


# Exploring Language Pedagogy with Virtual Reality and Artificial Intelligence

B. Michael<sup>1</sup> and N. Aburumman<sup>†2</sup> 

<sup>1</sup>Barnett Waddingham, London, UK

<sup>2</sup>Department of Computer Sciences, Brunel University London, London, UK



**Figure 1:** Our VR application, where the participant engages in learning Mandarin Chinese in a 3D immersive environment. Left. Free-roaming environment. Center. Calligraphy practice scenario. Right. Cafe practice scenario.

## Abstract

Virtual Reality (VR) is a highly immersive and interactive experience that renders users to be engrossed in a 3D virtual environment. The recent technological advancements with high-resolution headset display, and accurate tracking of six degrees of freedom paired with controllers allow life-like renditions of real-world scenarios as well as fictional scenarios without potential environmental risks. This paper explores the usage of Virtual Reality in education by incorporating current pedagogical approaches into an interactive 3D virtual environment. The focus of this study revolves around language pedagogy, in specific, the tool developed allows teach users fundamental Mandarin Chinese. This educational VR application enables users to practice their reading and writing skills through a calligraphy lesson and engages users in a listening and speaking lesson through natural conversation. To achieve an organic dialogue, phrases spoken by the user in a lesson are validated immediately through an intuitive phrase recognition system developed using machine learning. The developed prototype has undergone testing to ensure its efficacy. An initial investigation into this prototype found that the majority of participants were supportive of this concept and believe that it would improve the engagement of digital education.

## CCS Concepts

• **Computing methodologies** → *Virtual reality; Neural networks;*

## 1. Introduction

Education is the core of a thriving society, not only does it succeed one in fulfilling their career paths, but it has also paved the way for many civilisations to dawn a new age. Hence, the importance of constantly innovative pedagogical approaches is crucial [Vel16; ANT\*13]. In light of recent events, COVID-19 and pandemics have urged the importance of remote learning; this has accelerated the implementation of digital learning whereby students are staying home and having lessons taught through online mediums [JLY\*22]. Virtual Reality (VR) is the use of computer technology to create an immersive three-dimensional digital environment with the use of a head-mounted display, this device is otherwise

known as a VR headset. The VR experience can be paired with controllers to provide users with the ability to interact with objects within the environment [FO15]. VR has the natural capability to fully immersed one's vision, enabling their full attention span. The rapid adoption of VR across several industries proves effective usage in many aspects such as training, designing, and visualising [ZKK\*21]. A screen at home does not provide the same interactivity as traditional classrooms and online digital learning lacks the engagement that can contribute to a conducive learning environment [WBD\*19]. Moreover, the decline in the number of teaching staff requires a solution to maintaining a steady level of education quality [CD17]. In addition, the current approach to digital pedagogy shows many gaps that can be improved upon through the use of VR. Thus, VR can be leveraged in digital pedagogy to immerse students in a particular subject, resulting in increased interest and

<sup>†</sup> Department of Computer Sciences, Brunel University London, UK

focus. VR invokes multiple senses, when students provoke more than one sense simultaneously or over a short period of time, the interaction is associated more intensely and thereby retain what they have learned for longer [Bai08]. This means as VR becomes more readily available for consumers, it becomes more feasible to consider it in digital pedagogy, which makes VR a valuable asset in the educational industry.

This paper presents a VR tool which improves the typical approach to digital learning. Our VR application is able to provide users with a highly immersive introductory level of speaking, listening, reading, and writing in Mandarin Chinese (see Fig 1). Through the use of VR controllers, users may practice their calligraphy skills and with the implementation of a phrase recognition service, they may also practice their conversational skills. These aspects were implemented to innovate the approach of digital learning whereby users would be less distracted by their surroundings and as a result, be more engaged in the subject. Our VR tool consists of two parts the front-end game engine and the back-end speech recognition system. Unity is the game engine used, paired alongside the Oculus integration software for VR capabilities. TensorFlow paired with Librosa forms the speech recognition system using neural networks. Finally, a Flask server is implemented to accommodate communication between the front-end and back-end. To summarise the results of the feedback questionnaire, the overall responses show good prospects in this concept and indicate success in meeting the aims of this application.

## 2. Related Work

Our work is informed by language pedagogy, learning in VR environments, and educational VR. We cover below the references in literature that are most relevant to our work.

### 2.1. Language Pedagogy

Language learning is a complex process that develops through the four basic language skills: listening, speaking, reading and writing. One of the best methods of learning a new language is arguably to go abroad, visit the country that speaks the target language and surround oneself with the native culture [CDE88]. When teaching a new foreign language, Bulgarian psychiatrist and educator Georgi Lozanov developed a teaching method called Suggestopedia [Ibr23]. It is a compound word of suggestion and pedagogy, where students learn quickly by being made to feel relaxed, interested and positive. Many AI-based web applications adopt Suggestopedia, and allow learners access the applications to study in their own time. Applications such as Duolingo [Neu23] and Babbel [Ope16] have speech recognition software for pronunciation practice or spaced repetition practice for vocabulary development. Moreover, in terms of language pedagogy, Brown proposed 12 principles of language teaching, which are related to cognitive, affective, and linguistic aspects [Bro\*00]. These principles have been employed through gameplay in a wide variety of ways, from a planned learning activity in an instructional environment to an incidental by-product of a gamer's interactions with the game and its associated online activities [KK17]. An overview survey on language pedagogy techniques can be seen in [Flo09].

### 2.2. Learning in VR environments and Educational VR

VR enables users to engage in a computer-generated, interactive environment through diverse multisensory devices. The ultimate goal of VR is to present users with an immersive experience that disconnects them from the physical world and transfers them to an alternative reality where they can communicate, interact and have a sense of presence in a virtual environment [HMEW21]. Studies have shown that the use of VR in the classroom increases learners' engagement in an active learning environment [AvM18]. A survey by Lin and Lin [LL15] provides comprehensive details about learning languages in virtual reality environments.

## 3. System Overview

The main requirements for our VR tool dictate two core lessons that a user can engage in to practice the four language fundamentals (see Fig 2). To ensure that the user experiences their speaking and listening in a natural setting, an interaction representative of a real-world scenario will be written to then be implemented in Virtual Reality. This will be a short dialogue involving the user visiting a café and ordering a drink from a barista, a practice that will likely be useful in real life. Reading and writing are the second of two core lessons. By implementing an area within the virtual environment that allows the user to find a brush and paper, they may practice their calligraphy skills. This area should contain examples of Chinese characters along with a representation of their meaning so that the user may accomplish their understanding of reading and writing.

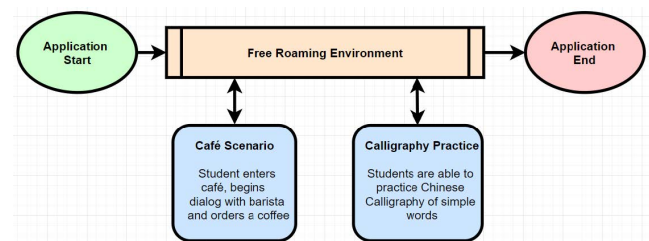


Figure 2: Overview of the entire VR tool.

Once the application is started, users may roam freely and explore the virtual environment. Doing so will allow the user to adjust themselves into virtual reality and let them absorb the cultural aesthetics designed into the software. Once ready, users may then enter a lesson of their choice and begin their learning. When a lesson has concluded, users may then visit the free-roaming environment again where they can either choose to repeat the lesson or try a different one. The user may exit the application at any point in time.

### 3.1. Listening and Speaking Scenario

This lesson aims to enhance the user's ability to listen to Chinese words and phrases as well as enable them to practice simple replies. This scenario requires the use of speech recognition to validate the accuracy of the spoken phrase. In this scenario (see Fig 3, on the right), the user is first greeted with "Nǐ hǎo" (Hello), in which a response is required by the user. The expected response is played

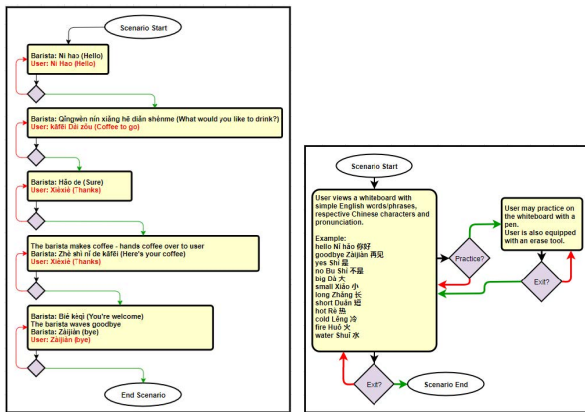


Figure 3: Operating logic of our VR tool. Left. Cafe scenario diagram. Right. Calligraphy practice diagram.

out for the user to listen, this may be repeated at the discretion of the user to familiarise themselves. Once ready, the user may then record an audio file which would then be analysed and validated by the speech recognition software. If the audio is not recognised by the software, the sequence would be repeated. If, however, the speech audio is recognised, the user may progress to the subsequent dialogues. When the sequence of dialogues comes to an end, the scenario will be complete, and the user will be returned to the free-roaming environment.

### 3.2. Reading and Writing Scenario

This lesson aims to improve the user’s reading and writing skills. Considering the vast differences between English and Chinese text, Chinese characters may initially be difficult for students who are native to English. This method of approach reinforces the character learning process through the practice of writing. The scenario begins in a classroom setting where Chinese characters (see Fig 3, on the left), their translation and pronunciation are presented on a whiteboard. The user can now take their time in examining the material and will be able to perform one of three options. Firstly, the user may choose to exercise their calligraphy skills by picking up a pen in the environment and writing on the board; mistakes may be corrected using an erase tool. Once satisfied, the user may put the pen down to stop their writing practice. Secondly, an audio sample of the respective phrase/word will be provided alongside the examples on the board; users may play this audio by interacting on the board. This feature will be optional as the main focus of the calligraphy lesson is to practice reading and writing. Finally, the user may choose to exit their calligraphy practice which will return them to the free-roaming environment. To ensure a smooth implementation process it is important to have a good visual understanding of the system architecture. As the Café scenario requires a phrase to be validated, the end-to-end service for validation can be depicted in Fig 4.

## 4. Phrase Recognition System

As there are no Mandarin phrase recognition assets on the Unity asset store, a decision was made to build this component using an

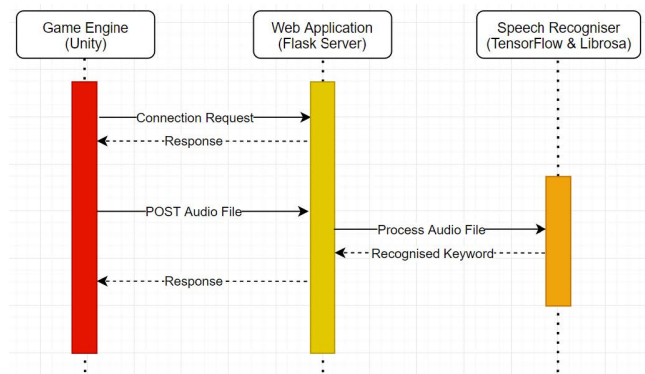


Figure 4: the end-to-end system architecture for the Café scenario.

open-source Machine Learning library, TensorFlow. We employed a supervised learning approach that is able to then predict a class label based on the method proposed by [SVR\*20], where we labelled the spoken phrase in the input audio files. Classification of the audio files is supervised by a subject expert to gauge the validity of each file; this is used as the training data for the phrase recognition service. This implementation is split into several parts: preparing the dataset, training the model, developing a keyword spotting service and lastly, establishing the Flask server.

### 4.1. Dataset Preparation

The first part of this requirement is preparing the dataset (see Fig 5), a solution is required to read, process, and identify key features of the input files. Mel-Frequency Cepstral Coefficients (MFCCs) are coefficients that form a Mel-Frequency Cepstrum [KRR12]. Based on a linear cosine transform of a log power spectrum on a non-linear Mel scale of frequency, it represents the short-term power spectrum of a sound. The usage of MFCCs are common among speech recognition systems, and music genre classification and will serve an important role in this requirement. A Python script was written to prepare the dataset of .wav audio files. Essentially, the audio files contained and sorted into appropriate folders are read, their MFCCs are identified, and a dataset of the MFCCs is created and stored into a JSON file where it can be used to develop the neural networks.

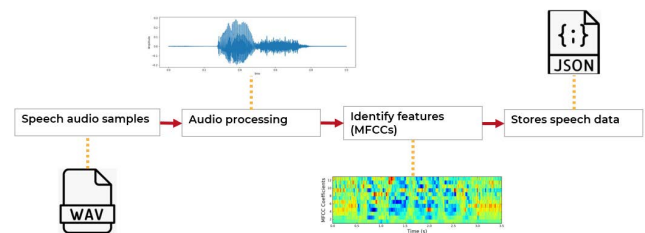


Figure 5: Overview of the data preparation process.

### 4.2. Machine Learning Model

With the prepared dataset stored in a JSON format, the numerical values can now be fed into the neural networks. A script was writ-

ten to build a Convolutional Neural Network (CNN) of 3 hidden layers, and train the model based on the prepared dataset; this is made possible with the Python module: TensorFlow. Each entry of data consists of the label along with the MFCCs and will be represented as the X and Y values in the neural networks. A percentage of the training data is used to test and validate the data, this percentage will be configured during testing to establish an effective training. There are several variables used to train the model such as the epochs, batch size, patience and learning which will be adjusted and tested during the testing phase. This training is a prerequisite for identifying a phrase and the trained model will be saved for later usage. With the model trained, it is now possible to refer back to it with an audio file to identify the spoken phrase. A script was written to parse in an audio file, extract the key features (MFCCs) and predict the phrase based on the trained prediction model. When the phrase is identified, the script would return a string of the predicted phrase.

### 4.3. Flask Server for End-to-End Communication

Now that the phrase detection is implemented, a solution must be in place to establish communication between the front-end Unity game engine and the back-end phrase detection. This architecture is depicted in the previous section and requires the Flask library to be imported as well as the phrase-detecting class. A script was written to instantiate a Flask server, which is able to receive an audio file through a POST request and call the predict function. The function is then able to return the string-predicted phrase in a JSON format.

## 5. Results

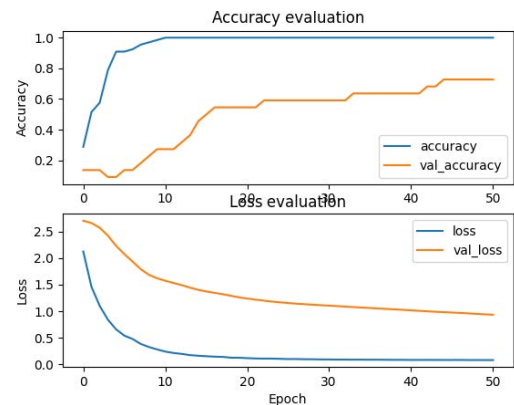
### 5.1. The Phrase Recognition Service

The phrase recognition service is a core component of our VR tool; therefore, it is important to ensure that it is working accurately. Audio machine learning for phrase recognition is a complex problem as there are many different tones of voices and speaking techniques. To develop a service that caters to a wide variety of users, audio samples from a wide variety of participants were gathered to train this system. Training requires audio samples from various participants, meaning research ethics approval has been obtained prior to any participant involvement. Participants were tasked to record audio samples of a few specified phrases in Mandarin. These phrases were then classified into sub-directories by a subject expert according to the identified spoken phrase where it was then used for training. Around 543 audio samples were recorded, which is a relatively low sample size for training data. Once the input data had been organised, the implemented training sequence was executed to prepare the dataset and build a trained TensorFlow model. For optimisation purposes, this training process is repeated several times and adjusted on each iteration. The optimisation was monitored using the Python module: Matplotlib [Hun07], which allows the creation of data visualisations that are used to evaluate the effectiveness of the trained TensorFlow model [Hun07]. An accuracy evaluation and loss evaluation are used to show the training process over the number of epochs.

#### 5.1.1. Accuracy Evaluation

Fig 6 (top) represents the accuracy of the training process over the number of epochs. Accuracy refers to the percentage in which data

is correctly predicted, typically, a higher accuracy percentage is preferred in neural networks. As mentioned previously, the training dataset is split to accommodate both testing and validation datasets, this is represented by the blue and orange lines respectively. The blue line represents the accuracy of the training model when compared with the testing data; the graph shows a sharp initial increase to the peak, which is an indicator that the model is training successfully. The orange line represents the accuracy of the training model when compared to the validation dataset. As the validation dataset is a sample of the training data that is retained from the training of the model, it provides an unbiased evaluation of the model fit.



**Figure 6:** Top. Accuracy evaluation of the phrase recognition system. Bottom. Loss evaluation of the phrase recognition system.

### 5.1.2. Loss Evaluation

Fig 6 (bottom) represents the loss of the training process over the number of epochs. Loss (or error) refers to the scale in which data is wrongly classified. Neural networks generally seek to minimise the loss, in order to gain better predictions. The descending loss evaluation graph indicates that the training is indeed improving over the number of epochs.

### 5.1.3. Learning Rate

One of the configurable hyperparameters is the learning rate. This is a figure that is usually set within the range of 0.0 – 1.0 and controls the amount in which weights are adjusted during the training process. Lower learning rate values would mean small changes to the weights on each epoch, resulting in poor predictions if the number of epochs is also low. A high learning rate value would mean dramatic changes to the weights on each epoch and could result in unstable training and poor predictions. In order to find a viable learning rate value, an experiment was conducted to monitor the accuracy and loss over different learning rates. This is a necessary step to find an optimised learning solution. Each test was repeated 5 times so that averages are recorded.

## 5.2. Prototype Testing

A total of 12 participants (8 males and 4 females) have been tested, and 10 of them are within the 18-24 age group and 2 are within the 25-30. After participants tried the VR, they filled out a feedback questionnaire. The results are shown below.

**Question: “Are you interested in learning another language?”** 83.3 responded with yes with the remainder 16.7 responding with maybe.

**Question: “What are you using to learn another language?”** The most popular platform indicated by the research is a Mobile Application with 91.7 with the next most popular being Textbook and Classes at 16.7.

**Question: “I think that Virtual Reality would improve the engagement of digital education”.** Responses were again taken on a scale of 1 (strongly disagree) to 5 (strongly agree) where participants can express their agreement on the particular statement. 66.7 voted 5 with the remainder 33.3 of votes at 4. The average score for this question is 4.67.

**Question: “I found trying out this VR application useful and engaging”.** On a scale of 1 (strongly disagree) to 5 (strongly agree), participants can express their agreement on the particular statement. 75 voted 5 with the remainder 25 of votes at 4. The average score for this question is 4.75, which again encourages the continuance development of this application.

## 6. Conclusion

This paper presented a VR application which improves the typical approach to digital learning. With the success of the implemented prototype, this VR application is able to provide users with a highly immersive introductory level of speaking, listening, reading, and writing in Mandarin Chinese. Through the use of VR controllers, users may practice their calligraphy skills and with the implementation of a phrase recognition service, they may also practice their conversational skills. These aspects were implemented to innovate the approach of digital learning whereby users would be less distracted by their surroundings and as a result, be more engaged in the subject. This application will highly benefit from additional scenarios as it will extend the amount of Mandarin one can learn in this environment. Moreover, the vision of this tool is to innovate the current approach to digital pedagogy. With language being the focus and Mandarin being taught, this application provides the foundations for an immersive language-learning platform. Once the fundamentals of one language can be established, this application may also delve into teaching other languages. In the next step, we plan to employ believable virtual characters [CLM\*19; RSB15], and give the characters in the application a deeper conversation.

## References

- [ANT\*13] ABOU EL-SEOUD, M SAMIR, NOSSEIR, ANN, TAJ-EDDIN, ISLAM, et al. “A proposed pedagogical mobile application for learning sign language”. *Tablet* 4 (2013), 32 1.
- [AvM18] ALLCOAT, DEVON and von MÜHLENEN, ADRIAN. “Learning in virtual reality: Effects on performance, emotion and engagement”. *Research in Learning Technology* 26 (2018) 2.
- [Bai08] BAINES, LAWRENCE. *A Teacher’s guide to multisensory learning: Improving literacy by engaging the senses*. ASCD, 2008 2.
- [Bro\*00] BROWN, H DOUGLAS et al. *Principles of language learning and teaching*. Vol. 4. longman New York, 2000 2.
- [CD17] CARVER-THOMAS, DESIREE and DARLING-HAMMOND, LINDA. “Teacher turnover: Why it matters and what we can do about it.” *Learning Policy Institute* (2017) 1.
- [CDE88] CARRELL, PATRICIA L, DEVINE, JOANNE, and ESKEY, DAVID E. *Interactive approaches to second language reading*. Cambridge University Press, 1988 2.
- [CLM\*19] CASTI, SARA, LIVESU, MARCO, MELLADO, NICOLAS, et al. “Skeleton based cage generation guided by harmonic fields”. *Computers & Graphics* 81 (2019), 140–151 5.
- [Flo09] FLOWERDEW, LYNNE. “Applying corpus linguistics to pedagogy: A critical evaluation”. *International journal of corpus linguistics* 14.3 (2009), 393–417 2.
- [FO15] FREINA, LAURA and OTT, MICHELA. “A literature review on immersive virtual reality in education: state of the art and perspectives”. *The international scientific conference elearning and software for education*. Vol. 1. 133. 2015, 10–1007 1.
- [HMEW21] HAMILTON, DAVID, MCKECHNIE, JIM, EDGERTON, EDWARD, and WILSON, CLAIRE. “Immersive virtual reality as a pedagogical tool in education: a systematic literature review of quantitative learning outcomes and experimental design”. *Journal of Computers in Education* 8.1 (2021), 1–32 2.
- [Hun07] HUNTER, J. D. “Matplotlib: A 2D graphics environment”. *Computing in Science & Engineering* 9.3 (2007), 90–95 4.
- [Ibr23] IBRAGIMOVA, SEVINCH. “ENGLISH LANGUAGE TEACHING METHODOLOGY FOR STUDENTS IN NON-PHILOLOGICAL EDUCATION USING SUGGESTOPEDIA TEACHING METHOD”. *Modern Science and Research* 2.6 (2023), 758–760 2.
- [JLY\*22] JIN, QIAO, LIU, YU, YAROSH, SVETLANA, et al. “How will vr enter university classrooms? multi-stakeholders investigation of vr in higher education”. *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 2022, 1–17 1.
- [KK17] KLIMOVA, BLANKA and KACET, JAROSLAV. “Efficacy of computer games on language learning.” *Turkish Online Journal of Educational Technology-TOJET* 16.4 (2017), 19–26 2.
- [KRR12] KOOLAGUDI, SHASHIDHAR G, RASTOGI, DEEPIKA, and RAO, K SREENIVASA. “Identification of language using mel-frequency cepstral coefficients (MFCC)”. *Procedia Engineering* 38 (2012), 3391–3398 3.
- [LL15] LIN, TSUN-JU and LAN, YU-JU. “Language learning in virtual reality environments: Past, present, and future”. *Journal of Educational Technology & Society* 18.4 (2015), 486–497 2.
- [Neu23] NEUSCHAFER, TOM. “DUOLINGO SPANISH USERS: DISCUSSION BOARDS USE OVER TIME”. *LLT Journal: A Journal on Language and Language Teaching* 26.1 (2023), 134–141. ISSN: 2579-9533 2.
- [Ope16] OPEN BABEL DEVELOPMENT TEAM. *Open Babel*. Version 2.4.0. Sept. 25, 2016. URL: [http://openbabel.org/wiki/Main\\_Page](http://openbabel.org/wiki/Main_Page) 2.
- [RSB15] RUMMAN, NADINE ABU, SCHAEFER, MARCO, and BECHMANN, DOMINIQUE. “Collision Detection for Articulated Deformable Characters”. *Proceedings of the 8th ACM SIGGRAPH Conference on Motion in Games*. MIG ’15. Paris, France: Association for Computing Machinery, 2015, 215–220. ISBN: 9781450339919 5.
- [SVR\*20] SOLOVYEV, ROMAN A, VAKHRUSHEV, MAXIM, RADIONOV, ALEXANDER, et al. “Deep learning approaches for understanding simple speech commands”. *2020 IEEE 40th international conference on electronics and nanotechnology (ELNANO)*. IEEE. 2020, 688–693 3.
- [Vel16] VELETSIANOS, GEORGE. *Emergence and innovation in digital learning: Foundations and applications*. Athabasca University Press, 2016 1.
- [WBD\*19] WONG, JACQUELINE, BAARS, MARTINE, DAVIS, DAN, et al. “Supporting self-regulated learning in online learning environments and MOOCs: A systematic review”. *International Journal of Human-Computer Interaction* 35.4-5 (2019), 356–373 1.
- [ZKK\*21] ZIKAS, PAUL, KAMARIANAKIS, MANOS, KARTSONAKI, IOANNA, et al. “Covid-19-VR strikes back: innovative medical VR training”. *ACM SIGGRAPH 2021 Immersive Pavilion*. 2021, 1–2 1.