Digital manipulatives: new toys to think with. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 281–287. ACM Press/Addison-Wesley Publishing Co., 1998.

[158] Andreas Riener. Gestural interaction in vehicular applications. *Computer*, 45(4):42–47, 2012.

[159] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *European Conference on Computer Vision*, pages 17–35. Springer, 2016.

[160] Marco Roccetti, Gustavo Marfia, and Cristian Bertuccioli. Day and night at the museum: intangible computer interfaces for public exhibitions. *Multimedia tools and applications*, 69(3):1131–1157, 2014.

[161] Maria Roussou. Learning by doing and learning through play: an exploration of interactivity in virtual environments for children. *Computers in Entertainment (CIE)*, 2(1):10–10, 2004.

[162] Szymon Rusinkiewicz, Olaf Hall-Holt, and Marc Levoy. Real-time 3d model acquisition. In *ACM Transactions on Graphics (TOG)*, volume 21, pages 438–446. ACM, 2002.

[163] James M Sabatier and Alexander E Ekimov. Range limitation for seismic footstep detection. In *Unattended Ground, Sea, and Air Sensor Technologies and Applications X*, volume 6963, page 69630V. International Society for Optics and Photonics, 2008.

[164] Dan Saffer. *Designing gestural interfaces: touchscreens and interactive devices.* " O'Reilly Media, Inc.", 2008.

[165] D. Salvati and S. Canazza. Incident signal power comparison for localization of concurrent multiple acoustic sources. *The Scientific World Journal*, 2014:13 pages, 2014.

[166] Elizabeth B-N Sanders. Postdesign and participatory culture. In *Proceedings of Useful and Critical: The Position of Research in Design*, Helsinki, 1999. University of Art and Design.

[167] Elizabeth B-N Sanders and Pieter Jan Stappers. Co-creation and the new landscapes of design. *Co-design*, 4(1):5–18, 2008.

[168] Angela Sasse, Chris Johnson, et al. Coordinating the interruption of people in human-computer interaction. In *Human-computer interaction, INTERACT*, volume 99, page 295, 1999.

[169] Roger C Schank, Tamara R Berman, and Kimberli A Macpherson. Learning by doing. *Instructional-design theories and models: A new paradigm of instructional theory*, 2(2):161–181, 1999.

[170] Bert Schiettecatte and Jean Vanderdonckt. Audiocubes: a distributed cube tangible interface based on interaction range for sound design. In *Proceedings of the 2nd international conference on Tangible and embedded interaction*, pages 3–10. ACM, 2008.

[171] Natasha Dow Schüll. Data for life: Wearable technology and the design of self-care. *BioSocieties*, 11(3):317–333, 2016.

[172] Phoebe Sengers. The engineering of experience. In *Funology 2*, pages 287–299. Springer, 2018.

[173] Ben Shneiderman. Touch screens now offer compelling uses. *IEEE software*, 8(2):93–94, 1991.

[174] Ben Shneiderman and Catherine Plaisant. *Designing the user interface: strategies for effective human-computer interaction*. Pearson Education India, 2010.

[175] Siddharth Sigtia, Emmanouil Benetos, and Simon Dixon. An end-to-end neural network for polyphonic piano music transcription. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(5):927–939, 2016.

[176] Tyler Simpson, Michel Gauthier, and Arthur Prochazka. Evaluation of tooth-click triggering and speech recognition in assistive technology for computer access. *Neurorehabilitation and neural repair*, 24(2):188–194, 2010.

[177] Jongseo Sohn, Nam Soo Kim, and Wonyong Sung. A statistical model-based voice activity detection. *IEEE signal processing letters*, 6(1):1–3, 1999.

[178] Yale Song, Louis-Philippe Morency, and Randall Davis. Multimodal human behavior analysis: learning correlation and interaction across modalities. In *Proceedings of the 14th ACM international conference on Multimodal interaction*, pages 27–30. ACM, 2012.

[179] Adam J Sporka, Sri Hastuti Kurniawan, and Pavel Slavik. Whistling user interface (u 3 i). In *ERCIM Workshop on User Interfaces for All*, pages 472–478. Springer, 2004.

[180] Barrett Stout. The harmonic structure of vowels in singing in relation to pitch and intensity. *The Journal of the Acoustical Society of America*, 10(2):137–146, 1938.

[181] G. Succi, D. Clapp, R. Gampert, and G. Prado. Footstep detection and tracking. In *in Proc. of SPIE*, volume 4393, page 22, 2001.

[182] M. Syczewska and T. Oberg. Mechanical energy levels in respect to the center of mass of trunk segments during walking in healthy and stroke subjects. In *Gait & Posture*, page 131, 2001.

[183] Ana Tajadura-Jiménez, Francisco Cuadrado, Patricia Rick, Nadia Bianchi-Berthouze, Aneesha Singh, Aleksander Väljamäe, and Frédéric Bevilacqua. Designing a gesture-sound wearable system to motivate physical activity by altering body perception. In *Proceedings of the 5th International Conference on Movement and Computing*, page 46. ACM, 2018.

[184] L. Terrenghi and A. Zimmermann. Tailored audio augmented environments for museums. In *In Proceedings of the 9th international Conference on intelligent User interfaces*, pages 334–336, IUI '04. ACM, New York, NY, 334-336., 2004.

[185] Sam Thellman, Annika Silvervarg, Agneta Gulz, and Tom Ziemke. Physical vs. virtual agent embodiment and effects on social interaction. In *International Conference on Intelligent Virtual Agents*, pages 412–415. Springer, 2016.

[186] Jenifer Tidwell. *Designing interfaces: Patterns for effective interaction design.* " O'Reilly Media, Inc.", 2010.

[187] Matthew Turk. Multimodal interaction: A review. *Pattern Recognition Letters*, 36:189–195, 2014.

[188] Michael Van den Bergh, Daniel Carton, Roderick De Nijs, Nikos Mitsou, Christian Landsiedel, Kolja Kuehnlenz, Dirk Wollherr, Luc Van Gool, and Martin Buss. Real-time 3d hand gesture interaction with a robot for understanding directions from humans. In *2011 Ro-Man*, pages 357–362. IEEE, 2011.

[189] R. Vera-Rodriguez, N.W.D. Evans, R.P. Lewis, B.Fauve, and J.S.D. Mason. An experimental study on the feasibility of footsteps as a biometric. In *in Proceedings of 15th European Signal Processing Conference (EUSIPCO'07)*, page 748–752, 2007.

[190] István A. Veres and Mahir B. Sayir. Wave propagation in a wooden bar. *Ultrasonics*, 2004.

[191] Arnold P. O. S. Vermeeren, Effie Lai-Chong Law, Virpi Roto, Marianna Obrist, Jettie Hoonhout, and Kaisa Väänänen-Vainio-Mattila. User experience evaluation methods: current state and development needs. In *Proceedings of the 6th Nordic conference on human-computer interaction: Extending boundaries*, pages 521–530, New York, NY, USA, 2010. ACM, ACM.

[192] Alex Waibel, Minh Tue Vo, Paul Duchnowski, and Stefan Manke. Multimodal interfaces. *Artificial Intelligence Review*, 10(3-4):299–319, 1996.

[193] Marcelo Mortensen Wanderley and Nicola Orio. Evaluation of input devices for musical expression: Borrowing tools from hci. *Computer Music Journal*, 26(3):62–76, 2002.

[194] Wei-Chih Wang. *Electromagnetic wave theory*. Wiley, New York, 1986.

[195] Daniel Wigdor and Dennis Wixon. *Brave NUI world: designing natural user interfaces for touch and gesture*. Elsevier, 2011.

[196] Gregg Williams. Apple macintosh computer. *Byte*, 9(2):30–31, 1984.

[197] Frank R Wilson. *The hand: How its use shapes the brain, language, and human culture*. Vintage, 1999.

[198] S. Winiarski and A. Rutkowska-Kucharska. Estimated ground reaction force in normal and pathological gait. In *Acta of Bioengineering and Biomechanics*, volume 11, 2009.

[199] Jacob O Wobbrock, Meredith Ringel Morris, and Andrew D Wilson. User-defined gestures for surface computing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1083–1092. ACM, 2009.

[200] Yansun Xu, John B Weaver, Dennis M Healy, and Jian Lu. Wavelet transform domain filters: a spatially selective noise filtration technique. *IEEE transactions on image processing*, 3(6):747–758, 1994.

[201] Gareth W Young and Dave Murphy. Hci models for digital musical instruments: Methodologies for rigorous testing of digital musical instruments. In *International Symposium on Computer Music Multidisciplinary Research (CMMR)*, 2015.

[202] Panagiotis Zaharias, Despina Michael, and Yiorgos Chrysanthou. Learning through multi-touch interfaces in museum exhibits: An empirical investigation. *Journal of Educational Technology & Society*, 16(3):374–384, 2013.

[203] Oren Zuckerman, Saeed Arida, and Mitchel Resnick. Extending tangible interfaces for education: digital montessori-inspired manipulatives. In *Proceedings of the*

SIGCHI conference on Human factors in computing systems, pages 859–868. ACM, 2005.