Preface

The De Vinci Innovation Center (DVIC) is a community of makers that develops technologies within philosophical and critical frameworks to shape our societies' futures. The objective is to implement real-world solutions as well as design projects to enhance public engagement, improve education, and overall provide scientific knowledge. Our researchers contribute actively to top-level international research in multiple fields, including artificial intelligence, human-computer interactions, education, and ecology. We believe that these objectives require a transdisciplinary approach, that bridges the gap between sciences, techniques, sociology, and philosophy. This is performed by collaborating with other scientists and industrial and startup sharing our values, to form strong research partnerships...

The Artificial Lives group, led by Dr. Clement Duhart, aims to develop the next generation of machines and Human-Machine Interfaces. The group members strongly believe that through the combination of Design and Engineering, human-centered technologies can blend into our environments to become invisible, vastly improving daily lives. To achieve this vision, the members contribute to human-computer interactions, cognitive enhancement through new forms of extended intelligence, learning platforms, and cobotic. Our bio-inspired, multidisciplinary approach couples AI and virtual reality with intelligent materials, robotics and the Internet of Things.

For the past two years, De Vinci Innovation Center (DVIC) students following the Creative Technologies curriculum had the opportunity to develop their vision on technology, innovation, and society. This proceeding is a composition of six master's theses, ranging from Machine Learning, Human-Computer-Interaction to Robotics. The authors strongly believe that developing alternative futures requires new types of engineering that take into consideration both the people's needs and the environment. These documents have been written to reflect this vision and refined over several months with an iterative reviewing supervised by the Principal Investigators.

The Authors, the Principal Investigators and the whole DVIC community is proud of releasing this first proceeding. We dedicate this first edition to Pascal Brouaye and Nelly Rouyres, without whom nothing would have been possible.

DE VINCI INNOVATION CENTER

Artificial Lives

DE VINCI HIGHER EDUCATION

List of Theses

Gamified Interfaces for Music and Sign-Language Learning and Performance Enhancement **1**



Adrien LEFEVRE - has always sought to learn different ways of being creative. He discovered electronics, computer graphics and AI at DVIC, which allowed him to carry out projects linked to his passion for music, psychology and technology.

Gamified Interfaces for Music and Sign-Language Learning and Performance Enhancement

Adrien LEFEVRE

This thesis explores the impact of gamified interfaces on learning and performance in the fields of music theory, singing, theremin, and sign language. By integrating augmented reality and tangible interfaces, the study is guided by principles from gamification, performance psychology, and multiple intelligences theory. The research yields valuable insights into optimizing engagement and enhancing information retention through these interfaces.
The presented projects include an interactive musical score, an American Sign Language learning game, and applications for singing and theremin practice. Each project employs innovative approaches to render complex subjects appealing while upholding teaching effectiveness. Future endeavors include refining these projects to further elevate user engagement and information retention, contributing to the ongoing evolution of enhanced learning interfaces.

Contents

1 Introduction **3**
1.1 Context . . . 3
1.2 Approach . . . 3
1.3 Field of research . . . 4
1.4 Contributions . . . 6

2 Interactive Musical Score **7**
2.1 Introduction . . . 7
2.2 Related work . . . 7
2.2.1 Context . . . 7
2.2.2 Approaches . . . 8
2.2.3 Musical Mental Projection . . . 9
2.2.4 Tangible Interactive Media for Music Practise . . . 9
2.3 General Architecture . . . 11
2.3.1 Overview . . . 11
2.3.2 System Design . . . 12
2.3.3 Electronic Music Box . . . 13
2.3.4 Paper Score Manufacturing . . . 14
2.3.5 Conductive Sheet Manufacturing . . . 14
2.3.6 Integration and Usability . . . 15
2.4 Applications and Evaluation . . . 16
2.4.1 Set up . . . 16
2.4.2 Results . . . 16
2.5 Discussion . . . 16
2.5.1 Attractivity . . . 16
2.5.2 Musical Imagery learning . . . 17
2.5.3 Mobilizing multiple types of attention . . . 17
2.5.4 Limitations . . . 17
2.6 Conclusion . . . 18
2.6.1 Future works . . . 19

3 Sign Language Learning Game in AR **20**
3.1 Introduction . . . 20
3.1.1 GOSAI for Augmented Mirroir . . . 21
3.1.2 Context . . . 23
3.2 Related work . . . 24
3.2.1 Sign Language Recognition . . . 24
3.2.2 Sign Language Learning Video Games . . . 26
3.2.3 Visual Novel Engines . . . 29
3.2.4 Animated 3D Avatar . . . 30
3.3 Gameplay . . . 31
3.4 General Architecture . . . 32
3.4.1 The Sign Language Video Game in GOSAI . . . 32
3.4.2 Visual Novel Engine . . . 33
3.4.3 Animations . . . 35
3.5 Sign Language Recognition AI . . . 39
3.5.1 Overview . . . 39
3.5.2 Integration in GOSAI . . . 40
3.5.3 Structure . . . 40
3.5.4 Configuration . . . 44
3.5.5 Visualization of the results . . . 45
3.5.6 Limitations . . . 46
3.6 Usage scenario . . . 46
3.7 User study . . . 47
3.7.1 Set Up . . . 47
3.7.2 Results . . . 48
3.8 Discussions . . . 51
3.9 Conclusion . . . 53
3.9.1 Playful Learning and ASL . . . 53
3.9.2 Future Works . . . 53

4 Music learning applications in AR **55**
4.1 Introduction . . . 55
4.2 Related work . . . 56
4.2.1 Singing Learning Applications . . . 56
4.2.2 Music Practise . . . 57

4.3 Singing Learning Program in AR 60
4.3.1 Context 60
4.3.2 Overview 61
4.3.3 General Architecture 63
4.3.4 User Tests 63
4.3.5 Results 65
4.3.6 Discussion 65
4.3.7 Future Work 66
4.4 Theremin Learning Program in AR 67
4.4.1 Context 67
4.4.2 Overview 68
4.4.3 General Architecture 69
4.4.4 Results 71
4.4.5 Discussion 72
4.4.6 Future work 72
4.5 Conclusion 73

5 Conclusion **74**

Introduction 1

1.1 Context

In the field of educational technology and human-computer interaction, the distinction between learning and performance forms a critical foundation for understanding how interfaces can enhance both aspects.

Learning, in essence, is the process of automating activities through repeated exposure and practice. It involves acquiring the knowledge and skills necessary to perform a task efficiently. Acquiring proficiency in a novel discipline may pose challenges attributable to the requisite integration between the processes of learning and performance.

This goes beyond the mere execution of actions. It encompasses the internalization of information, allowing individuals to reproduce movements or recall information without external prompts.

With training, the learner can perfect his ability to link an idea, a wished sign, a sound, a mental image with a movement, or a "physical" sound. This ability to translate mental concepts into concrete practice is called performance psychology.

As an illustration, it transcends the ability to simply make a movement or read a sentence, delving into the domain of retaining and reproducing these actions independently [1].

1.2 Approach

This thesis studies the influence of gamified interfaces on user performance and learning. The areas concerned include music theory, singing, theremin, and sign-language. It presents several projects that aim to exploit specific

human ways of interacting to promote the use of different intelligences, leading to the retention of information and the ability to practice efficiently.

The devices used in these projects incorporate tangible interfaces, augmented reality, sound, visual feedback, haptics, and enhanced features through electronics or software to help motivate users.

1.3 Field of research

This thesis draws inspiration from gamification, performance psychology, and the theory of multiple intelligences, all interconnected with augmented reality and tangible interfaces.

Gamification amification is a domain that seeks to make people learning subjects more engaging by incorportating playful mechanisms [2].

In recent years, gamification has revolutionized the field of professional training, accompanying the digital transformation of training and modernizing existing modules. The gamification of training involves incorporating playful mechanisms into learning content to personalize the relationship between the learner and the training material.

This technique improves learner engagement by arousing their interest and encouraging positive emotions that lead to an improvement in learning. Being immersed in an enjoyable educational experience enrich the memorization process thanks to the emotional trigger provided by the game. The playfulness of training will encourage positive emotions that lead to the improvement of learning. The effect is engagement and motivation improvement.

Augmented Reality and Tangible Interfaces Another way to improve engagement is through activating different senses. In the HCI field, Augmented Reality and

Tangible Interfaces make digital information more immersive by projecting it into the real world, engaging sight, sound, and kinesthesia [3].

The Augmented Reality (AR) experience is thriving as a significant trend. Around 2.4 billion people use AR on their mobile worldwide in 2023. AR can augment computer-generated graphics into the natural environment on screen. This augmentation can serve gamification, improving the education system efficacity and making students' attitudes more positive. It makes learning interesting, fun, and effortless, improving collaboration and capabilities.

These paradigms make the link between abstract information and the body more legible.

HCI optimizes the symbiosis between user and technology (Human-Computer Confluence) or how the elements of the human ecosystem cooperate to optimize their interaction with humans. Communication Technology (ICT) can be based on radically new forms of sensing, perception, interaction, and understanding [4]. The particularity of digital learning environments lies in the fact that they can accommodate diverse users' needs [5].

Multiple Intelligences and Multimodality The primary purpose of an interface is to allow the link with the user to a piece of equipment or an application. It is built with the aim of best optimizing its primary mission. These technologies can be used for the purpose of assisting learning. They must therefore adapt as best as possible to the specificities of human brain capacities for learning (i.e. stimulate cerebral plasticity, synaptogenesis or even neurogenesis). [6]

According to Howard Gardner [7], there are 8 forms of intelligence, each of which has certain preferential strengths. These intelligences are: logical-mathematical, verbal-linguistic, musical-rhythmic, bodily-kinesthetic, visual-spatial, interpersonal, intrapersonal and naturalist. Multiple intelligences are often used to identify the

profiles and intelligences of students, to offer them appropriate support. The use of multiple intelligences in learning or practice allows for the stimulation of different areas of the brain and thus promotes and optimizes the retention of conscious and unconscious information in many ways.

1.4 Contributions

This thesis focuses on several areas of study, including the basics of music theory using an interactive electronic score, practicing music through augmented reality learning applications, and learning and practicing sign language through a video game and an augmented reality training application.

The contributions of this work include insights into how to effectively learn and perform using augmented and tangible interfaces, promoting engagement, and enhancing the retention of information. The presented interfaces are designed to exploit brain plasticity and improve the learning process by integrating complex cognitive mechanisms. By using tangible devices, these interfaces promote sensorimotor engagement, involving the user's senses and physical movements. Gamification, based on principles of positive reinforcement and reward, activates the brain's reward system, releasing neurotransmitters such as dopamine, which are associated with motivation and learning. Synaptogenesis and long-term synaptic plasticity (LTP) are also stimulated, strengthening neuronal connections relevant for information retention.

Furthermore, by creating immersive learning environments, these interfaces exploit spatial memory and memory consolidation, thus promoting more effective knowledge retention [8]. Our contribution creates enriched learning contexts, exploiting the neurobiological bases of the cognitive process to optimize the assimilation and retention of information.

Interactive Musical Score 2

2.1 Introduction

Reading music from the score is essential to Western classical music training. Traditionally, children learn the different musical notes by singing or playing notes on an instrument guided by a teacher. We envision a way for children to learn the correspondence between notation and sound by directly touching the score. The Interactive Score is effortless and allows children to make discoveries independently. The correspondence between the visual, the tactile, and the sound can aid in learning.

In this work, we introduce the Interactive Score, a novel instrumental device for children's solfege learning. Paper scores lay onto a staff drawn with conductive ink and connected to an Adafruit musical box. Pressing a note in the score triggers its sound, and running fingers over the notes play a melody.

2.2 Related work

2.2.1 Context

A review of music theory pedagogy over the past decade reveals many criticisms about music theory courses. Other concerns include taking a harmonic, melodic, or compositional approach to teaching theory. The early involvement of students in creative thinking about harmony, melody, and rhythm partially determines their success in the academic program [9]. Developing a feeling for the mastery of the preliminaries to music and the presentation of the basics are the prerequisites for a successful future theoretical education.

However, out of the large number of students who

embark on learning music, only a few retain the courage and motivation to continue their studies to the end. For example, out of 100 French people aged 15 and over, a study recorded that only 30% of the musicians who have learned music continue practicing their instrument during their lives [10].

It is pointless and frustrating for students to be pushed too quickly into advanced and sophisticated theory without musically visualizing what they are studying.

2.2.2 Approaches

Music theory courses can take many forms. There are many approaches to developing musicianship skills. Teachers often choose between traditional or more contemporary approaches.

Traditional approach Traditional music lessons prepare students to read, write and perform music from the work of great academic composers. It produces excellent results in terms of sight-reading skills. However, many students leave these music studies because they need more enjoyment of the method.

Contemporary approach Contemporary music lessons are paired with instrumental lessons (usually piano or guitar). They do not require music reading or theory. These lessons focus on skill, musicality, and more intuitive learning methods. They require the student to have a good ear for music and a sense of rhythm. The contemporary approach benefits students with difficulty visualizing with theory classes. They can play real pieces only a few months after starting their lessons.

The primary advice if a student has any problem understanding music theory is to study with a piano. Intervals and harmonies are easier to understand on a piano first, as is all the rest of the theory.

2.2.3 Musical Mental Projection

Musical imagery, or the ability to create an image of sound in our minds, is an essential skill for all musicians. For example, brass, winds, strings, and singers imagine the pitch of an upcoming note to make it easier to play it and determine the distance from the previous note [11]. Composers and arrangers also use musical imagery when creating a new piece. Musical imagery training improves the ability to follow the upward and downward movements of the tonal contour of a musical phrase or imagined tune [12].

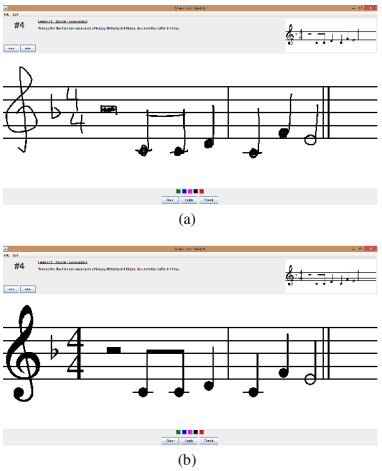
Ear training" (or solfege) has traditionally been part of the curriculum of most music schools. An essential part of solfege is the ability to read music notation and imagine how it is supposed to sound. We are interested in teaching this skill to children.

2.2.4 Tangible Interactive Media for Music Practise

Many projects aim at getting a child involved in the world of music. However, only some consist of tangible interactive media for music discovery.

For example, Zigelbaum et al. investigated how electronic instruments can engage young learners in learning to make music. Their project was the development of different tools involving movement, linking it to a sound. They created a trampoline, an interactive matrix, or musical bracelets [13].

Xiao Xiao et Al. propose an understanding of the essential workings of music without going into the details of music theory [14]. The authors expose a new technique for visualizing musical motion on a piano keyboard. The technique, called Andante, uses walking figures that move along the keyboard to represent the movement of musical phrases. The results of their user test showed that the participants found the Andante animation to be significantly more informative and engaging than the video without animation.



(a) (b)

Figure 2.1: Maestoso Educational Sketching Tool for Learning Music Theory

The work of Taele et al. [15] 2.1 describes the practical and cognitive benefits of learning music theory for both musicians and non-musicians. The paper proposes an intelligent educational tool to help students learn music theory. The tool, called Maestoso, utilizes sketch-based interaction and machine-learning techniques to provide personalized feedback to the user. The paper first introduces the importance of music education and the challenges students face when learning music theory, such as the abstract nature of music concepts and the difficulty of translating musical ideas into notation. Maestoso is a sketching interface that allows users to draw musical notes, chords, and melodies using a stylus or finger. The program uses machine learning algorithms to recognize the user's sketches and provide feedback on their accuracy and completeness.

Amico et al. discuss the development of a new type of MIDI controller designed for music education [16]. The device, called Kibo, is a tangible user interface that allows students to interact with music physically and intuitively. The authors introduce the concept of tangible user interfaces and explain how they can be used to create more immersive and interactive learning experiences. The device includes a set of modular blocks that the user can rearrange and customize to create different musical experiences.

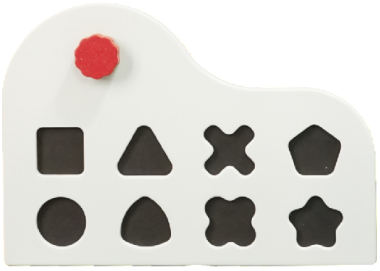

Figure 1: The interface of Kibo.

Figure 2.2: The interface of Kibo

Implementing such systems is often fully digital and interactive through a screen. Some projects aim to teach children music theory or interact with notes to compose, learn and experiment. Most of these projects are applications that users can download on smartphones or tablets. The relationship with the tangible paper score is gradually lost, and digital interaction replaces it more and more.

One solution is to use conductive ink to keep a tool in paper form without losing its interactive aspect. Conductive inks, paints, and varnishes are liquids containing metal particles, conductive polymers, or graphite. They have the specific ability to conduct electricity.

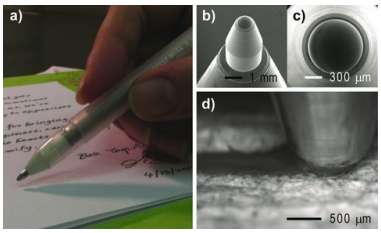


Figure 2.3: Pen-on-Paper Flexible Electronics. a) Optical image of a rollerball pen loaded with conductive silver ink. b) and c) side and top views of the rollerball pen. d) Optical image of the rollerball pen tip writing a conductive silver track

Inkjet printing allows to pattern of organic semiconductors [17], metal contacts on organic semiconductors [18] [19], and metallic structures that require minimal further processing. Researchers such as Ahn et al. used conductive ink printing to realize metallic connections between functional components of flexible devices.

Another project like that of Russo et al. has resulted in an optical image of a flexible paper display containing a LED array [20]. The prototype is a multi-color 25 × 16 LED array connected to the printed silver electrodes by depositing a drop of concentrated silver ink.

2.3 General Architecture

This project was published and demonstrated at the International Conference on New Interfaces for Musical Expression (NIME) in June 2022. NIME is a research field and conference series that explores innovative ways of creating, performing, and experiencing music through technology and interactive interfaces. It brings together experts from various disciplines such as music, engineering, design, and computer science.

2.3.1 Overview

Our design augments a traditional paper score in many digital music learning applications on screen-based devices. Children already spend a considerable amount of time in front of screens, which can harm their eyes from a young age. Paper is flexible, lightweight, and easily transportable, and incorporating electronic circuits in the paper has shown its attractiveness to children [21].

The project does not aim to teach a user how to play music. It seeks to link music theory directly with music practice without requiring knowledge.

The user has to supply the electronic part (substrate and PCB), then he places a partition (a cardboard stencil) on top of the substrate (where the conductive lines are

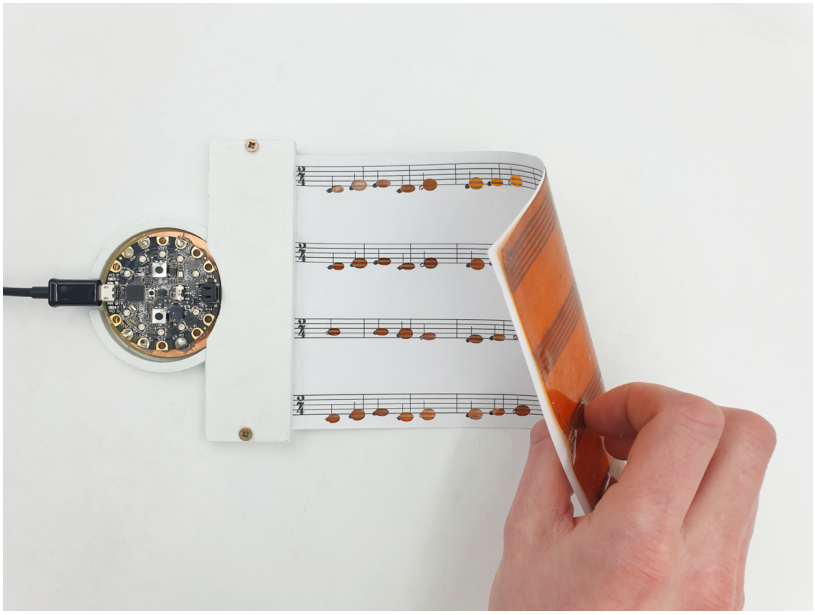


Figure 2.4: Interactive Score Prototype

located) 2.4. He can play the music and change it to another one.

2.3.2 System Design

The Interactive score consists of two thin layers. The first layer is the traditional sheet music, printed on cardstock paper, with holes punched for each note. Under this sheet is a polyimide substrate with conductive lines printed on it 2.5.

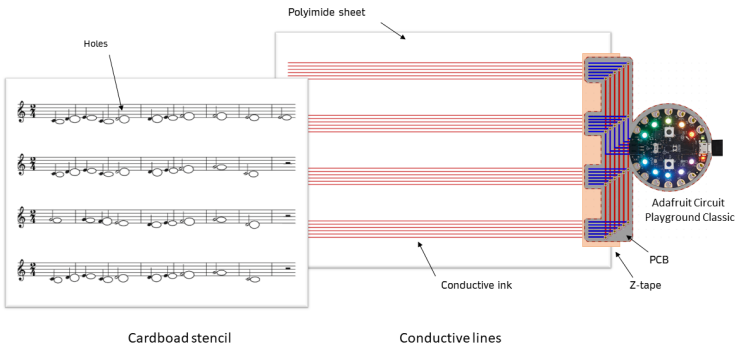


Figure 2.5: Interactive Musical Score architecture.

The conductive lines are connected to an Adafruit Circuit Playground printed circuit board (PCB) using a double-sided "z-tape".

When the user touches a note on the top layer, the finger makes contact with the conductive lines through the holes in the cardstock. The signal travels through the ink paths and the z-tape to the PCB, which detects a potential difference using capacitive touch and plays the relevant note. Detecting several simultaneous signals on multiple pins allows playing eleven different notes with only six lines 2.5.

The system is kept in a 3D printed case, maintaining contact between the polyimide, the z tape, and the PCB. The user can easily open and close the case to change the music.

2.3.3 Electronic Music Box

The signal is recovered and used in capacitive touch with an Adafruit Circuit Playground Classic 2.6. An Arduino program allows for generating a vast number of different notes. The code is retrievable on GitHub [22].

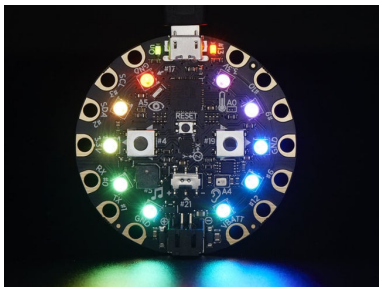


Figure 2.6: Adafruit Circuit Playground Classic

About the sound, the Adafruit Circuit Playground embarks a built-in buzzer. The implemented Mini Speaker is the SQMS5002S4036A. This device is a miniature magnetic speaker not adapted for playing detailed audio but for beeping, buzzing, and simple bleepy tunes. It is also possible to change the tone by changing the name of a variable in the code on the microcontroller.

The Adafruit circuit playground has ten mini NeoPixels of all colors, which can be animated with light when a user plays a note. The controller includes motion, temperature, light, sound sensors, a switch, and a mini speaker. This device is ideally suited for use on the score and integrates many features.

2.3.4 Paper Score Manufacturing

Notes, staves, lines, hyphenation, and cutouts were all placed on the same project for perfect dimensional compatibility between the different elements of the score. It

allows the elements to fit together perfectly and export the cutout locations at the correct size relative to the partition's rest.

An inkjet printer draws all cardboard lines, hyphens, notes, and numbers.

A score of "J'ai du bon tabac" (on the left) lies on the substrate. Then, the notes were isolated in another PNG file, selected on Cricut design, and cut directly into the cardboard with Cricut maker 3.

2.3.5 Conductive Sheet Manufacturing

The conductive ink paths are 1mm thick and 15cm long, with a resistance of 0.07 Ohms. A simple inkjet printer equipped to print with silver nanoparticles conductive ink drew the lines directly on the substrate. A hoven sintered the printed patterns at 180°C for 73 minutes. This process allows the quick production of flexible circuits [18].



Figure 2.7: Interactive Score Conductive Sheet

This prototype comprises four staffs of 6 lines made of conductive ink on a Kapton polyimide sheet substrate. This polymer does not melt at high temperatures and has excellent mechanical strength, and is very dimensionally stable and creep resistant at temperatures above 260°C. The lines are 1mm thick, and at a distance of 15cm in length, it gave a resistance of 0.07 Ohms.

An Epson WF-2010 printer printed the lines on the substrate [22].

Conductive ink allows the score to be flexible, like an actual musical score. The use of ink rather than flexible brass strips has the advantage of quickly printing interactive scores on an industrial scale. The material needed is only simple printers and ink and cleaning products for the printers. This process facilitates and accelerates production possibilities.

2.3.6 Integration and Usability

Ability to change the score

Its ease of interchangeability and production characterizes the paper score. The user can easily remove the paper sheet from the box and place another to play a different melody. All the electronic parts (substrate, PCB, microcontroller) are independent of the paper score. The sheet music has the exact dimensions of the substrate. Therefore, placing the two precisely on each other to align the holes with the ink lines is simple.

Ability to improvise

The user can improvise by not playing the notes in the same order. The project allows a great deal of modularity in its use. Just by touching specific notes at certain times, users can experiment with different rhythms and melodies and reconstruct a piece from a few notes. With a simple score, including an ascending scale, he can try his hand at composition. It is impossible to create dissonance as the Circuit Playground plays melodies in a fixed tonality set in the code. Unfortunately, this fact also limits the improvisation capacity.

2.4 Applications and Evaluation

2.4.1 Set up

Four users interact with the prototype already prepared for use. They then answer several questions. The project is plugged in and set up as a base for further use. Therefore, measuring the openness and visualization of music theory given by the project to the user is interesting. The questions concern understanding use, practicality, interest, innovation, attractiveness, and playfulness.

2.4.2 Results

The users did not encounter any particular problems during the test. The prototype worked well. The testers gave interesting feedback. The majority of the comments were positive, as the project considered most of their comments since a previous test.

The users proposed many ideas to make the project evolve. They advised adding indications, potentially with LEDs, to indicate actions to be carried out. One idea was to create a binder with different stencils inside. They would have appreciated a correction of the electronics, which would warn them if they made a mistake. The testers liked that the person playing has to press the right notes, unlike some music toys that automatically correct the sound to play the right notes wherever the user presses. The users were also very interested in the fact that the project is portable. They would have liked to be able to take it with them to play music everywhere.

2.5 Discussion

2.5.1 Attractivity

The project could be attractive for children as a tangible object. Its gamified aspect makes it close to a toy. This aspect contributes to its attractiveness, especially to young children. Its musical property favors the user's

creativity in addition to being portable, tactile, fun, and easy to use.

2.5.2 Musical Imagery learning

An essential aspect of musical practice is the ability of the musician to decipher a score. When this one practices with visual support, his brain must take the exercise to connect the note, which he links visually with the position of his fingers on the instrument. The transcription of the note read is an intermediate the brain uses, deducing a position and calculating an interval from it. The ability to translate a visual note into a sound is a reflex to be trained and very relevant in the practice of music. The basics of this ability must be acquired very quickly when learning the basics of music.

2.5.3 Mobilizing multiple types of attention

This kind of device has a genuine interest in terms of learning. The user practices bodily intelligence with touch, visual intelligence since he has a support in front of him, and musical-rhythmic intelligence.

2.5.4 Limitations

The project has some limitations being a prototype. The device is limited in terms of the number of playable notes. The conductive sheet has six lines of conductive ink, which allows for playing 13 different notes between C5 and F6. It is not possible to generate alterations while playing. This fact means that if the playground circuit is set in a specific key, it is impossible to play a note with an alteration not present in the score's framework. The key must also be changed manually at each score change (paper sheet).

The major problem of the prototype is in its transmission of electrical signals. The most crucial problem of the project is its sensitivity to the electric field due to the

use of capacitive touch. Near electric fields, the microcontroller can detect false contacts. The device must also be connected to a power outlet so that the potential difference calculated between the user's finger and the ground is remarkable. Otherwise, the playground circuit may not detect the touch of a line or start playing by itself because of the detection of surrounding electric fields. A user should therefore keep the device at a distance from electronic devices and metal surfaces.

Z-tape transmits the signal from the conductive sheet to the PCB (inside the box). This conductive tape can be damaged and cause false contacts after many uses and transport. The tape can not transmit the current to the lines, and some notes stop working.

Another problem is the conductive ink which can crumble, preventing the current from crossing certain lines. The ink traces are indeed quite fragile and easily damaged. The ink is only dried on the substrate, so its adhesion can weaken. As the user has to run his finger over the ink, he can also remove thin layers, thus causing some lines to break.

The prototype also presents a difficulty in overlaying the sheet paper on top of the polyimide sheet to align the notes with the ink lines while keeping a flexible system. The case was designed for this purpose but could be more effective.

The device also needs to be improved in terms of sound quality. The speaker driver circuitry is an on/off transistor, so the device can only play square waves. The device's loudest frequency is around 4 KHz. The sound generated is shrill, very electronic, and of poor quality. It can thus slow down the desire to practice and does not resemble the sound of an instrument.

2.6 Conclusion

In conclusion, the interactive score project has a bright future, as it implements a very recent technology to

popularize it. This process could be industrialized and even replicable for individuals with some improvements and optimizations. In the future, our goal is to make the score more accessible. Features such as flexibility and a simple "plug and play" aspect make it attractive even to children.

2.6.1 Future works

Adding the play of an alteration with three "buttons" usable thanks to the capacitive touch: flat, sharp, and natural, which the user can trigger with the index finger of the left hand, is an idea to answer the problem of changing the key. These buttons will allow the user to discover the notion of these three tools and their meaning. They would be an interesting tool that would add an improvisation capacity to the system.

A "musical tutorial" should be added so the user can listen to the score's music before playing it. The tutorial would allow the user to assimilate the musical rhythm (the time between playing each note) with the physical rhythm (the time between pressing notes).

The electronic PCB/microcontroller part should also be redesigned to no longer integrate an Adafruit Circuit Playground but a much smaller, handmade circuit. It would be possible to improve the speaker's quality and connect the device to Bluetooth or Wi-Fi to play music at a distance. The next steps are also about looking for partnerships in children's education to research experimentations on the impact of this interactive score on music assimilation. Several parameters would be evaluated, such as concentration level, playing time, and knowledge retention. It is necessary to consider different strategies to transcript musical-rhythmic on this interactive score.

Sign Language Learning Game in AR

3

3.1 Introduction

Sign Language (SL) serves as the primary means of communication for individuals who are deaf or hard-of-hearing, utilizing visual movements to facilitate interaction with the world. However, the practice of SL requires a significant investment in learning, resulting in a communication gap between individuals who are mute and those who are able-bodied. According to the World Health Organization (WHO), as of 2020, the number of people affected by hearing disabilities was estimated at 466 million, including 34 million children [23]. This number is projected to exceed 900 million by 2050.

In France alone, approximately 4 to 5 million individuals experienced difficulties or inability to communicate through speech due to hearing impairment in 2020. However, the exact figures for French Sign Language (LSF) users remain uncertain, with estimates ranging from 80,000 to 120,000, depending on the sources.

The human brain exhibits a natural affinity for learning through gestures, thanks to its innate mimicry reflex [24]. Consequently, an effective learning solution should leverage this mimetic capability to guide learners in performing accurate gestures. With the advancements in augmented reality, the possibilities of dynamic and engaging learning experiences have significantly expanded. These techniques facilitate the acquisition of skills involving spatial movements through compelling visualizations.

Enhancing the learning process of SL can be achieved by fostering user engagement, surpassing the limitations of traditional methods such as watching sign language tutorial videos, utilizing apps, or attending lessons with an instructor. Augmented reality offers the potential

to develop interactive characters that can effectively demonstrate gestures to learners. Combining such a program with gamification elements holds promise for promoting both user involvement and comprehension of movements.

The present work introduces a short video game designed for teaching sign language. While the application is intended to be used with the DVIC Interactive Mirror, it is fully functional on a standard personal computer as well.

The DVIC Augmented Mirror is a project developed with the goal of enabling the display of graphical elements overlaid onto the real environment as observed through the mirror glass. This device empowers users to monitor their own movements and compare them with a model, while also employing computer vision techniques to analyze the position of their limbs and verify the accuracy of performed gestures.

Consequently, the architecture of the learning game incorporates an artificial intelligence (AI) system for Sign Language Recognition (SLR), utilizing the Mediapipe framework to extract the user's coordinates and analyzing them through a Pytorch-based model.

3.1.1 GOSAI for Augmented Mirroir

GOSAI

GOSAI [25] is a new framework to help the development of augmented interfaces on the computer with a display. This framework targets all developers, from beginners to experts. GOSAI offers basic and often used augmented reality functionalities. Thomas Juldo is a DVIC alumnus who developed the project, the augmented mirror system, and some applications.

The system's structure allows it to reuse components between its applications and thus build a general catalog of tools that grows over time.

In addition, the framework uses mainstream program-

ming languages to allow a wide range of developers to use it. The framework is written in Python and Javascript. Python is used for the framework's core components, while Javascript is used for display.

JavaScript is a flexible programming language. It is one of the core technologies of web development, and everyone can use it on both the front-end and the back-end. It is a versatile and robust language for video games. The developers can use JavaScript to make games using a variety of platforms and tools. They can use 2d and 3d libraries combined with JavaScript to create fully-fledged games in the browser or external game engine platforms [26].

The interactive mirror

The following projects presented in this thesis are implemented on an augmented mirror running on the GOSAI software system.

The augmented mirror is a platform that provides extensive interaction between the real and the virtual world. The objective is to create a recreational, medical, and educational platform.

A one-way mirror is placed against a screen. The mirror reflects perfectly where the screen is black and can display information when the pixels emit light. A camera is placed at the top of the mirror, facing slightly down. A laptop is placed at the back of the screen (see figure 3.1).

The mirror can scan the room and detect and position objects the user interacts with. The augmented mirror uses the Intel D435 camera to estimate the position of a user in front of the mirror with Mediapipe. Thus, it can add a new dimension of interaction by exploiting kinesthesia. This dimension is an advantage over traditional interfaces using touch or a mouse.



Figure 3.1: The Augmented Mirror

3.1.2 Context

The DVIC augmented mirror project aims to change the way we perform certain tasks and disciplines in everyday life. It adds a dimension of interactivity through our reflection, our position, our movements.

Applications allow for example to have a return of its position with the display of a mediapipe skeleton above its reflection, to learn elements of sign language thanks to a training module, as well as to test other games. One of the primary uses of the mirror is to enable much better learning, using as many forms of intelligence as possible.

The mirror's augmentation gives people more awareness of their position and their movements. It allows full-body interactions with digital information, rather than accessing it through the small screens of computers and mobile devices. The added value of the project is to exploit mainly kinesthesia (which is not possible with other more traditional devices) as well as other forms of interaction. This is in addition to the various ways of interacting with the human body (human inputs/outputs) that can be exploited by the electronic devices we use every day.

This interface is ideal for the development of applications requiring movement. It is relevant for the implementation of AI-assisted sign language learning modules.

3.2 Related work

3.2.1 Sign Language Recognition



Figure 3.2: Polhemus

A wide range of domains uses SLR for different purposes. It can be found in robotics, human services, games, virtual reality applications, direct or remote communication, or HCI projects [27] (see figure 3.2).

Many early SLR systems used data gloves and accelerometers to acquire specifics of the hands. The devices measure the x/y/z positions, the orientation, and the velocity directly using a sensor such as the Polhemus tracker [28] [29] or DataGlove [30] [31] including accelerometers, gyroscopes, and electromyography sensors (see figure 3.3). Those techniques did not allow entire natural movement and constricted the mobility of the signer, altering the signs performed and being restrictive because of the need for supplementary material.

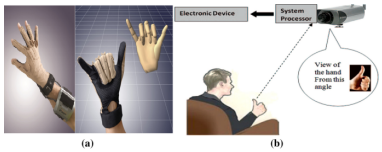


Figure 3.3: Human–computer interaction using: a CyberGlove-II [32], b vision-based system

Most techniques based on data gloves convert the position of fingers and hands according to their angles into electrical signals to obtain the desired sign.

In 2010, the ImageNet files appeared [33]. They provided a basis for the current CNNs and deep learning models, which was the beginning of computer vision. In 2012, AlexNet appeared and dramatically reduced the error rate for image recognition [34]. After the appearance of these models, the use of data gloves is gradually abandoned to focus on the implementation of modules using computer vision.

Modern Computer vision-based techniques use pose estimation on the face, body, hands, and fingers to detect their position. This method uses images or videos of the signs through a camera and calculations on the images assisted by artificial intelligence [27].

The identification of signs must take into account many different parameters. Facial expressions and body posture are vital in determining the meaning of sentences; e.g., eyebrow position can determine the question type. Some signs are distinguishable only by lip shape, as they

share a common manual sign [35].

The speed of the sign realization can change the speed induced by the performed sign. A sign can also depend on its position on the body. All limbs must therefore be taken into account during the analysis. These challenges include sensor placement, data collection and preprocessing, and model training and evaluation [36] (see figure 3.4).



Figure 3.4: Various custom gloves constructed by researchers in the sign language recognition field.

Sign language recognition systems based on computer vision and wearable sensors have been proposed by several researchers [37] [38] [39] [40] [41] [42] [42] [43], [44] [45] [46].

Most recent SLR techniques use various image or vision-based SLR systems comprising feature extraction and classification following a classical architecture [47] (see figure 3.5).

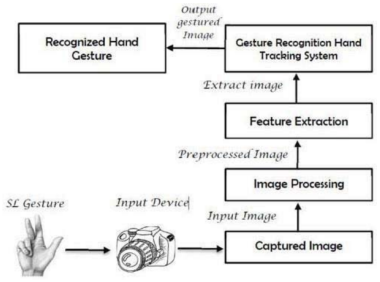


Figure 3.5: Typical Vision Based Sign Language Recognition architecture.

Many other projects using only computer vision assisted SLR also exist [47] [48] [49] [50] [51] [52] [53] [54].

Many of these projects use a Convolutional Neural Network (CNN) model for predicting the American Sign Language (ASL) alphabet [55]. Previously, classifiers like support vector machine [56], random forest, multilayer perceptron, transfer learning, and fine-tuning [57] were introduced for sign language recognition on simple images. Recently, shallow CNN and Capsule Networks have obtained better results [58].

Skeleton coordinate-based action recognition with coordinates has recently been attracting more and more attention to compute sign language videos because of its invariance to the subject or background. In contrast, skeleton coordinate-based SLR only takes the crucial data for its learning.



Figure 3.6: Top: Multi-person pose estimation. Body parts of the same person are linked, including foot key points (big toes, small toes, and heels). Bottom left: Part Affinity Fields (PAFs) corresponding to the limb connecting the right elbow and wrist. The color encodes orientation. Bottom right: A 2D vector in each pixel of every PAF encodes the position and orientation of the limbs.

The most commonly used pose estimation frameworks that extract coordinates from a person using pose estimation are, for example, OpenPose [59] (see figure 3.6), MoveNet [60], PoseNet [61] and MediaPipe [62].

Techniques using computer vision or data gloves recover

and process the coordinates with training methods. Various machine learning algorithms are used for sign language recognition, including neural networks, support vector machines, and hidden Markov models [36].

Adeyanju et al. comprehensively reviews the state-of-the-art techniques used in sign language recognition using machine learning [63] 3.7. The paper highlights the significance of sign language recognition and its potential to revolutionize communication between the deaf and hearing communities. The authors then review the different machine-learning techniques used for sign language recognition, such as Hidden Markov Models (HMMs), Support Vector Machines (SVMs), and Deep Neural Networks (DNNs). The authors also highlight the importance of datasets in sign language recognition and review some of the commonly used datasets for sign language recognition. They emphasize the need for large, diverse datasets to train machine learning models effectively.

However, sign language is more than a collection of well-specified gestures.

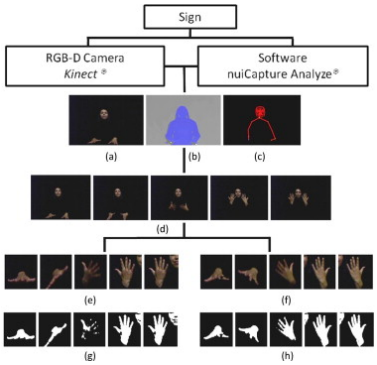


Figure 3.7: Feature extraction in Brazilian Sign Language Recognition based on phonological structure and using RGB-D sensors

3.2.2 Sign Language Learning Video Games

The video game is a dynamic audiovisual entertainment platform, accessible and stimulating the imagination of players. Using them to strengthen skills and abilities within society is possible. Video games are becoming increasingly valuable for children's development and youth culture. They enhance the function of the attentional system, stimulate visual attention, reduce reaction time, and improve the ability to discriminate shape and color, plus efficiency when following multiple objects [64].

They are an efficient didactic way to promote interest and motivation by linking playfulness and pedagogical functions [65].

The augmentation of interfaces thanks to technology, causes a better attractiveness of the learning method

and thus an increase of the time voluntarily dedicated to self-learning and the motivation to concentrate on the method [66].

Very few games use sign language as a primary element in the gameplay. We can cite Moss [67] (see figure 3.8), a video game on PlayStation VR in which the hero communicates with the player with ASL.



Figure 3.8: Moss hero communicating with the player through American sign language

Zahoor Zafrulla et al. present Copycat, a game designed to improve the American Sign Language (ASL) skills of deaf children [68]. A team of researchers developed the game at the University of California in collaboration with members of the deaf community. The game, called CopyCat, is a digital game that uses machine learning to provide feedback to the players. The game consists of a series of mini-games that focus on different aspects of ASL, such as finger spelling, vocabulary, and grammar (see figure 3.9). In each mini-game, the player watches a video clip of someone signing a word or phrase in ASL. The player is then asked to copy the sign or phrase using their signing.



Figure 3.9: Screenshot of ASL Game Interface and the input devices for user

CopyCat uses machine learning to analyze the player's signing and provide feedback on their performance. The authors designed the game to be adaptive, meaning it adjusts the difficulty level based on the player's skill level. The game also tracks the player's progress and provides feedback on improvement areas. CopyCat developers then enhanced their SLR system with the Kinect depth-mapping camera, which uses colored gloves and embedded accelerometers to track children's hand movements [69].

Bouzid et al. explore using a learning game for SignWriting, a system for writing sign languages, to enhance sign language learning for students [70]. The authors designed the game to be played on a computer or tablet. It includes various activities such as matching signs to their written symbols and translating written symbols into signs. The authors used a 3D human character to interpret the SignWriting notation. An avatar-based system called tuniSigner [71] (see figure 3.10).



Figure 3.10: The interpretation of the sign "house" via tuniSigner

Lesmes et al. discuss the development of educational video games for deaf children in order to facilitate their inclusion in mainstream educational settings [72]. The paper outlines the design and development process of educational video games, including using a participatory design approach that involves deaf children and educators throughout the design process. The authors designed the games to incorporate sign language, visual cues, and other features that would make them accessible to deaf children. The user can see an objective written on the left part of the interface, a character in the center, and an interpreter on the right (see figure 3.11).

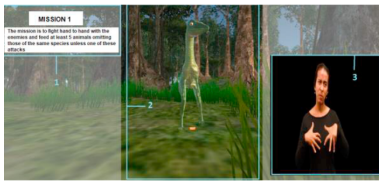


Figure 3.11: Start interface of the videogame "Life of the Dinosaurs".

Most mobile applications for sign language learning are simple quizzes with a lesson and a questionnaire. PocketSign is an application for learning American Sign Language through interactive activities (see figure 3.12). The project offers learning lessons, a dictionary for translating words into sign language videos, and a "fingerspelling" mode in which the user watches video tutorials of words or letters of the alphabet and then has to repeat them.

The application uses the phone's camera to film the user and verify that he or she is doing a word correctly before validating it and moving on to the next word.



Figure 3.12: example of exercise with a tutorial in Pocket Sign

3.2.3 Visual Novel Engines

P5VN is an open-source engine created by Mitchell G. in the Interactive Design and Media Major at NYU Tandon School of Engineering student project. The engine allows the creation of a visual novel in the web using P5.js. It enables the display of dialogues, sprites, and backgrounds and using menus by clicking (see figure 3.13). The game scenario is easily editable through a script that the engine parses. The project initially aimed to implement a prototype engine based on p5.js with a custom scripting language.



Figure 3.13: Presentation of the visual novel engine P5VN

Another visual novel engines, Tuesday JS [73] or Mono-

gatari [74] are simple web-based visual novel editors that can be used in a web browser. They are written in JavaScript, allowing the use of vector graphics svg, gif animations, and CSS styles.

Tuesay.js is an easy-to-use visual novel editor, free and open source. It runs on any web browser. The engine is written in JavaScript, does not use third-party libraries, and does not require additional software installation. It uses a drag-and-drop interface for editing scenes and creating interfaces. The script is displayed as a flowchart with all the elements and branches of the plot. The navigation is easy, and the editor helps to create stories with many plot options.

Monogatari.io is similar to Tuesday.js. The platform supports different media (images, videos, music) and multiple languages. It is highly customizable, open source, and multi-platform.

JavaScript is a language adapted for the creation of projects like this one. The language is used extensively in games that only require a few resources. A 2D interactive game displays only a few details and elements.

Frameworks like Phaser JS library or p5.js are suitable for coding a simple game. p5.js is a JavaScript library for creative coding, focusing on making coding accessible and inclusive for artists, designers, educators, beginners, and anyone else. It can display graphics and process many different elements.

3.2.4 Animated 3D Avatar

Most systems implement the display of an animated model from animation software (Blender, 3DS Max, Maya, Unity, Houdini) to create an animated avatar. The animations are worked directly in the software and then imported into a program for display in a game. Some projects allow the animation of a 3D model directly from the motion capture. In the project "Real-time Avatar Animation from a Single Image" [75] Saragih et Al. realize the modeling of 3D models from a simple



Figure 3.14: example of a real-time avatar animation from a single image thanks to semantic expression transfer.

photograph (see figure 3.14).

A tool such as Unity or Maya can make its animation from MediaPipe coordinates. Pose Animator is a tool that gives rendering for 2D animation. A demonstration works with FaceMesh and PoseNet (from MediaPipe) online [76] (see figure 3.15). Unfortunately, this one does not understand the finger movements necessary for precise sign language tutorials.



Figure 3.15: Animating full body character using FaceMesh and PoseNet with TensorFlow.js.

Another way is to adapt the coordinates in real-time to animate a character on Blender using Daz Studio as Nguyen et Al. [77] (see figure 3.16). Most interactive avatars with pose estimation use MediaPipe and TensorFlow.js (namely FaceMesh, BlazePose, and HandPose) [78] (see figure 3.17).

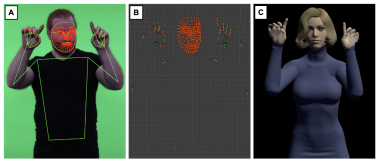


Figure 3.16: Holistic tracking applied to a video frame. A) is the annotated original footage where the red dots are the tracked landmarks, while the green lines connect the joints. B) is the same frame in Blender with landmarks plotted as orange spheres and cyan-colored bones. C) shows the motion capture data applied to an avatar. [77]

Posenet_demo [79] is a project allowing moving a 3D model with Three.js and Tensorflow's Pose Estimation model (PoseNet). The model uses Tensorflow and PoseNet to detect the critical points of the joints in each frame and then send those points over to the Model file.

The project uses an Adobe Mixamo [80] 3D model in FBX format and Blender to import the FBX format and export a GLB format [81]. The avatar follows the head and shoulder's inclination (see figure 3.18).

A last interesting example is that of Kalidokit [82]. Kalidokit is a tool that uses MediapipeTensorflow.js models for tracking face, eyes, pose, and hand movements. It is compatible with various models such as Facemesh, Blazepose, Handpose, and Holistic. The tool calculates simple Euler rotations and blendshape face values based on the predicted 3D landmarks.

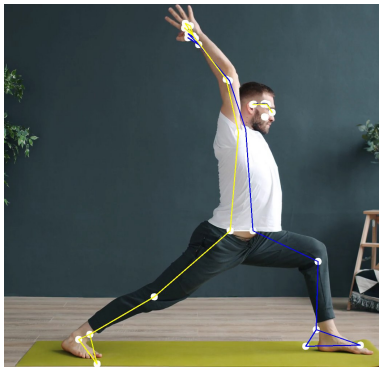


Figure 3.17: BlazePose results on yoga use-cases

Kalidokit is the core component for Vtuber web apps, such as Kalidoface and Kalidoface 3D. Its purpose is to rig 3D VRM models and Live2D avatars. The project is a JS library intended for developers using Mediapipe pretrained models rather than a complete app. The library still has to be adapted to run on different platforms. The project is based on Three.js, a powerful library for

creating three-dimensional models and games. With few lines of JavaScript, Kalidokit allows the creation of simple 3D patterns for photorealistic, real-time scenes. The library can create complex 3D geometrics and animate and move objects.

Three.js is a JavaScript library for creating 3D scenes in a web browser. It can be used with the HTML5 canvas tag without needing a plugin. The library enables the application of textures and materials. It also provides various light sources to illuminate scenes, advanced postprocessing effects, custom shaders, and load objects from 3D modeling software. It is easy to use, intuitive, and a well-documented library.

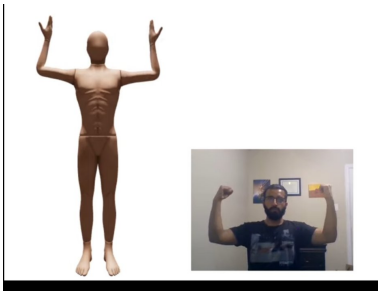


Figure 3.18: React project that will allow us to move a 3D model with Three.js (React Three Fiber) and TensorFlow's Pose Estimation model (PoseNet).



Figure 3.19: KalidoKit can move 3D avatars by tracking face and body movement with a simple webcam.

3.3 Gameplay

The sign language game is an augmented reality game, a visual novel on the augmented mirror. We follow a character during a short adventure where the user can make choices by making American Sign Language signs in front of the mirror. He can answer the characters, interact with objects, choose actions, and choose a path. Sometimes the player sees the tutorial of one, two, or three signs simultaneously. The user must then make the sign of the choice he takes. For example, he can turn right or left by making the appropriate sign (see figure 3.20).



Figure 3.20: Choice to go left or right in the sign language learning game.

3D avatars are the models for all the characters. Their limbs (including fingers) are animated. Only the character Aria (the hero's friend) is performing the sign tutorials. The camera captures the signs, and AI in the back end guesses them. Dialogue lines appear during the adventure to guide and discuss with the player. The user must make the "ok" sign to scroll the text with his hand.

3.4 General Architecture

3.4.1 The Sign Language Video Game in GOSAI

The application is connected to the get_sign and SLR drivers of GOSAI to estimate the sign of a person in front of the mirror. The socket.io then transmits the data to the engine.js, which runs the game, considering the user's movements (see figure 3.21).

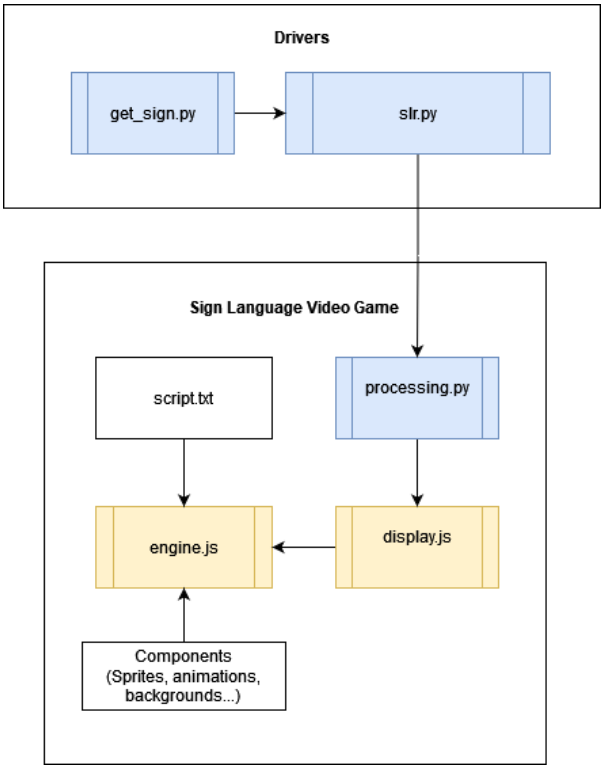


Figure 3.21: The Sign Language Video Game architecture in GOSAI.

The engine.js accesses the script containing the whole game and the components contained locally (sprites, animations, or backgrounds).

3.4.2 Visual Novel Engine

Engine implementation

A visual novel engine had to be implemented to enable the game to function. The P5VN engine is precisely adaptable for the mirror because the engine works with, library already installed on the mirror and used by the majority of its applications.

Previously, p5VN could load and display background, sprites of characters, some text interactive with the mouse click, and menus with multiple buttons. The engine was running synchronously in a single thread. Thanks to an important adaptation, the module is now running asynchronously to be compatible with GOSAI. The engine loads video animations automatically at launch and can then play them. The user can now interact and select menus thanks to sign language and added many other features.

Menu

In a visual novel, the player does not interact directly with the keyboard but must click on menu buttons to make choices. As the user interacts with the mirror only by the estimated pose, the menu system had to be adapted to enable debugging by clicking and making choices with movements.

Each time a menu appears, it takes the form of two or three words aligned and distributed horizontally on the screen. An avatar of Aria (the principal character) appears behind each word to demonstrate the sign related to this word (see figure 3.20). The location of the words and the 3D tutorials are spread over the screen's width according to the number of buttons in the displayed menu.

Aria's avatar animation plays in a loop until the player makes a sign detected by the SLR module and included in the menu.

Script

A script.txt file contains the game script in the application components. It contains a set of commands starting with the $ symbol telling the game to do a particular action. These commands can be :

- $tag indicating a specific point in the script
- $jump to jump to a tag
- $defineC to define a character (name, sprite address, text color)
- $defineImg to define an image
- $bg to display a background
- $show to display a character
- $addAnimation to play an animation to a character
- $setSprite to display the sprite of a character
- $menu to create a new menu
- $hide to hide a character
- $setvar to set a variable to a value
- $if to manage if then else commands

All the loading of sprites, animations, playback, and display are now managed automatically at startup in asynchronous instead of being set manually in the script or the code as before. The engine was running in synchronous mode at the beginning. The game starts on a loading time at startup corresponding to the loading of sprites and animation videos in the memory of GOSAI. This loading allows the game to play each animation instantly. The game directs the chosen path toward the continuity of the chosen branch. The player can choose different paths to finish the game.

3.4.3 Animations

A module was implemented on the mirror to create accurate custom 3D animations for the animated sign language tutorials. The program allows one to control an avatar in 3D through pose estimation and then record the movements produced or a video of the animation.

An application to control an avatar remotely has been developed on the augmented mirror thanks to the esti-

mation pose. The avatar can track and copy the position of the user's limbs accurately (see figure 3.22). The created application is now a free demonstration on the mirror. It also allowed the creation of all the 3D character animations on the sign language learning game.

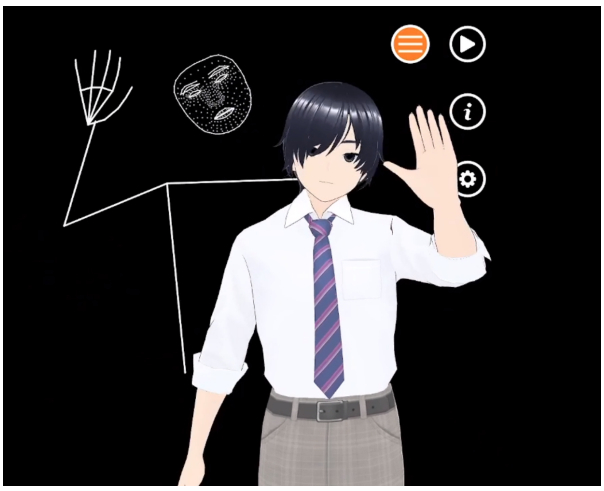


Figure 3.22: The first version of the Augmented Reality Interactive Avatar (ARIA).

Esla

The Educative Sign Language Avatar (ESLA) is a 3-dimensional avatar controllable in a GOSAI application following the estimated MediaPipe pose.

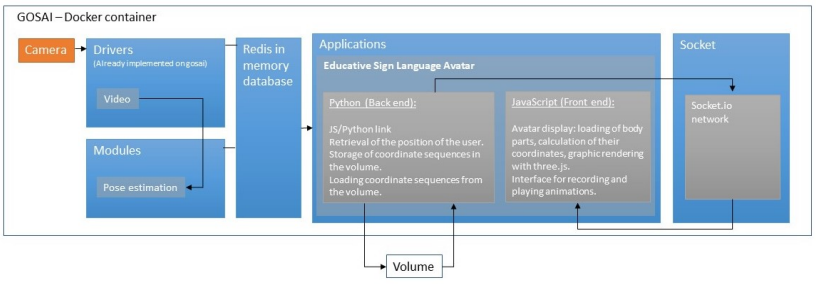


Figure 3.23: ESLA architecture on GOSAI.

The program uses the driver of the mirror camera, processed by an estimation module (see figure 3.23). The Redis database then transmits the data to other modules of the back end in Python. The front end, written in Javascript, manages the display, particularly the calculations related to three.js. Its movements copy live those of the user in front of the mirror [83]. This version was the first version of the interactive avatar on the augmented mirror.



Figure 3.24: The Educative Sign Language Avatar (ESLA).

An animated 3D character appears on top of the scenery at the launch of the application (see figure 3.24). Its animation is made possible by playing features using MediaPipe and OpenCV. The program can retrieve the coordinates of a user's position in real-time. These coordinates are processed, and the avatar copies the movements thanks to Animation Retargeting, a video game animation technique to map movements on 3D objects.

The program retrieves the links between each point of the pose provided by MediaPipe, and Three.js calculates and displays the 3D part.

Three.js entirely manages the loading of the character, its display, its animation, the rendering, the light, and the camera.

Three.js uses a pivot system with matrixes and quaternions. However, some functions to transfer rotations from one base to another still need to be included (even though they are very much in demand by the community). Therefore, the entire limb rotation system was implemented by hand. MediaPipe first retrieves the (real) coordinates of the user. Transformations convert these coordinates into the world base (x-axis to the right, y-axis to the top, z-axis to the camera) of the three.js scene. The program retrieves the coordinates of the top of the spine, the shoulder, and the elbow (see figure 3.25). It creates two vectors spine/shoulder and shoulder/elbow. The coordinates of these two vectors are then read from the base of the parent of the bone to be rotated

(here, the left shoulder). An algorithm then calculates a quaternion containing the rotation data from the first vector to the second.

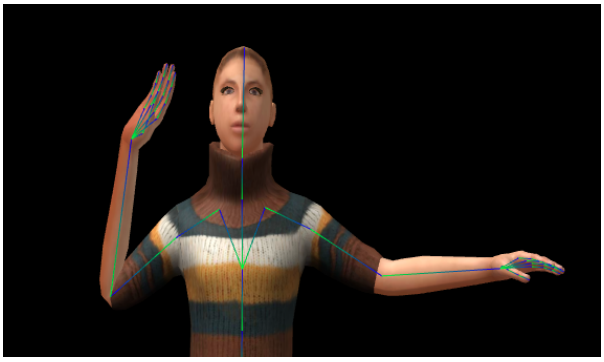


Figure 3.25: ESLA's squeletton.

Three.js then applies this quaternion to the limb, which rotates the same as the user's arm with his shoulder in its local base (see figure 3.26). This rotation allows precise rotation transfers from one base to another, thus animating all parts of the avatar.

Unfortunately, the project lacks much precision in the movements, especially at the level of the fingers, which is very inconvenient for creating accurate tutorials in sign language. Another more efficient solution for creating 3D animations with the estimated pose had to be implemented.

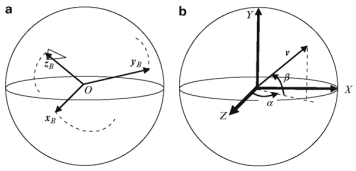


Figure 3.26: Spherical parameterization of rotations: (a) Movement of unit vectors attached to body axes during a rotation of the object. (b) Parametric representation of unit vectors on a sphere [84]

Kalidokit module implementation

The creation of an avatar controllable with Mediapipe on the mirror has been by the implementation of the Kalidokit solver. This JS library allows the animation of arms, hands, fingers, face, mouth, eyelids, and pupils of VRM models (Virtual Reality Models). The VRM is based on the standard 3D format glTF2.0 to manage humanoid models. It aims to be particularly expressive. This model has many articulations and can blink and animate its mouth. It is often used in VR games (VR chat) or by Vtubers (entertainment broadcasters who use a virtual avatar).

The models used in the game come from the site Vroid Hub cite[85]. They must respect specific conditions of

use: use by a third party, downloadable, use as an avatar, and commercial use by a company. The avatar must also be sober, as it should be a model taking part in a demonstration on the augmented mirror deployed at the De Vinci Innovation Center. The one selected in the demo application allowing to control the avatar in augmented reality is called "papa_de_him_chan".

The model is easily changeable locally in the application, allowing the user to manipulate different characters. This application makes creating sign language tutorial animations for the learning game possible.

The implementation of Kalidokit on the mirror has been reworked to suit the operation of GOSAI. The joints are set to greater freedom to allow more flexible animations (see figure 3.22). Such an application allows recording qualitative signs and poses for the characters and tutorials of the game.

The models used for the characters of Aria, the grandmother, and the salesman were recovered on Vroid Hub and are free to use. The VRM models are mainly adapted for Vtubers, explaining why the characters look childish and look like Japanese anime characters.



Figure 3.27: Aria's character performing the "I love you" sign.

All the avatars' sprites and 3D video animations were recorded from the "Aria" application on the mirror. These animations include 22 video animations and 14 sprites for the character of Aria in the sign language game (see figure 3.27).

3.5 Sign Language Recognition AI

3.5.1 Overview

A program using computer vision was developed and implemented in the project to enable SLR. This AI allows the recognition of choices in the game thanks to the estimation of the user's pose and the control of the game using commands linked to signs.

Sign language involves using the upper part of the body, such as hand gestures, facial expressions, lip-reading, head nodding, and body postures to disseminate information [27]. The model must be able to recognize these different elements.

Many models with sign language recognition already exist. The work done here proposes a model and an easy system of creation, dataset management, training, and visualization of the data.

3.5.2 Integration in GOSAI

The project of creating the SLR AI is apart from GOSAI [86]. GOSAI integrates only the weights in a module, making the comparison between a sign made in front of the camera and the values of the weights recorded locally. This architecture means that not all of SLR's AI code is integrated into gosai's code, and only lightweights can be accessed for the comparison.

The video game retrieves the user's coordinates and places them in tensors containing the data of a video of 30 frames. The tensors are then passed to the model, which returns a table of probabilities concerning each action. When the most probable sign is performed for one second, it is transmitted from the SLR module through GOSAI's Redis database to the game engine. The game engine considers a sign to be validated if it is received more than 5 times in a row. This prevents accidental signing.

3.5.3 Structure

The project contains six local modules and a main file that initializes the parameters and launches the different processes. The processes are an optional phase of dataset creation, a tutorial phase that displays the skeleton for data visualization, a preprocessing phase that retrieves the data from the dataset and formats them, a data augmentation module, and a weight calculation and exportation in different file types.

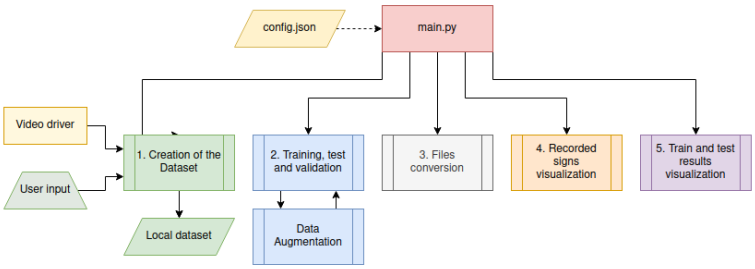


Figure 3.28: Structure of the AI for sign language recognition.

The developed AI is a full-fledged project that does not require a goal. It takes five optional steps (see figure 3.28). These steps are: the creation of a dataset, the training, test, and validation (with data augmentation or not), the conversion of the created files, the visualization of the recorded signs, and the visualization of the training and test results.

Dataset If a movement is included in the known actions but is not present in the files, the creation of the dataset of this movement launches automatically.

A user creates the database on-live during a dataset recording phase. The data is stored in a folder "MP data" created and separated into three folders: train, valid, and test containing the actions (name of the movement). The recorded sequences are automatically separated between the three folders train (80% of them), test (10%), and valid (10%). The program automatically completes the existing dataset by default according to the number of sequences entered in the configuration file.

Each recorded sign sequence lasts one second, or 30

frames by default. The frames include 58 point coordinates, (or 116 data per frame) after removing 431 points of the face and just Keeping 4. The four kept points are one for the chin, forehead, right, and left cheek. It enables signs using the head to be better detected without recording a large number of points that would not be useful for detection. There are 1600 sequences in all, for a total of 16 different signs, i.e. 100 sequences for each sign (see figure 3.29).

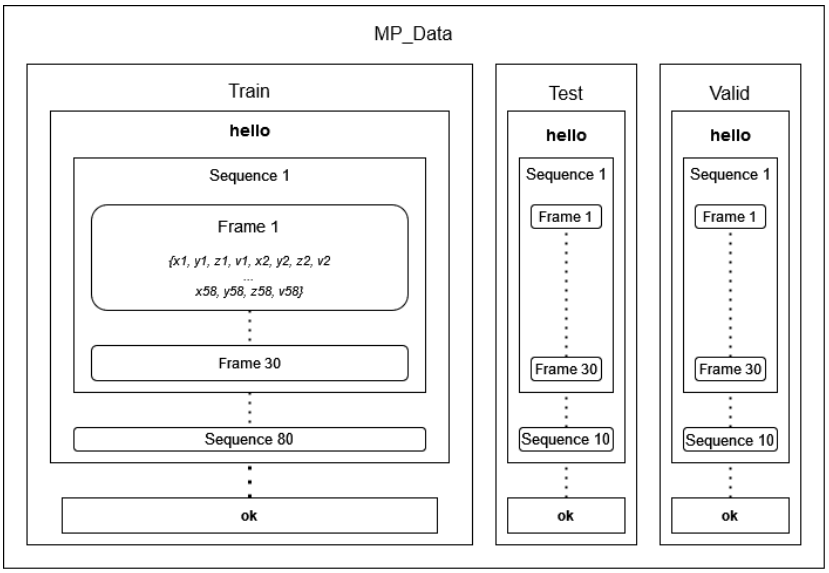


Figure 3.29: Sign language dataset structure.

The program retrieves the coordinates from Mediapipe during the creation of the dataset. If the Intel camera is undetected, the device's webcam record, an Intel real sense, or the default pc webcam. When implemented on another device, such as the interactive mirror, this driver must be modified to match the camera used.

At the start of the dataset creation, the user alternates one second of video recording, then two seconds to reposition in a loop for each sign. He has video feedback on himself to check if he is well-positioned in the plan.

The recorded signs visualization The recorded signs visualization module (data_vis) shows a visualization of every recorded sign (stored in the dataset). This process verifies that the signs have been made and are stored correctly on site. The program retrieves the coordinates of a person stored in the dataset (for each video), sorts

the coordinates of the points of each body part, and finds the different links between each point frame by frame.

The visualization displays all the links between the points stored in the dataset. The program also displays the action with which the visualization of the sign is associated above the drawn character.

Preprocess The preprocessing phase retrieves all the data from the dataset and places them in tensors to provide them to the model. After creating the preprocess instances (according to their type: train, test, valid), we pass them to the data loader.

These pass an index to them (concerning a sequence among all the sequences attributed to the type of the preprocess). Their program calculates which action is concerned by the desired sequence and recovers all the coordinates of the frames of this sequence. Some papers express that sign language movement is only possible by indicating with lips or facial expressions [35]. Therefore, the dataset should keep some lip coordinates integrated during the preprocessing. For some other papers [87], keeping facial and lips expressions is unnecessary to keep these points, which takes much storage for nothing. Here, the preprocess takes into consideration only four coordinates of the face. This selection allows for much faster train loops because all these points are unnecessary.

Data augmentation The preprocessing phase optionally calls the data augmentation just after retrieving the data of the concerned sequence. The functionality reviews the data and applies a random horizontal and vertical shift and scale to it. This process artificially increases the number of positions the user can stand.

Model The model comprises a bidirectional LTSM, a linear, a dropout, a batchnorm1d, a relu, and another

Linear. The model reaches an accuracy of 87% on the test set and 96% on the training (see figure 3.30).



Figure 3.30: Visualization of testing and training accuracy over 200 epochs. In yellow is the accuracy without data augmentation, and in green with data augmentation.

Recurrent Neural Networks (RNN) allow the processing of temporal sequences (language, videos, numerical data). They keep the memory of past data to predict data sequences shortly. LSTM (Long-Short-Term-Memory) and Bidirectional layers are RNNs used in the current model to consider the different coordinates and frames already passed for better performance. The LSTMs are composed of several gates (respectively 3 and 2), which allow one to forget or to selectively memorize the information of the previous temporal sequence in dynamic memory. This RNN is particularly adapted to multiple input dimensions. For example, here, the data is composed of sequences containing frames and coordinates (3 dimensions).

The training uses the optimizer AdamW, a stochastic gradient descent method based on adaptive estimation of first-order and second-order moments with an added method to decay weights [88].

The program exports the output models in pth format. Pth is a standard PyTorch convention to save models. They can also be exported to onnx (Open Neural Network Exchange) [89], an open format designed to represent

machine learning models.

Onnx extension enables greater interoperability in the AI tool community, allowing many AI frameworks to be more compatible by allowing them to share models. This interoperability allows users to deploy trained models in different software platforms like GOSAI quickly.

Here, the onnx models retrieved from the project output are directly implemented in the SLR module of GOSAI.

3.5.4 Configuration

All the processes mentioned above are optional. The user can enable or disable many project features in the config.json file.

All the options are: "actions" (all the signs recognized by the AI, "adapt_for_mirror" (to transform the coordinates specifically for the augmented mirror), "convert_files" (to convert the files from pth to onnx), "erase_dataset" (erasing the dataset before recording a new one), "erase_runs" (erase the local TensorBoard files), "make_data_augmentation", "make_dataset", "make_train", "make_tuto" (visualization of the recorded movements), "nb_epochs", "nb_sequences" (total video number the user wants to record for a sign), "sequence_length" (one video frame number).

The default actions are "nothing", "empty", "ok", "yes", "no", "left", "right", "house", "store", "hello", "goodbye", "television", "leave", "eat", "apple", and "peach". Those are all the signs used inside the SL video game.

The project aims to make it as easy as possible for the user to create weights according to his desired characteristics.

3.5.5 Visualization of the results

TensorBoard is a tool providing visualization solutions to machine learning tests. It allows for tracking and

visualizing metrics such as loss and correctness. TensorBoard allows displaying the exported model's graph, histograms of weights, biases or other Tensors, and many other functions.

Just by entering the "make set-up_tensorboard" function in the terminal (or by doing a "make first_boot"), the user launches the creation of a local server running on port 6006, retrieving the weights as the model trains.

A browser page then opens automatically to access the local server address. The user can study a graph of the training and test phase (percentage of success according to the number of epochs) on this page.

The logs of these results are stored locally. They are accessible on a webpage, and the user can easily compare them. This visualization makes experimentation easier.

3.5.6 Limitations

In the course of numerous experiments, several biases have emerged that directly impact the outcomes of AI. For instance, when there is a significant disparity in body size between the user and the individual who generated the dataset, it often poses a challenge, leading to variations in recognition performance among users with similar or distinct morphologies.

Furthermore, the results of the systematic literature review (SLR) are contingent upon the parameters of Mediapipe, as the AI relies on data obtained from pose estimation for training. The training outcomes indicate that Mediapipe exhibited superior performance in identifying a person's limbs when there was a strong contrast with the background. The selection of the background resulted in a notable performance discrepancy of approximately 15% between the training and test sets. Mediapipe demonstrated improved capability in detecting the contrast between white skin and a black shirt, and subsequently recognizing instances where a person created the dataset or conducted SLR while wearing

such attire.

This contrast renders the SLR more effective for detecting signs produced on the belly, which serves as a plain background devoid of intricate details, compared to signs created in other areas with more background complexity.

3.6 Usage scenario

One of the primary objectives of the augmented mirror is to facilitate interactive learning in augmented reality. It encompasses multiple applications that integrate entertainment and practice functionalities. Users can select and launch specific applications from the mirror's menu. In the context of the sign language game application, users engage with Aria's instructions to embark on an adventure. This game is particularly designed to foster practice and engagement among children.

The augmented mirror eliminates the need for additional equipment, although it requires a sufficient amount of space for optimal usage. Users should be able to stand two meters away from the device and freely execute various movements without any disturbances. It is important to note that the mirror is not suitable for use in dimly lit environments. Adequate lighting conditions are necessary for the camera to accurately detect the user's movements.

3.7 User study

3.7.1 Set Up

As the sign language learning game primarily targets young users, this study focuses on a sample of 6 individuals aged between 18 and 25 years old.

Six participants train with a sign language practice program available on the "Pocket sign" application. The application offers effective training by watching sign

language tutorials and then taking quizzes asking a user to choose the right sign for a word. Learning takes place through simple viewing and repetition. Participants practice 10 signs with the application ("same", "different", "forget", "remember", "orange", "key", "mouse", "box", "spray", "glasses"). They are then asked to perform 4 of these signs and are evaluated on them.

Then, the same participants interact with the sign language learning game using the mirror interface. Before using the game, they receive an explanation of its principles. Two of these participants navigated through the house scenario, while the remaining four navigated through the store scenario during the game. Navigating through the house displays the signs: "ok", "yes", "left", "right", "house", "store", "hello", "goodbye", "television", "leave", "eat" and the store path displays the signs: "ok", "yes", "no", "left", "right", "house", "store", "goodbye", "apple", "peach". The participants are then asked to perform 4 of these signs and are evaluated on them.

Both groups are informed from the outset that they will be assessed on specific signs at the conclusion of the study. The duration of each session, from the beginning to the end, is recorded for both groups.

In conclusion, users were asked to provide their opinions on various aspects of the game using a rating scale of 10.

The survey measures:

- The average motivation to learn sign language through the augmented mirror and video game approach rather than with "Pocket sign".
- The level of fun in interacting rather than with "Pocket sign".
- Ease of understanding the tasks.
- The convenience of the augmented mirror as a learning method.
- The overall effectiveness in learning with the game.

3.7.2 Results

Pocket Sign Application The average time required to complete the Pocket Sign application was 6 minutes and 28 seconds. The average score achieved by participants was 3.54 out of 4 signs learned. The application is interesting for learning to recognize sign language gestures, as this is what the application trains the user to do. Most of the errors made by the participants are linked to a problem of gesture in certain complex movements (e.g. "box" or "remember"). Users appreciated the fact that the tutorial was given by a human, which enabled them to see things clearly. On the other hand, they criticized the speed at which the signs were executed, which made them more difficult to reproduce.

Sign Language Video Game Certain users encountered difficulties associated with inadequate AI recognition. For instance, one user required 8 minutes and 25 seconds to complete the game, significantly longer than the average completion time of 4 minutes and 44 seconds. This discrepancy can be attributed to the substantial height difference between the dataset creator (1.84m) and the user (1.60m), thereby introducing biases into the system.

To assess the users' performance, they were evaluated on a random selection of signs following their gaming session. On average, participants achieved a score of 2.33 out of 4 signs learned. The errors observed can be attributed to various factors. In some instances, the AI prematurely identified a sign before the user had accurately positioned their hands, leading to discrepancies between the expected and detected gestures.

Subsequently, users required additional time to observe the signs and make their choices. Another error stemmed from the presence of three simultaneous tutorials (menus with three choices), which hindered users' ability to concentrate on each tutorial, resulting in less efficient sign memorization.

The results from the survey are available on the figure

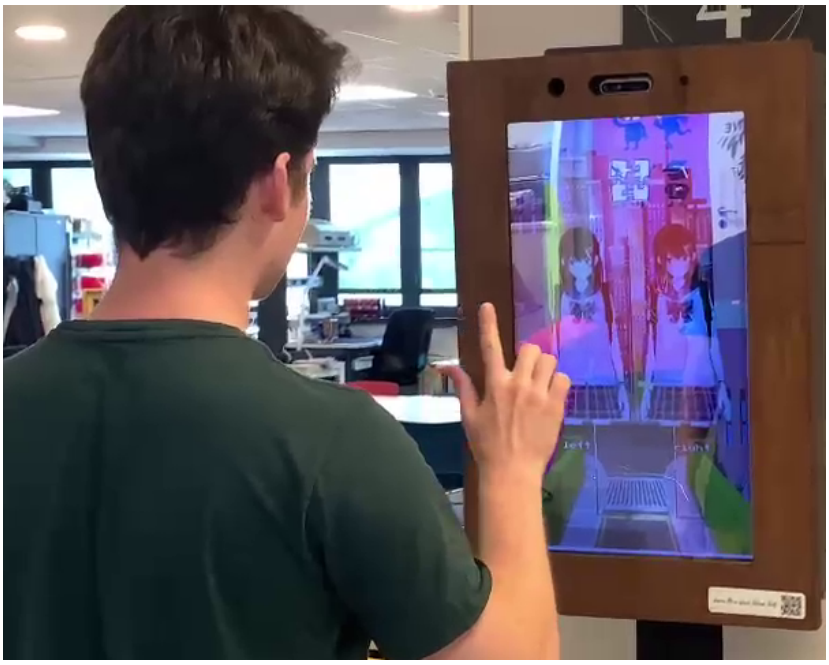


Figure 3.31: Sign Language Video Game user study.

3.32.

Overall feedback indicated that the avatars' excessive finger movements compromised precision. The use of a 2D display on the mirror screen, instead of 3D (see figure 3.31), posed challenges in understanding the signs. Some users were deterred by the anime-like appearance of the characters, which dampened their enthusiasm for practicing.

A few users forgot the "ok" sign to skip dialogues during the game, despite it being shown twice at the beginning. This forgetfulness hindered their progress and required external intervention for sign reminders. One user faced learning difficulties due to perceived ambiguity in the avatar's movements. However, interacting with a real person facilitated more effective learning. Users strongly criticized the AI's lack of accuracy, rating it at 6.11/10. Addressing the issue of AI detection speed can also be achieved by increasing the detection threshold value, offering an more effective solution.

The 3D animated tutorials garnered positive feedback and were considered more engaging than simple video tutorials. Users expressed satisfaction with using the game on an augmented mirror, deeming the mirror's

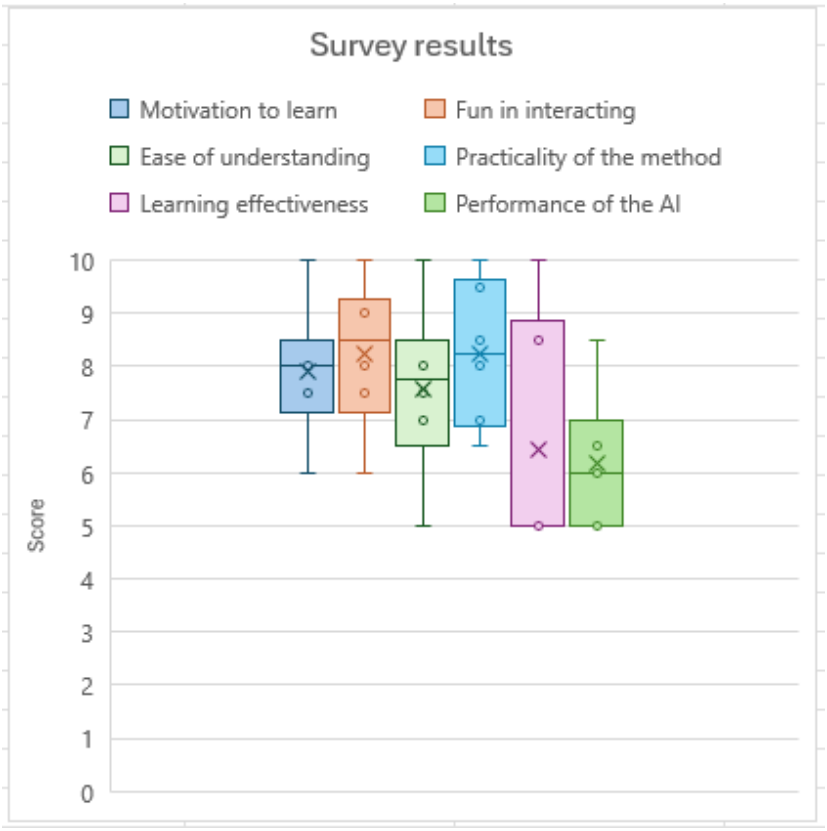


Figure 3.32: Results of the survey.

interface mainstream and easily integrable into a living room setting.

3.8 Discussions

Participants reported slight demotivation after using the app and practicing 10 signs. The application is effective, but the learning curve is intense. Some signs reappear many times in the app's questionnaire. It becomes laborious to validate them several times, and the learning curve can become long. Even so, this method guarantees effective information retention.

The advantage of the sign language learning game is the marked improvement in a player's motivation to learn and practice SL. The Sign Language Game effectively stimulates visual, verbal, and kinesthetic intelligence among its users. Visual intelligence encompasses the ability to mentally construct images and accurately perceive the three-dimensional visual world. By engaging with 3D sign language tutorials and deciphering the

diverse movements and limb positions of the avatars, users' visual intelligence is stimulated. They must perceive and mentally project the identified signs before reproducing them.

Verbal intelligence pertains to rapid reading, speaking, and the capacity to articulate mental concepts into words. Connecting visual images to corresponding words within the project enhances users' fluency in expressing themselves and linking mental ideas to gestures.

Kinesthetic intelligence involves the refined utilization of one's body, expressing oneself through movement, and precise control over bodily motions. The project's gamified approach stimulates this form of intelligence as users must translate mental images of body positions and movements into real-world actions.

The simultaneous stimulation of these three intelligences, coupled with the motivational aspects of gamification, enhances information retention. Users retain sign language gestures more effectively through the combination of visual observation, mental projection, and physical execution in space. Studies have demonstrated the influential impact of non-verbal behaviors, such as gestures, on comprehension and memory in communication [90].

Limitations The findings from the user studies highlight several limitations within the project.

One limitation pertains to the size discrepancy between the individual who created the dataset and potential users with varying heights. While the accuracy is high for the dataset creator, it diminishes for users with different morphologies. To address this issue, one potential solution is to introduce random compression of coordinates along the horizontal and vertical axes during data augmentation or to include individuals of diverse sizes during dataset registration.

The dataset utilized in the project is relatively small,

exhibiting limited noise, slight variations, and minimal mixing. Although the model demonstrates moderate efficiency, even when trained on multiple distinct signs, its accuracy would further decline with dataset expansion. Therefore, the model should be reworked in the future. The implementation of transformers could prove beneficial, enhancing performance. Transformers have proven successful in various domains such as image processing, biological sequence analysis, and video understanding.

Another area for improvement concerns the clarity of avatars in sign language tutorials, which can sometimes be challenging. The pivot angles of the avatar's bones are constrained to prevent glitches and ragdoll-like movements. Consequently, capturing clear signs, particularly at the finger level, can be difficult. As a result, users must imitate and learn from an inaccurate model. To address this issue, slight modifications can be made to widen the allowable angles of the avatar's bones, thereby enhancing precision.

Learning is also not as efficient as desired despite the increase in motivation to learn with this method. Learning with the application resulted in a score of 3.54 signs learnt out of 4, for a score of 2.33 with the video game. Memory stimulation exercises would have to be added in a future work. Certain limitations therefore slow down the experience such as the lack of diversity of the dataset, the lack of precision of the 3D tutorials, and the lack of memory stimulation.

3.9 Conclusion

3.9.1 Playful Learning and ASL

Playful learning and gamification are pedagogical approaches that leverage elements of play to enhance the acquisition of knowledge. These approaches are frequently employed in educational contexts, including the use of video games, to facilitate the teaching of practical skills such as sign language. By integrating augmented

reality, pose estimation, and artificial intelligence, the Sign Language Video Game offers an immersive and interactive learning experience, enabling learners to practice reproducing signs accurately in real-time.

This approach facilitates learning through immediate feedback, allowing learners to engage in repeated practice sessions at their own pace. Furthermore, the incorporation of cutting-edge technologies in the game design tends to generate curiosity and captivate learners, thereby fostering increased motivation and perseverance in the acquisition of sign language skills.

3.9.2 Future Works

The current game is suitable for direct practice. However, the learning system is too simple because users can afford to copy the sign without actually learning it. It would be necessary to add an evaluation dimension with signs that come up in the potential choices without giving the tutorials every time. Thus the user will have to force himself to remember and learn better by repetition.

Current work This dimension of training is already present in another mirror-based application, resembling sign language training. The application comprises consecutive viewings of sign language tutorial videos, practice sessions, and a correction phase. During the correction phase, frozen Mediapipe bones and points, forming a skeletal structure, manifest on the mirror and perform the initial stage of a sign.

The user is provided with a Mediapipe skeleton displayed on the mirror, which tracks their movements. When the user overlays their skeleton onto the model's skeleton, the model's skeleton undergoes slight adjustments to present the subsequent frame of the recorded sign movement. Subsequently, the user must align their skeleton with the second frame until they validate the 30 frames constituting a sign. Specifically, the person needs to align certain points of their body, particularly their

hands and fingers. These corrections facilitate precise learning of the executed sign. The training then proceeds to a tutorial of another sign.

Implementing the correction phase of this training within the sign language game would yield significant benefits. The corrections enhance the user's sign retention by requiring them to repeatedly perform the signs while focusing on the positioning of each finger. The clear demonstrations of the signs enable the user to visualize the accurate spatial positioning they need to adopt. Consequently, this implementation would address some of the current issues in the game. Although the correction phase is presently excluded due to its disruption of the game's flow, integrating it at the game's conclusion as a reminder of the learned signs would serve as an effective learning method.

Music learning applications in AR

4

4.1 Introduction

The DVIC augmented mirror overlays various digital information on top of the natural reflection of the user's head and upper body. Such overlaid information can be useful for the learning of embodied skills, and applications have already been created for the mirror to help people learn sign language and dance.

Currently the most common way to learn singing is through:

- Recording and listening to one's own performances.
- Taking a good posture, which is important to learn to sing in tune.
- Paying attention to one's breathing, which is the basis of vocal technique and can be improved through breathing exercises.
- Performing vocalization exercises, which require a lot of patience.
- A capella singing, practised by listening to and learning a piece many times before reinterpreting it.

This project explores how the mirror can help people learn music, and in particular, singing. Seeing one's reflection is useful when practicing singing because it allows the monitoring of one's posture. The mirror adds the ability to monitor one's pitch accuracy. This is done by adding a frequency tracking module that estimates notes sung by the user and an interface that gives the user real-time visual feedback. The user can then train his or her ability to sing in tune by comparing the target frequency with that actually sung. This ability is called musical imagery, the ability to "imaginate" a note before playing/singing it [91]. It can be perfected with practice. This ability is very important for a musician as

it allows him/her to visualise a gap between two notes he/she wants to play/sing, and to translate it on his/her instrument or by voice.

Practicing music can be very laborious, long, and difficult. For example, practicing singing requires attending singing classes that teach you to visualize the right note in your head before you sing it. Singing also requires good posture and breath control to be effective. The interactive mirror has the potential to give users the ability to monitor their posture, an important element in singing. The user can get double feedback by singing in front of the augmented mirror. He can see the position of the notes in relation to the ones he is singing and can correct himself. At the same time, he sees his reflection in the mirror and can correct his posture to sing more effectively. In order to implement the singing application, a frequency tracking module estimates the notes sung by the user, and the user interface gives feedback on the accuracy of his pitch. The singer can also activate a tutorial with falling notes, activate the playback of music to learn it before singing it, and use a calibration to set the singable range of notes for him.

4.2 Related work

4.2.1 Singing Learning Applications

The development of new electronic musical interfaces gives rise to potential novel approaches in music education. The gestual interactive systems and interfaces have great potential for practice in music pedagogy [92]. A number of smartphone apps exist for people learning to sing that focus on giving feedback on correct intonation. Yousician is a fairly comprehensive application that allows you to quickly learn and master piano, guitar, bass, singing, and other instruments by practicing on your real instrument. It offers to learn music theory through special exercises and lessons. The application listens to the user singing/playing, evaluates his/her performance with a frequency estimation, and proposes

to play different songs with tutorials and games.

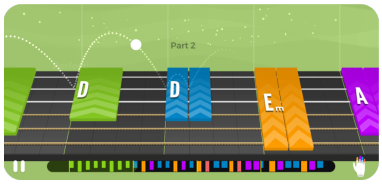


Figure 4.1: Yousician interface.

It offers a system of notes in the form of bars coming from the right, which the user must sing afterwards (see figure 4.1). If he sings right, the bar turns green, if not, red.

Other applications are very similar such as Sing Sharp, LearnSing... Riyaz is another application that allows you to load existing music to generate the notes of the song, and then view a tutorial of these falling notes. The application then gives statistics on the performance: pitch accuracy, control and breathing capacity, vocal range. These applications often start by determining a range of frequencies that the person can perform with the voice before starting the singing training. This allows the tutorial to be tailored to the user's vocal abilities.

The general strategy of these applications is to use learning methods by practice, measurement, and correction. The classic operation is the realization of tutorials on the placement of the hands for an instrument, or the position of the notes to be sung for the note, then by a graphic feedback with the visualization of the fact that a note has been correctly sung or not. It is possible to learn how to play music by means of video games created for this purpose. Among the best known video games in this field are Guitar Hero and Rock Band. These two games allow you to learn to play the guitar in a fun way, always using a system of falling notes [92].

4.2.2 Music Practise

Music practise assisted by AR Augmented reality, which involves the integration of virtual objects into the real world and enabling interactivity through sensor technology, presents a novel paradigm for practical applications.



Figure 4.2: Teacher and student using the system during a music class [93].

Frederic Bevilacqua & al. proposed a project with miniature wireless sensor system that is used with accelerometers and gyroscopes to make gesture recognition in musical pedagocial scenarios (see figure 4.2). Music

teaching methods are often based on playful approaches and exercises focusing on body movements. The project aimed to experience and practice "smoothness" and "fluidity" of musical gestures by analizing the curves of the movements [93].

Yu Ting Hang et al. proposed a visual interface for learning singing on the Internet. The aim is to link pitch and musical notation through frequency estimation and waveform recognition. The visualization of the sung notes in relation to the desired ones allows to correct vocal accuracy errors and thus to practice singing by sight [94] (see figure 4.3).

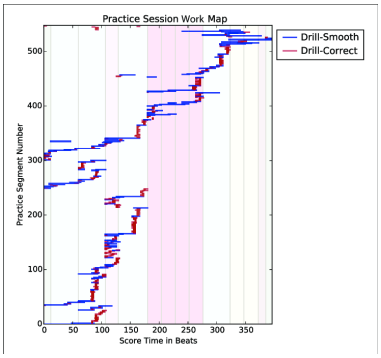


Figure 4.3: Practice Session Work Map of an expert player (E-1)'s hour-long practice of Chopin's "Mazurka in A minor, Op. 17 No. 4".

In this context of music assisted with AR,Giovanni Santini [95] delve into sound processing and synthesis facilitated by sensors, thereby unveiling innovative avenues for musical experimentation. Their work sheds light on the promising prospects that augmented reality offers in the augmentation of musical instruments, emphasizing the pivotal role of interface design in shaping the musical experience. The article highlights several notable projects at the intersection of augmented reality (AR) and musical instrument augmentation pertinent to our research endeavor.

ChromaChord [96] is another example standing out as an initiative enabling users to engage with virtual interfaces through motion capture devices like Leap Motion and Oculus Rift (see figure 4.4). In this setup, virtual objects produce sound upon interaction with the user's hands. The users can play music by interacting with virtual strings in a three-dimensional environment. ChromaChord has potential applications in music education, composition, live performance, art creation, and other areas of music and VR.



1. Performer operating *ChromaChord*. Oculus Rift and

Figure 4.4: Performer operating ChromaChord. Oculus Rift and Leap Motion.

VRMin [97] is a virtual reality application that employs principles akin to AR interfaces to enhance the Theremin, thereby elevating gaming precision through virtual spatial representations 4.5. The project combines sensor input from the physical theremin with visual cues and feedback in a virtual environment to provide an en-

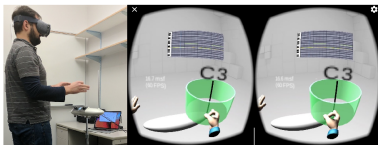


Figure 4.5: A student practicing the theremin using VRMin with capture of the view on the right.

hanced learning experience.

Pianolens is an augmented reality system offering a dynamic and interactive sheet music display interface. The program displays intricate musical elements through intuitive adaptations to the sheet music and imparts comprehensive guidance on mastering new compositions, including precise timing and fingering techniques.

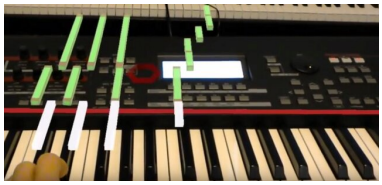


Figure 4.6: PianoLens demonstration.

The approach involves the utilization of colored blocks in an interactive musical setting. These blocks possess widths commensurate with the corresponding piano keys and lengths proportional to the note duration (see figure 4.6). They are set in motion towards the performer and intersect with a specific reference line, denoting the "current" moment. At this juncture, the pianist is required to depress the relevant piano key and sustain the keypress until the block's conclusion. Additionally, these systems offer real-time feedback regarding the accuracy of note execution and rhythmic precision.

The features accelerate the acquisition of novel musical pieces but also elevate learner engagement during the rehearsal phase, standing as an effective tool in the field of music education and performance.

These illustrative examples underscore the diversity of projects that unite music, AR, and instrumental augmentation, paving the way for fresh opportunities in interaction and musical expression. Each of these projects tackles the design of AR interfaces in distinctive ways, all aimed at enhancing musical practice.

Music practice assisted by frequency estimation Kin Wah Edward Lin et al. presented a self-learning tool for singing on the smartphone. It allows the user to improve timing and intonation in a similar way to the previous project. It uses an intonation level classifier, a user performance rating mechanism, and a pitch training tool.

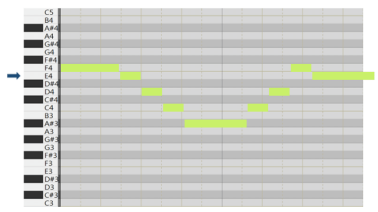


Figure 4.7: Intonation Level Classifier with Song - Fly Me to the Moon [98].

The mix of visual demonstration/feedback is natural to understand and useful (see figure 4.7). The user's

intonation score improved by an average of 94.81 % after training.

Philip McLeod [99] presents a program called "Tartini" developed by Philip McLeod of the University of Otago in New Zealand.Tartini is a program intended to provide real-time music analysis tools to musicians. It uses a microphone and computer to provide precise visual feedback on elements such as intonation, intensity and shape of vibrato.The project aims to help musicians improve their musical technique, focusing mainly on pitch-related aspects of music.

4.3 Singing Learning Program in AR

4.3.1 Context

The DVIC augmented mirror project aims to change the way we perform certain tasks and disciplines in everyday life. It adds a dimension of interactivity through our reflection, our position, our movements.

The mirror's 3D camera makes it possible to precisely estimate the position of the user in space. The person can then interact remotely, without touching the screen. A Plexiglas screen positioned above the screen of the device reflects the reflection of the user. This allows the display of elements over a reflection and thus interact with elements in augmented reality. The mirror's augmentation gives people more awareness of their position and their movements. It allows full-body interactions with digital information, rather than accessing it through the small screens of computers and mobile devices. The added value of the project is to exploit mainly kinesthesia (which is not possible with other more traditional devices) as well as other forms of interaction. This is in addition to the various ways of interacting with the human body (human inputs/outputs) that can be exploited by the electronic devices we use every day.

Applications allow for example to have a return of its position with the display of a mediapipe skeleton above

its reflection, to learn elements of sign language thanks to a training module, as well as to test other games. One of the primary uses of the mirror is to enable much better learning, using as many forms of intelligence as possible.

4.3.2 Overview

The objective of the project is to be an example of a module allowing people to learn the practice of a musical instrument on the mirror. As the device only requires the person in front of it without any material, a musical module would not have been suitable for most musical instruments.



Figure 4.8: A user testing the musing training application.

The project is based on frequency estimation for practice. A user plays an instrument or sings in front of the mirror (see figure 4.8). The mirror collects the frequency emitted and displays it on the screen.

The user sees the frequency he or she is emitting represented by a dot on the screen with a particle effect. The lower the frequency, the further to the left the dot

is, and vice versa. Some options can be launched from the sub-menu of the "music_training" application already implemented on the mirror. The user can launch different music tutorials.

For example, he can launch the tutorial "La vie en rose", notes fall from the top of the screen and he must play them in rhythm. The module no longer has any aspect related to movement or position but only to the frequency emitted by the instrument or by the voice. When the user sings correctly, the bars falling from the top change colour, gradually turning green and the particle effect on the dot intensifies (see figure 4.9). They become increasingly red when the user sings out of tune. The score achieved is displayed for a few seconds at the end of the song.

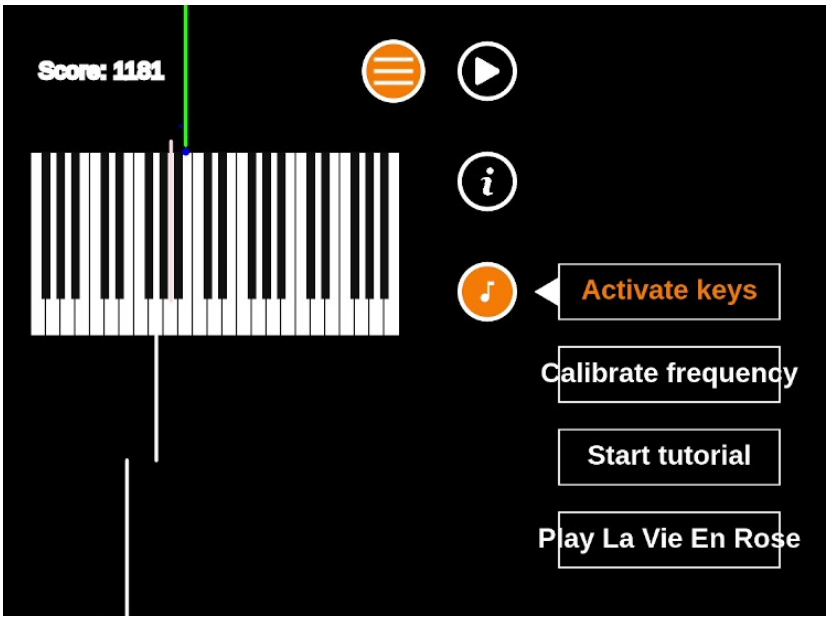


Figure 4.9: The interface of the music learning application.

The vision of the project is to train the user to more or less consciously link the manipulation of the instrument or his voice with the sound he visualises in his head. The visual present on the mirror allows him to know where he is, if he has played/sung the right note and how to correct it if not. This approach is different because it offers different options in the display. The user can launch different music tutorials, display bars to locate notes, erase them, as well as display the position of his hands. Practising on a mirror is a plus as the user

practices standing up unlike with a telephone. This makes it easier to practice singing because the body remains upright. This project is intended to be suitable for multiple instruments, while being entertaining and instructive.

4.3.3 General Architecture

The augmented mirror system runs on an operating system software called GOSAI. The music practice module takes the form of a JavaScript application coded directly on it. It uses the P5 library to display 2D elements and create the sketch. The backend uses two drivers coded on GOSAI: a new driver allowing the OS to use the microphone of the device and another driver allowing to make frequency estimation listening to the microphone. The latter takes a sound wave generated as an input and returns a musical frequency.

The GOSAI microphone driver takes the audio stream from the device hardware. It passes the data to the "frequency_estimation" module which detects the frequency emitted in the audio stream as well as the gain and other parameters. This module places this data in the Reddis data. The processing.py file retrieves the frequency estimation data from the data and sends it to the display.js with socket.io. The display initializes a new music workout and sends it the necessary data to run.

New various tools such as the sub-menu system are coded on GOSAI for the project. The sub-menu allows the applications to propose options to be activated or deactivated manually by the user thanks to the estimation pose according to his desires.

4.3.4 User Tests

The user tests in this project aim to assess a user's facilitation of learning to sing with the mirror compared to other conventional tools.

Conduct of the user learning test We evaluate the difference between classical learning through "a capella" practice and learning with dynamic correction on the mirror.

A randomly chosen user listens to the first few seconds of "La vie en rose - Carolina Eyck - theremin" [100] one time. This cover played on the theremin was chosen in particular because its sound is very precise and the accompaniment is very discreet, which promotes the user's understanding. The user must then sing it in playback while listening to it in his headphones . We record his singing performance to evaluate it later.

In a second part, he tests the application on the mirror. He then follows the 10 seconds tutorial with the falling notes. We record his singing performance to evaluate it later.

We realize these experiments in this order with 2 users each. We also conduct this experiment by reverting the order of experiments (first the program with music, and then the singing without visual support) with 2 other users.

Evaluation of the user experience Here are the areas evaluated and the questions asked to the program testers. Their opinion is measured with a score out of 10 points. Evaluation of the motivation to learn with this type of interaction: -Average motivation to learn this particular way. -Average fun in interacting. Evaluation of the comprehensibility of the module: -Ease of understanding what needs to be done Evaluation of the estimation/display system: -Usability of the overall system, responsiveness, accuracy, smoothness of display. -Practicality of the design, position of the notes, graphics chosen for the different elements. The rendering estimated and displayed in the application should be smooth and not display any other spurious frequencies measured by the microphone. Problems of this kind can lead to the display of unsung frequencies as well as "skipping" of the cursor displaying the sung note.

Evaluation of the tutorial method with falling notes. -Relevance and effectiveness of the tutorial method with falling notes rather than another. Evaluation of the use of the reflection. -Does the use of the reflection in the glass seem to be a good potential for improving one's vocal practice. Finally, the subject is asked about their remarks, suggestions and other constructive comments about the project.

4.3.5 Results

After testing four users singing the same music in playback, we observed a total number of errors of 6 imprecise semitones out of the 20 notes sung. We observe an inconsistent rhythm with successive accelerations and slowdowns. From the middle of the piece (containing the most complicated parts to reproduce), the number of errors increases considerably.

Users singing with the program achieved a score of 5 inaccurate semitones out of the 20 notes of the song. We observe a much better consistency of rhythm with little shift for each note. The user generally corrects the slight shift in the rhythm from one note to the following. During the two tests, we observed most errors during the large grade intervals (E3 to C3, then C3 to B3).

Users of the program gave an average rating of 6/10 for fun with the device, a rating of 8/10 for ease of understanding what needs to be done, a rating of 7/10 for accuracy and responsiveness of the program, a score of 7/10 in the practicality of the user interface and a score of 9/10 in the relevance of a falling notes tutorial for singing.

4.3.6 Discussion

Two of the four people interviewed had a beginner level in singing, and two had a more advanced level. During the first tests without the program, the novice users interviewed could not repeat the piece they listened to orally. They could not redo the correct frequency intervals, and

a third person could not sing another specific one after listening. His results were not included in this analysis, but it is essential to consider this example.

After tests by two other users with a more advanced level of music, on singing the same music in playback, we observed a total number of errors of 6 imprecise semitones on the 20 notes sung on average. We also observe a somewhat inconsistent rhythm with successive accelerations and slowdowns. From the middle of the piece (containing the most complicated patterns to reproduce), the number of errors increases considerably.

On the other hand, the two beginners could use the program on the mirror, which allowed them to sing approximately but with more success than without support. Their results were around twenty missed semitones (an approximate figure due to significant gaps and missing notes). Users with an advanced level of singing with the program achieved a score of 5 inaccurate semitones on the 20 notes of the song. We observe a much better consistency of rhythm with little shift for each note. Small rhythm shifts are generally corrected instantly. We observed most errors during both tests during the large grade intervals (E3 to C3, then C3 to B3).

4.3.7 Future Work

The current project does not currently allow feedback on the user's position. He can consult the mirror's reflection if he stands correctly, but the program does not assist him at this level. It would be interesting to offer feedback on the user's position. This feedback will help him stand correctly and sing better.

The project should use the pose estimation model to extract features and then analyze them with a simple algorithm. The program could capture a reference pose of the user when they stand correctly and then compare it to their posture when they sing. The program will then display notifications containing notifications to help restore its correct position in real-time. The applica-

tion could similarly communicate more general singing tips.

This system would provide feedback on the user's position and helps them sing better.

4.4 Theremin Learning Program in AR

4.4.1 Context

The theremin is a unique electronic musical instrument known for its ethereal sound and absence of physical touch.

The classical method of learning the theremin traditionally involves a structured and hands-on approach to understanding this instrument. Given its non-traditional interface, where pitch and volume are controlled by the proximity of the player's hands to two antennae, novice players often begin with fundamental exercises to develop precise hand control and sensitivity to the electromagnetic fields. These exercises typically include practicing scales, arpeggios, and simple melodies, with an emphasis on achieving accurate pitch and volume modulation. Additionally, students may study the intricacies of body movements and gestures required for nuanced expression in theremin performance.

Mastering the theremin poses a considerable challenge due to its lack of tangible keys or strings. In this context, the integration of augmented reality (AR) into theremin learning emerges as a promising avenue. This research explores the development and effectiveness of a tutorial program utilizing falling notes displayed on a mirror through AR technology. By leveraging the immersive capabilities of AR, this study aims to enhance the learning experience, providing a novel approach to bridge the gap between aspiring theremin players and the complexities of this unconventional instrument.

4.4.2 Overview

The theremin learning application is one of the augmented mirror applications. It consists of emulating a theremin functioning with a tutorial of notes similar to the project to help learn singing.



Figure 4.10: A user testing the theremin learning application.

The user can control a virtual theremin on the mirror, their right hand controlling the pitch horizontally and their left hand the volume vertically. Raising the left hand raises the volume, and lowering it decreases it. Similarly, moving the right hand to the right increases the pitch, and moving it to the left decreases it. To launch a note tutorial, people can choose different "scores" in the submenu. When choosing one of them, colored "lines" representing notes come down from the top of the screen. The user must then position their right hand in space so that the point representing the position of their hand corresponds to the position of the colored line. When the dot is aligned with the color line, the line turns green. The user must hold their hand in this position until the color line disappears. When the color line disappears, the user must reposition it on the following line. When the user successfully aligns the point on their right hand with a colored line, their score increases 4.11.

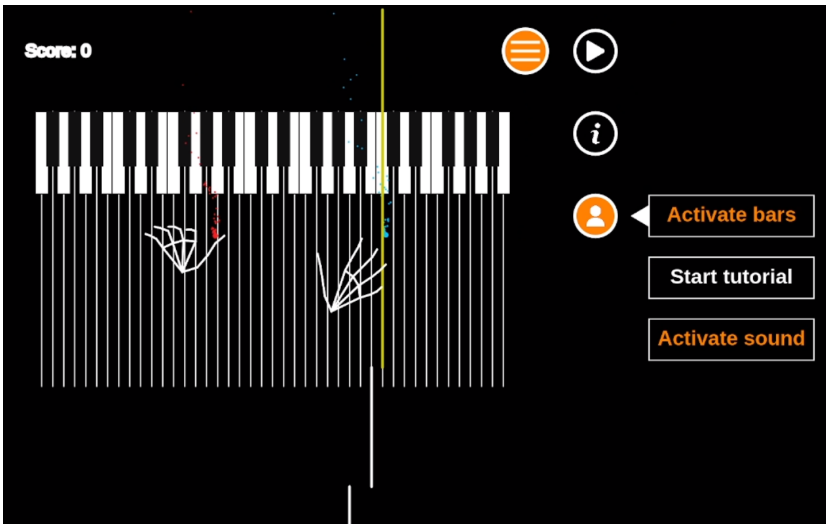


Figure 4.11: The interface of theremin training application.

The point representing the right hand constantly emits particles. These move with acceleration and velocity to mimic an attraction to the ground. Their volume increases and the lines color change when the right hand point is correctly positioned on the falling line. The line turns red when green or red if the user's hand is in the correct or incorrect position.

The screen displays a piano keyboard and white lines indicating its continuity. It allows the user to place his right hand at the desired height. The user can also hide the keyboard and guide bars using a button in the submenu.

4.4.3 General Architecture

The theremin practise module functions similarly to the singing module. The structure is a JavaScript application using the P5 library. The application uses the GOSAI pose estimation driver, subscribing to events providing left and right-hand points provided by Mediapipe.

The program calculated the hand movement values in space one meter from the mirror to match a user's real spatial hand movement values using an actual theremin. For example, moving an note or volume gap with the app means making the same movements as with a theremin. The program maps a real to virtual space regarding notes F2 to B5 (notes before and after this range are not

playable). The mirror calculates and displays a projection of two points of the hands on the mirror following the position of the user's hands.

The application calculates the exact positions of the two points from the highest point of the left hand and the rightmost point of the right hand.

User Tests

The user tests in this project aim to assess a user's facilitation of learning to play the theremin with the mirror compared to other conventional tools.

Evaluation of the user learning The tests are carried out on 4 people who try to play the theremin without support, and 4 others who try with the notes tutorial on the mirror as support. The basics of theremin operation are explained to each user. They can test it as much as they like beforehand. We give them tips and techniques to help them play correctly. We show and listen to the first few seconds of "La vie en rose - Carolina Eyck - theremin" (see figure 4.12). The subject is asked to try to play on the theremin exactly what he or she has just heard, with the right frequencies. The entire sequence of the user study is explained to the user beforehand.



Figure 4.12: Carolina Eyck playing "La vie en rose" with a theremin.

Conduct of the tests without the program Four users take turns. Each has six tries and can listen to the music in between. The last performance is recorded by capturing the mirror screen. We don't take rhythm into account, but concentrate on the precision of note deviations. We therefore evaluate the user's ability to play valid note deviations without visual support. An audio recording of the performance is made for later evaluation.

Conduct of the tests with the program Four other users take turns. Each user plays the theremin in front of the program, with the note tutorial correctly calibrated. We record his last performance by capturing a

video of the mirror screen. We don't take rhythm into account, concentrating instead on the accuracy of note deviations.

We then carry out an evaluation of the user experience similar to that of the singing learning application.

4.4.4 Results

Test without program The total error is calculated by comparing the frequency difference between each note played and the previous one. We compare each gap between the notes played to the expected intervals in the essential pieces. We consider previous offset errors, i.e., the user may have shifted the frequency by making errors. The sum of all missing or excess semitones in each interval played gives the total error.

We evaluate five people with several training sessions limited to 6 trials. We retrieved the results of the last test. We get an average of 45 semitones offset and many unplayed notes. That is an average of 2.25 semitones of shift per interval. There were many overall frequency shifts in the song (shifts not subsequently corrected).

Test with program We calculate the total error by counting the offset in semitones between each note played by the user and the program's objective note. The sum of all offsets gives the total error. After three trials using the program's tutorial with the theremin, we measured an average offset of 19.66 semitones on the fourth trial. That's an average of 0.98 semitones difference for each note. Those tested using the program rated their level of fun, giving an average of 5.33/10 after being asked about their experience using it. They gave an average score of 9.33/10 to the understanding of what needs to be done, a score of 5.66/10 to the fluidity of calculations and display, a score of 8/10 to the practicality of the interface and display, and a score of 9.33/10 for the relevance of the tutorial method with the falling scores.

4.4.5 Discussion

We note that the control group (without the program) made approximately twice as many errors as the group using the program despite twice as much training. The program also allows you to avoid global frequency shifts of the songs. We observed a clear progression in user performance ed during the tests. Using the application allows you to learn to play the theremin more quickly and gives a much better visualization of the notes played. The application also allows you to understand better the errors made and correct them more easily.

The most critical problem with this method is the very high accuracy of the theremin, even though the pitch was set to a minimum. It is challenging to play precise notes with the theremin used. The gap between each note is small, the keyboard displayed is very tight, and it isn't easy to place your hand correctly. Lack of precision harms the quality of learning. The use of tutorials from a notes tutorial with augmented reality was considered relevant and effective.

4.4.6 Future work

An interesting capability of the Theremin practice aid app is its ability to turn off its sound to let users practice their theremin in front of the mirror. The application detects the position of the user's hands at the same time as the user plays their instrument. This feature allows him to visualize the note intervals much better, so he can move his hands and position himself more precisely.

A typical theremin (such as the Moog Music Etherwave Standard used in this project) has a dial for adjusting the accuracy of the detected pitch. This setting set to a minimum allows one to play precise notes, but the difference in physical distance between two notes remains small. Therefore, the keyboard displayed on the screen must be perfectly adapted to match the distance between the notes. Several parameters must be taken

into account, such as the mirror/theremin distance, the pitch detection accuracy of the theremin, and the hand detection ability of the mirror camera being slightly tilted downwards. A missing feature of the project is the possibility of configuring the size of the displayed keyboard to perfectly adapt to the actual dimensions of the range of notes playable by the theremin according to its placement. We can think about an automatic calibration using the GOSAI frequency estimation and the pose estimation drivers. The application would automatically shift the keyboard displayed on launch, so that the frequency measured by the microphone is the same as the one associated with the user's hand placement.

4.5 Conclusion

This innovative method not only caters to visual learners by providing a dynamic and interactive representation of musical notes, but it also fosters a more intuitive understanding of pitch and timing. The falling notes simulate the real-time progression of a musical piece, allowing learners to synchronize their vocal and musical performance with the visual cues. This visual feedback not only enhances the comprehension of musical nuances but also accelerates the learning curve by reinforcing the connection between auditory perception and vocal execution. Moreover, the program's adaptability to different musical genres and the ability to customize difficulty levels make it an inclusive and personalized tool for learners of various skill levels. As technology converges with the art of singing and practising music in general, the incorporation of falling notes tutorials emerges as a promising avenue for more engaging, efficient, and tailored vocal education.

5 Conclusion

This thesis contributes to developing and studying new interfaces that enhance learning and practice. It studies how to transform a device through various technologies to make learning a complex subject more attractive, mainly using gamification. We first explored the improvement of a tactile interface through the interactive musical score project. The use of conductive ink has made it possible to differentiate the project from traditional smartphone applications by introducing children to the basics of music theory with innovative technology. This score takes advantage of its digital interactivity, flexibility, and sound generation to position itself as an object that is at once attractive, educational, and intuitive. Future work on the project will need to add buttons to the interface, activating certain features such as changing keys or playing associated music. The electronics should also be redone and reduced to improve the embedding of the device.

The American Sign Language (ASL) Learning Game Project is an AR application that uses the GOSAI framework's modularity on the DVIC augmented mirror to enable fun and practical learning. The application uses different intrinsic functionalities of the OS, such as augmented reality, a robust backend, a permissive frontend, and its suitability for supporting computer vision projects. Exploiting this capability made implementing a custom ASL recognition model possible, which produces an attractive tutorial/correction application when combined with 3D ASL animations. Future work should improve the learning function of the application. Its current effectiveness related to user information retention needs to be improved. Future fixes and lessons features are relevant to the project.

Similarly, the singing and theremin practice projects aim to correct a user during musical practice on the

augmented mirror. They allow a user's performance correction through a frequency estimation module and a GOSAI pose estimation module. The application adds a dimension of improvement in precision, visualization, and attractiveness of musical performance. However, future work concerns the correction of a singer's posture using the mirror's reflection and calibrating the graphical interface with the assistance of a theremin player. This thesis contributed to studying different ways of stimulating interest and investment in learning and practice. The projects studied here make complex fields more attractive while maintaining teaching effectiveness using innovative technology.

Acknowledgments

I want to thank my supervisor, Dr. Clément DUHART, for guiding me throughout my projects and, more broadly, introducing me to the world of research and innovation by integrating me into the DVIC. The work environment of this space, combining self-improvement, knowledge sharing, and camaraderie, allowed me to flourish during these two wonderful years.

I thank Dr. Marc TEYSSIER for his valuable advice, particularly in electronics, and his feedback on my various projects. I also thank Dr. Xiao XIAO for her advice on writing this thesis, her supervision on my musical projects, and for sharing her passion for music.

I am grateful to Dr. Yliess HATI for introducing me to and sharing his passion for machine learning and computer graphics, to Dr. Grégor JOUET for his valuable courses and software advice, and to Ph.D. student Madalina NICOLAE for her friendship and support over these two years.

I thank Maxime BROUSSART for his shared collaboration, investment, and contributions to GOSAI. Our continuous sharing of new ideas for improving our respective platforms and rivalry have been a real source of inspiration and surpassing myself in my projects.

Finally, I can never thank Ph.D. student Thomas JULDO enough for his investment in my projects, for sharing my passions in AI and Computer Graphics, for all his knowledge sharing, mentorship and friendship during these two years.

References

[1] Shailesh S Kantak and Carolee J Winstein. 'Learning–performance distinction and memory processes for motor skills: A focused review and perspective'. In: *Behavioural brain research* 228.1 (2012), pp. 219–231 (cited on page 5).

[2] Awaz Naaman Saleem, Narmin Mohammed Noori, and Fezile Ozdamli. 'Gamification applications in E-learning: A literature review'. In: *Technology, Knowledge and Learning* 27.1 (2022), pp. 139–159 (cited on page 6).

[3] Hartmut Seichter. 'Augmented reality and tangible interfaces in collaborative urban design'. In: *Computer-Aided Architectural Design Futures (CAADFutures) 2007: Proceedings of the 12th International CAADFutures Conference*. Springer. 2007, pp. 3–16 (cited on page 7).

[4] Alois Ferscha, Stefan Resmerita, and Clemens Holzmann. 'Human computer confluence'. In: *Universal Access in Ambient Intelligence Environments: 9th ERCIM Workshop on User Interfaces for All, Königswinter, Germany, September 27-28, 2006. Revised Papers*. Springer. 2007, pp. 14–27 (cited on page 7).

[5] Constantine Stephanidis et al. 'Seven HCI grand challenges'. In: *International Journal of Human–Computer Interaction* 35.14 (2019), pp. 1229–1269 (cited on page 7).

[6] Pattie Maes and Robyn Kozierok. 'Learning interface agents'. In: *AAAI*. Vol. 93. 1993, pp. 459–465 (cited on page 7).

[7] Howard Gardner and Yves Bonin. *Les intelligences multiples*. Retz Paris, 2004 (cited on page 7).

[8] Joep Janssen et al. 'Gamification in physical therapy: More than using games'. In: *Pediatric Physical Therapy* 29.1 (2017), pp. 95–99 (cited on page 8).

[9] Leland D Bland. 'The college music theory curriculum: the synthesis of traditional and comprehensive musicianship approaches'. In: *College Music Symposium*. Vol. 17. 2. JSTOR. 1977, pp. 167–174 (cited on page 9).

[10] Olivier Donnat. 'Les amateurs, enquête sur les activités artistiques des Français'. In: *Les amateurs, enquête sur les activités artistiques des Français*. 1996, p. 198 (cited on page 10).

[11] Robert J Zatorre and Andrea R Halpern. 'Mental concerts: musical imagery and auditory cortex'. In: *Neuron* 47.1 (2005), pp. 9–12 (cited on page 11).

[12] Robert J Weber and Suellen Brown. 'Musical imagery'. In: *Music Perception* 3.4 (1986), pp. 411–426 (cited on page 11).

[13] Jamie Zigelbaum et al. 'BodyBeats: whole-body, musical interfaces for children'. In: *CHI'06 extended abstracts on human factors in computing systems*. 2006, pp. 1595–1600 (cited on page 11).

[14] Xiao Xiao, Basheer Tome, and Hiroshi Ishii. 'Andante: Walking Figures on the Piano Keyboard to Visualize Musical Motion.' In: *NIME*. Cambridge, MA. 2014, pp. 629–632 (cited on page 11).

[15] Paul Taele, Laura Barreto, and Tracy Hammond. 'Maestoso: An intelligent educational sketching tool for learning music theory'. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 29. 2. 2015, pp. 3999–4005 (cited on page 12).

[16] Mattia Davide Amico, Luca Andrea Ludovico, et al. 'Kibo: A MIDI controller with a tangible user interface for music education'. In: *Proceedings of the 12th International Conference on Computer Supported Education. 1: CSME*. SCITEPRESS. 2020, pp. 613–619 (cited on page 12).

[17] Dongjo Kim et al. 'Heterogeneous interfacial properties of ink-jet-printed silver nanoparticulate electrode and organic semiconductor'. In: *Advanced materials* 20.16 (2008), pp. 3084–3089 (cited on page 13).

[18] Arshad Khan et al. 'Soft inkjet circuits: rapid multi-material fabrication of soft circuits using a commodity inkjet printer'. In: *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology*. 2019, pp. 341–354 (cited on pages 13, 16).

[19] Michael Wessely et al. 'Sprayable user interfaces: Prototyping large-scale interactive surfaces with sensors and displays'. In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 2020, pp. 1–12 (cited on page 13).

[20] Analisa Russo et al. 'Pen-on-paper'. In: *Advanced materials* 23.30 (2011), pp. 3426–3430 (cited on page 13).

[21] Anneli Hershman et al. 'Light it up: Using paper circuitry to enhance low-fidelity paper prototypes for children'. In: *Proceedings of the 17th ACM conference on interaction design and children*. 2018, pp. 365–372 (cited on page 13).

[22] Adrien Lefevre. *capacitive$_t o_n otes$*. https://github.com/AdrLfv/capacitive_to_notes. 2022 (cited on pages 15, 17).

[23] Salma A Essam El-Din and Mohamed A Abd El-Ghany. 'Sign Language Interpreter System: An alternative system for machine learning'. In: *2020 2nd Novel Intelligent and Leading Emerging Sciences Conference (NILES)*. IEEE. 2020, pp. 332–337 (cited on page 22).

[24] Yin Wang and Antonia Hamilton. 'Social top-down response modulation (STORM): a model of the control of mimicry in social interaction'. In: *Frontiers in Human Neuroscience* 6 (2012). DOI: 10.3389/fnhum.2012.00153 (cited on page 22).

[25] *Gosai*. https://dvic.devinci.fr/projects/gosai (cited on page 23).

[26] *javascriptgaming*. https://medium.com/nerd-for-tech/javascript-and-the-gaming-industry-3d2ff7f102e9 (cited on page 24).

[27] IA Adeyanju, OO Bello, and MA Adegboye. 'Machine learning methods for sign language recognition: A critical review and analysis'. In: *Intelligent Systems with Applications* 12 (2021), p. 200056 (cited on pages 26, 41).

[28] M.B. Waldron and Soowon Kim. 'Isolated ASL sign recognition system for deaf persons'. In: *IEEE Transactions on Rehabilitation Engineering* 3.3 (1995), pp. 261–271. DOI: 10.1109/86.413199 (cited on page 26).

[29] B. Lekhashri and A. Arun Pratap. 'Use of motion-print in sign language recognition'. In: *2011 National Conference on Innovations in Emerging Technology*. 2011, pp. 99–102. DOI: 10.1109/NCOIET.2011.5738842 (cited on page 26).

[30] Mohammed Kadous. 'Machine Recognition of Auslan Signs Using PowerGloves: Towards Large-Lexicon Recognition of Sign Language'. In: (Feb. 1970) (cited on page 26).

[31] Dimitris Metaxas. 'Adapting hidden Markov models for ASL recognition by using three-dimensional computer vision methods'. In: (1997) (cited on page 26).

[32] *cyberglovesystems*. https://www.cyberglovesystems.com/products/cyberglove-II/photos-video (cited on page 26).

[33] F-F ImageNet Li. 'Crowdsourcing, benchmarking & other cool things'. In: *CMU VASC Semin* 16 (2010), pp. 18–25 (cited on page 26).

[34] Md Zahangir Alom et al. 'The history began from alexnet: A comprehensive survey on deep learning approaches'. In: *arXiv preprint arXiv:1803.01164* (2018) (cited on page 26).

[35] Helen Cooper, Brian Holt, and Richard Bowden. 'Sign language recognition'. In: *Visual Analysis of Humans: Looking at People* (2011), pp. 539–562 (cited on pages 27, 44).

[36] Karly Kudrinko et al. 'Wearable Sensor-Based Sign Language Recognition: A Comprehensive Review'. In: *IEEE Reviews in Biomedical Engineering* 14 (2021), pp. 82–97. DOI: 10.1109/RBME.2020.3019769 (cited on pages 27, 28).

[37] Bogdan Ionescu et al. 'Dynamic hand gesture recognition using the skeleton of the hand'. In: *EURASIP Journal on Advances in Signal Processing* 2005 (2005), pp. 1–9 (cited on page 27).

[38] Chenglong Yu et al. 'Vision-based hand gesture recognition using combinational features'. In: *2010 Sixth International Conference on Intelligent Information Hiding and Multimedia Signal Processing*. IEEE. 2010, pp. 543–546 (cited on page 27).

[39] Shao-Zi Li et al. 'Feature learning based on SAE–PCA network for human gesture recognition in RGBD images'. In: *Neurocomputing* 151 (2015), pp. 565–573 (cited on page 27).

[40] Jayesh S Sonkusare et al. 'A review on hand gesture recognition system'. In: *2015 International Conference on Computing Communication Control and Automation*. IEEE. 2015, pp. 790–794 (cited on page 27).

[41] Vladislava Bobić, Predrag Tadić, and Goran Kvaščev. 'Hand gesture recognition using neural network based techniques'. In: *2016 13th Symposium on Neural Networks and Applications (NEUREL)*. IEEE. 2016, pp. 1–4 (cited on page 27).

[42] Md Mohiminul Islam, Sarah Siddiqua, and Jawata Afnan. 'Real time hand gesture recognition using different algorithms based on American sign language'. In: *2017 IEEE international conference on imaging, vision & pattern recognition (icIVPR)*. IEEE. 2017, pp. 1–6 (cited on page 27).

[43] Himadri Nath Saha et al. 'A machine learning based approach for hand gesture recognition using distinctive feature extraction'. In: *2018 IEEE 8th annual computing and communication workshop and conference (CCWC)*. IEEE. 2018, pp. 91–98 (cited on page 27).

[44] Razieh Rastgoo, Kourosh Kiani, and Sergio Escalera. 'Hand pose aware multimodal isolated sign language recognition'. In: *Multimedia Tools and Applications* 80 (2021), pp. 127–163 (cited on page 27).

[45] Biao Xu, Shiliang Huang, and Zhongfu Ye. 'Application of tensor train decomposition in S2VT model for sign language recognition'. In: *IEEE Access* 9 (2021), pp. 35646–35653 (cited on page 27).

[46] Romain Nith et al. 'DextrEMS: Increasing Dexterity in Electrical Muscle Stimulation by Combining It with Brakes'. In: New York, NY, USA: Association for Computing Machinery, 2021 (cited on page 27).

[47] KP Nimisha and Agnes Jacob. 'A brief review of the recent trends in sign language recognition'. In: *2020 International Conference on Communication and Signal Processing (ICCSP)*. IEEE. 2020, pp. 186–190 (cited on page 27).

[48] Yonas Fantahun Admasu and Kumudha Raimond. 'Ethiopian sign language recognition using Artificial Neural Network'. In: *2010 10th International Conference on Intelligent Systems Design and Applications*. IEEE. 2010, pp. 995–1000 (cited on page 27).

[49] Mohamed Deriche, Salihu O Aliyu, and Mohamed Mohandes. 'An intelligent arabic sign language recognition system using a pair of LMCs with GMM based classification'. In: *IEEE Sensors Journal* 19.18 (2019), pp. 8067–8078 (cited on page 27).

[50] Tareq Z Ahram, Waldemar Karwowski, and Jay Kalra. *Advances in Artificial Intelligence, Software and Systems Engineering: Proceedings of the AHFE 2021 Virtual Conferences on Human Factors in Software and Systems Engineering, Artificial Intelligence and Social Computing, and Energy, July 25-29, 2021, USA*. Vol. 271. Springer Nature, 2021 (cited on page 27).

[51] Tao Song et al. 'Intelligent human hand gesture recognition by local–global fusing quality-aware features'. In: *Future Generation Computer Systems* 115 (2021), pp. 298–303 (cited on page 27).

[52] Carman KM Lee et al. 'American sign language recognition and training method with recurrent neural network'. In: *Expert Systems with Applications* 167 (2021), p. 114403 (cited on page 27).

[53] Kunyoung Lee et al. 'A comparative analysis on the impact of face tracker and skin segmentation onto improving the performance of real-time remote photoplethysmography'. In: *Intelligent Human Computer Interaction: 12th International Conference, IHCI 2020, Daegu, South Korea, November 24–26, 2020, Proceedings, Part II 12*. Springer. 2021, pp. 27–37 (cited on page 27).

[54] Liqing Gao et al. 'RNN-transducer based Chinese sign language recognition'. In: *Neurocomputing* 434 (2021), pp. 45–54 (cited on page 27).

[55] Lee Yi Bin, Goh Yeh Huann, and Lum Kin Yun. 'Study of convolutional neural network in recognizing static American sign language'. In: *2019 IEEE international conference on signal and image processing applications (ICSIPA)*. IEEE. 2019, pp. 41–45 (cited on page 27).

[56] Celal Savur and Ferat Sahin. 'Real-time american sign language recognition system using surface emg signal'. In: *2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)*. IEEE. 2015, pp. 497–502 (cited on page 27).

[57] Yaser Saleh and Ghassan Issa. 'Arabic sign language recognition through deep neural networks fine-tuning'. In: (2020) (cited on page 27).

[58] Md Mehedi Hasan et al. 'Classification of sign language characters by applying a deep convolutional neural network'. In: *2020 2nd International Conference on Advanced Information and Communication Technology (ICAICT)*. IEEE. 2020, pp. 434–438 (cited on page 27).

[59] Zhe Cao et al. 'Realtime multi-person 2d pose estimation using part affinity fields'. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 7291–7299 (cited on page 27).

[60] Rishabh Bajpai and Deepak Joshi. 'MoveNet: A Deep Neural Network for Joint Profile Prediction Across Variable Walking Speeds and Slopes'. In: *IEEE Transactions on Instrumentation and Measurement* 70 (2021), pp. 1–11. DOI: 10.1109/TIM.2021.3073720 (cited on page 27).

[61] Alex Kendall, Matthew Grimes, and Roberto Cipolla. 'Posenet: A convolutional network for real-time 6-dof camera relocalization'. In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 2938–2946 (cited on page 27).

[62] Camillo Lugaresi et al. 'Mediapipe: A framework for building perception pipelines'. In: *arXiv preprint arXiv:1906.08172* (2019) (cited on page 27).

[63] Sılvia Grasiella Moreira Almeida, Frederico Gadelha Guimarães, and Jaime Arturo Ramırez. 'Feature extraction in Brazilian Sign Language Recognition based on phonological structure and using RGB-D sensors'. In: *Expert Systems with Applications* 41.16 (2014), pp. 7259–7271 (cited on page 28).

[64] C Shawn Green and Daphne Bavelier. 'Enumeration versus multiple object tracking: The case of action video game players'. In: *Cognition* 101.1 (2006), pp. 217–245 (cited on page 28).
[65] Ricardo Tejeiro Salguero, Manuel Pelegrina del Rıo, and Jorge Luis Gómez Vallecillo. 'Efectos psicosociales de los videojuegos'. In: *Comunicación: revista Internacional de Comunicación Audiovisual, Publicidad y Estudios Culturales, 1 (7), 235-250.* (2009) (cited on page 28).
[66] Eva L Baker and Harold F O'Neil. *Technology assessment in education and training*. Vol. 1. Psychology Press, 1994 (cited on page 29).
[67] *Moss*. https://store.steampowered.com/app/846470/Moss/. 2018 (cited on page 29).
[68] Zahoor Zafrulla et al. 'CopyCat: an American sign language game for deaf children'. In: *Face and Gesture 2011*. IEEE Computer Society. 2011, pp. 647–647 (cited on page 29).
[69] Zahoor Zafrulla et al. 'American sign language recognition with the kinect'. In: *Proceedings of the 13th international conference on multimodal interfaces*. 2011, pp. 279–286 (cited on page 29).
[70] Yosra Bouzid et al. 'Using educational games for sign language learning-a signwriting learning game: Case study'. In: *Journal of Educational Technology & Society* 19.1 (2016), pp. 129–141 (cited on page 29).
[71] Yosra Bouzid and Mohamed Jemni. 'An Avatar based approach for automatically interpreting a sign language notation'. In: *2013 IEEE 13th International Conference on Advanced Learning Technologies*. IEEE. 2013, pp. 92–94 (cited on page 29).
[72] Clemencia Zapata Lesmes et al. 'Design and Production of Educational Video Games for the Inclusion of Deaf Children'. In: *Procedia Computer Science* 198 (2022), pp. 626–631 (cited on page 30).
[73] *TuesdayJS*. https://kirilllive.github.io/tuesday-js/ (cited on page 30).
[74] *Monogatari*. https://monogatari.io/ (cited on page 31).
[75] Jason M Saragih, Simon Lucey, and Jeffrey F Cohn. 'Real-time avatar animation from a single image'. In: *2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG)*. IEEE. 2011, pp. 117–124 (cited on page 31).
[76] *pose$_a$nimator*. https://blog.tensorflow.org/2020/05/pose-animator-open-source-tool-to-bring-svg-characters-to-life.html (cited on page 32).
[77] Lan Thao Nguyen et al. 'Automatic generation of a 3D sign language avatar on AR glasses given 2D videos of human signers'. In: *Proceedings of the 1st International Workshop on Automatic Translation for Signed and Spoken Languages (AT4SSL)*. 2021, pp. 71–81 (cited on page 32).

[78] *blazepose*. https://blog.tensorflow.org/2021/08/3d-pose-detection-with-mediapipe-blazepose-ghum-tfjs.html (cited on page 32).
[79] $posenet_demo$. https://github.com/aphrx/posenet_demo (cited on page 32).
[80] *mixamo*. https://www.mixamo.com/#/ (cited on page 32).
[81] *posenet*. https://github.com/aphrx/posenet (cited on page 32).
[82] *kalidokit*. https://github.com/yeemachine/kalidokit (cited on page 32).
[83] *esla*. https://github.com/AdrLfv/esla (cited on page 37).
[84] $what_when_how$. http://what-when-how.com/advanced-methods-in-computer-graphics/quaternions-advanced-methods-in-computer-graphics-part-4/ (cited on page 39).
[85] *vroid*. https://hub.vroid.com/en/ (cited on page 39).
[86] slr_mirror. https://github.com/AdrLfv/slr_mirror (cited on page 41).
[87] Philippe Dreuw et al. 'Speech recognition techniques for a sign language recognition system'. In: *hand* 60 (2007), p. 80 (cited on page 44).
[88] Ilya Loshchilov and Frank Hutter. 'Decoupled weight decay regularization'. In: *arXiv preprint arXiv:1711.05101* (2017) (cited on page 45).
[89] *onnx*. https://onnx.ai/ (cited on page 45).
[90] Spencer D Kelly et al. 'Offering a hand to pragmatic understanding: The role of speech and gesture in comprehension and memory'. In: *Journal of memory and Language* 40.4 (1999), pp. 577–592 (cited on page 53).
[91] Rolf Inge Godoy and Harald Jorgensen. *Musical imagery*. Routledge, 2012 (cited on page 57).
[92] Nicolas Farrugia et al. 'Tunes stuck in your brain: The frequency and affective evaluation of involuntary musical imagery correlate with cortical structure'. In: *Consciousness and cognition* 35 (2015), pp. 66–77 (cited on pages 58, 59).
[93] Frederic Bevilacqua et al. 'Wireless sensor interface and gesture-follower for music pedagogy'. In: *Proceedings of the 7th international conference on New interfaces for musical expression*. 2007, pp. 124–129 (cited on pages 59, 60).
[94] Yu Ting Huang and Chi Nung Chu. 'Visualized comparison as a correctness indicator for music sight-singing learning interface evaluation—a pitch recognition technology study'. In: *Frontier Computing: Theory, Technologies and Applications*. Springer. 2016, pp. 959–964 (cited on page 60).
[95] Giovanni Santini. 'Augmented Piano in Augmented Reality'. In: *Proceedings of the International Conference on New Interfaces for Musical Expression* (Birmingham, UK). Zenodo, June 2020, pp. 411–415. DOI: 10.5281/zenodo.4813449 (cited on page 60).
[96] John Fillwalk. 'ChromaChord: A virtual musical instrument'. In: *2015 IEEE Symposium on 3D User Interfaces (3DUI)*. IEEE. 2015, pp. 201–202 (cited on page 60).

[97] David Johnson and George Tzanetakis. 'Vrmin: using mixed reality to augment the theremin for musical tutoring.' In: *NIME*. 2017, pp. 151–156 (cited on page 60).

[98] Kin Wah Edward Lin et al. 'Implementation and evaluation of real-time interactive user interface design in self-learning singing pitch training apps'. In: *ICMC*. 2014 (cited on page 61).

[99] Philip McLeod. 'Tartini: The Music Analysis Tool'. In: () (cited on page 62).

[100] $carolina_e yck_t heremine$. https://www.youtube.com/watch?v=wrUo9hmNwts (cited on page 66).