ILLINOIS INSTITUTE OF TECHNOLOGY

eman ta zabal zazu

Universidad del País Vasco / Euskal Herriko Unibertsitatea

BILBOKO INGENIARITZA ESKOLA
ESCUELA DE INGENIERÍA DE BILBAO

Master of Information Technology and Management
IT Management and Entrepreneurship

Illinois Institute of Technology
College of Computing

Máster Universitario en Ingeniería de Telecomunicación
Telekomunikazio Ingeniaritza Unibertsitate Masterra

Universidad del País Vasco / Euskal Herriko Unibertsitatea
Escuela de Ingeniería de Bilbao / Bilboko Ingeniaritza Eskola

Research Project

Availability of Voice DeepFake Technology and its Impact for Good and Evil

Author: Naroa Amezaga Vélez
Project Director: Jeremy Hajek
Academic Year: 2020 - 2021

07/28/2021

Abstract

Artificial Intelligence and specially Machine Learning and Deep Learning techniques are increasingly populating today's technological and social landscape. These advancements have overwhelmingly contributed to the development of Speech Synthesis, also known as, Text-To-Speech, where speech is artificially produced from text by means of computer technology.

Despite existing such a variety of speech synthesis tools and systems, there is a fundamental common drawback: unnatural, robotic and impersonal synthesized voices. That's where Voice Cloning technology comes into play, which allows to generate an artificial synthetic speech that resembles a targeted human voice. This new practice offers many benefits in several fields, such as, healthcare, education, or advertising. However, there is a known fact that every coin has two sides. Likewise, with the development of voice cloning, the generation of fake video and voices, known as "deepfakes", has risen, causing a loss of trust and greater fear towards technology.

In this way, the objective of this project is to analyze the availability of deepfake voice technologies nowadays and its impact for good and evil. Therefore, we chose to focus on the educational field, by implementing a voice query assistant that answers to questions related to a course, like the professor's contact information or the date of the final exam. To enhance user experience, we added the extra feature of answering with the corresponding professor's cloned voice.

Moreover, as an initial point for a forthcoming investigation, we provide a design of a qualitative research study that allows participants to test the built framework and to give their views in order to gain a better understanding of the impact that voice cloning causes on people. Since all technology related concept, voice cloning is not exempt from discussion, so we also conduct an analysis about the misuse, the regulation, and the future of it.

The results of the case study show that it is possible to clone someone's voice based on just a few seconds of reference audio, which creates a superior user experience, but at the same time, reveals how easily can anyone have access to voice cloning. This expresses very well the importance of the new challenges opened by this potential technology and the need of safeguarding and regulation.

There is no doubt that to understand the dynamics and impact of voice cloning and reach more robust conclusions, future research is needed.

Keywords

Speech-To-Text, SQL query, Text-To-Speech, Speech Synthesis, Voice Cloning, Deepfake, virtual assistant, SV2TTS, similarity, Speech Recognition, speaker verification, síntesis de voz, clonación de voz, hizketa-sintesia, ahots-klonazioa.

Resumen

La inteligencia artificial y, especialmente, las técnicas de aprendizaje automático y aprendizaje profundo están cada vez más presentes en el panorama tecnológico y social actual. Estos avances han contribuido de manera abrumadora al desarrollo de la Síntesis de Voz, también conocida como Texto a Voz, donde, por medio de tecnología informática, el habla se produce de manera artificial.

A pesar de haber gran variedad de herramientas y sistemas de síntesis de voz, existe un inconveniente común: voces sintetizadas antinaturales, robóticas e impersonales. Ahí es donde entra en juego la tecnología de Clonación de Voz, que permite generar un discurso artificial que se asemeja a una voz humana especifica. Esta nueva práctica ofrece muchos beneficios en campos como la salud, la educación o la publicidad. Sin embargo, con el desarrollo de la clonación de voz, ha aumentado la generación de videos y voces falsos, conocidos como "deepfakes", provocando una pérdida de confianza y miedo hacia la tecnología.

De esta forma, el objetivo de este proyecto es analizar la disponibilidad de tecnologías de clonación de voz en la actualidad y su impacto tanto positivo como negativo. Para ello, optamos por centrarnos en el campo educativo, implementando un asistente de voz que responde a preguntas relacionadas con una asignatura, como la información de contacto del profesor o la fecha del examen final y, para mejorar la experiencia del usuario, hemos añadido una función adicional: responder con la voz clonada del profesor correspondiente.

Además, como punto inicial para una investigación futura, proporcionamos el diseño de un estudio cualitativo que permite a los participantes probar el sistema creado y dar su opinión, para así analizar el impacto que causa la clonación de voz en las personas. Junto a esto, y teniendo en cuenta que la clonación de voz no está exenta de discusión, también realizamos un análisis sobre el mal uso, la regulación y el futuro de la misma.

En suma, los resultados del proyecto muestran que es posible clonar la voz de alguien basándose en solo unos segundos de audio de referencia, lo que crea una experiencia de usuario mejorada, pero, al mismo tiempo, revela la facilidad con la que cualquiera puede tener acceso a la clonación de voz. Esto pone de manifiesto los nuevos retos a los que nos tenemos que enfrentar y la necesidad de prevención y regulación.

Aun así, no hay duda de que para comprender la dinámica y el impacto de esta tecnología y llegar a conclusiones más sólidas, son necesarias investigaciones futuras.

Laburpena

Adimen Artifizialak, eta bereziki Ikasketa Automatikoak eta Ikaskuntza Sakonak, gero eta garrantzi handiagoa dute gaur egungo ikuspegi teknologiko eta sozialean. Aurrerapen horiek izugarri lagundu dute hizketa-sintesiaren garapenean, non, ordenagailu baten bitartez, ahotsa artifizialki sortzen den.

Gaur egun hizketa sintesirako sistema ugari egon arren, gehienek eragozpen komun bat dute: ahots ez natural, robotiko eta inpertsonala. Horren konponbide gisa, ahots-klonazioa erabiltzen da, zeinari esker, norbaiten ahotsaren antza duen hizketa sintetiko artifiziala sortzen den. Teknologia berri honek hainbat abantaila eskaintzen ditu zenbait arlotan, hala nola, osasunean, hezkuntzan edota publizitatean. Hala ere, ahots-klonazioaren garapenarekin batera, "deepfake" izenarekin ezagutzen diren bideo eta ahots faltsuen sorrerak gora egin du, teknologiarekiko konfiantza galduz eta beldurra areagotuz.

Era horretan, proiektu honen helburua gaur egungo ahots-klonazioaren erabilgarritasuna eta eragin positibo eta negatiboak aztertzea da. Horretarako, hezkuntza eremuan zentratu gara, irakasgai batekin lotutako galderei (irakaslearen kontaktu informazioa edo amaierako azterketaren data adibidez) erantzuteko gai den ahots-laguntzailea garatuz. Horrez gain, erabiltzailearen esperientzia hobetzeko, erantzuna dagokion irakaslearen ahots klonatuarekin erreproduzituko da.

Gainera, etorkizuneko ikerlan posible bati begira, ikerketa kualitatibo baten diseinua eskaintzen da, non parte-hartzaileek garatutako sistema probatu eta beraien iritzia emango duten, ahots-klonazioak pertsonengan duen eragina hobeto ulertzeko. Bestalde, ahots-klonazioa, teknologiarekin loturiko kontzeptu guztiak bezala, ez dago eztabaidatik salbu, eta beraz, honen erabilera okerra, erregulazioa eta etorkizuna aztertuko dira.

Proiektu honen bidez norbaiten ahotsa klonatzea posible dela ondoriozta daiteke, erreferentzia gisa jatorrizko ahotsaren segundo gutxi batzuk oinarri hartuta. Horrek, aurreratu bezala, erabiltzaileen esperientzia hobetzen du, baina aldi berean, agerian uzten du ahots-klonazioa edonoren eskura dagoela. Beraz, teknologia honek sortutako erronka berriei aurre egiteko, prebentzio eta erregulazioaren beharra nahitaezkoa da.

Argi dago honen dinamika hobeto ulertzeko eta ondorio sendoagoetara heltzeko etorkizunean ikerketa gehiago behar direla.

Contents

Figures

Tables

Listings

1. Introduction

With the lately increase in the use and spread of technology across the world, more complex applications and tools of artificial intelligence are arising, and along with it, its underlying technologies, machine learning and deep learning [1]. These have constantly demonstrated significant potential for speech synthesis, also known as Text-To-Speech (TTS), which in recent years, has attracted increasingly more attention. This technology consists of converting text input into artificial human speech and has been utilized to enhance a wide range of application scenarios such as chatbots or virtual assistants [2].

One of the limitations of speech synthesis is that artificially created voices can sometimes sound very unnatural and robotic. Nevertheless, a relatively new technology called Voice Cloning allows to generate someone's cloned natural-sounding audio samples based on just a few seconds of speech [3]. Voice cloning is currently booming and therefore, a slew of companies are developing outstanding tools.

However, there is no light without darkness. The advancement of such technologies has led to the development of techniques for manipulation of video and audio, what is known as "Deepfake", which uses artificial intelligence, more specifically deep learning, to create fake videos and voices [4]. In the past, the creation of this type of content was available just to specialists, but nowadays, everyone has access to it, causing in some people a reluctance towards this technology [5]. Therefore, an assessment about the availability of voice deepfake technology is needed, along with an analysis about its impact for good and evil.

In relation to this, our objective is to enhance user experience by providing a framework that simulates a virtual assistant for an educational environment. Indeed, the education field is one of the much in which technology has been proven to be of great importance, for example, to boost the outcomes of student [6]. So, the aimed tool will be able to answer with the corresponding professor's voice to questions about course details coming from a student.

This framework proofs the leveraging of the most powerful technologies in improving the quality and experience of education for both students and teachers. In addition, the extra feature of voice cloning plays an important role in broadening the advancement of learning management systems, which together with eLearning, have experienced rapid growth due to the COVID-19 pandemic [6].

We also intend to design a qualitative research study to test the developed tool and get the first impressions of students and professors. This will allow us to know how the technology of voice cloning impacts people and what expectations and fears they have with regard to the future.

The structure of this document goes as follows. We begin with a review of the context and state of the art (section 2) for a better understanding of the topic. We then present the project goals and benefits in section 3. Specifications and requirements are defined in section 4, where the planning of the project, the risk analysis and the alternative analysis and selection are assessed. Section 5 describes the methodology and the details of the whole process. We later demonstrate the results of the work in section 6, and the market analysis in section 7. We conclude with a discussion, conclusion, and bibliography in sections 8, 9 and 10, respectively.

2. Context – State of the Art

2.1 Speech Synthesis - Text-To-Speech

Speech Synthesis is the computer-generated simulation of human speech. Text-to-Speech (TTS) refers to the artificial transformation of text to audio. This task is performed by a human simply by reading. The goal of a good TTS system is to have a computer do it, considering each time more the naturalness and expressiveness of the voice [7]. It is a cutting-edge technology in the field of information processing which involves many disciples such as acoustics, linguistics, digital signal processing or computer science [2]. A computer system used for this purpose is called a speech synthesizer and can be implemented in software or hardware [8].

The quality of a speech synthesizer is judged on the one hand, by its ability to be understood, and on the other hand, by its similarity to the naturalness of a human voice [8]. So, the most important qualities of a speech synthesis system are naturalness and intelligibility [8].

- **Naturalness**: how closely the output sounds like human speech.
- **Intelligibility**: ease with which the output is understood.

Speech synthesis systems usually try to maximize both characteristics since the ideal speech synthesizer is both natural and intelligible [8].

2.1.1 History

Mechanical synthesis

Before electronic signal processing was invented, speech researchers tried to build machines to create human speech. In St.Petesburg 1779, the scientist Christian Kratzenstein, explained the differences and built models of the five long vowels [9].

This was followed by von Kempelen of Vienna, Austria, in 1791, who added models of the tongue and lips, enabling the production of consonants as well as vowels [8]. In 1837, Charles Wheatstone produced a more sophisticated "speaking machine" based on von Kempelen's design [8], and in late 1800's, Alexander Graham Bell with his father, inspired by Wheatstone's speaking machine, constructed a similar kind of machine [9].

In the coming years, other experiments and further research with mechanical and semi-electrical analogs of vocal system were made, but with no remarkable success.

Electrical Synthesis

The very first full electrical synthesis device was introduced by Stewart in 1922. This machine was able to generate single vowel sounds, but not any consonants or utterances [9].

In the 1930s, Bell Laboratories developed the VOCODER (Voice Coder), a keyboard-operated electronic speech analyzer and synthesizer [8]. The speech quality and intelligibility were far from good but the potential for producing artificial speech were well demonstrated [9]. Homer Dudley refined this device into the VODER (Voice Operating Demonstrator), which was introduced at the 1939 New York World's Fair [8].

After the demonstration of the VODER, the scientific world became more and more interested in speech synthesis. It was finally shown that intelligible speech can be produced artificially [9].

The first computer-based speech synthesis systems were created in the late 1950s. In 1961, physicist John Larry Kelly used an IBM 704 computer to synthesize speech, which has become an event among the most prominent in the history of Bell Labs. Indeed, Kelly's voice synthesizer recreated the song "Daisy Bell". Coincidentally, Arthur C. Clarke was visiting Bell Labs facility at that moment, and he was so impressed by the demonstration that he used it in the climactic scene of his screenplay for his novel 2001: A Space Odyssey. In one of its scenes, the HAL 9000 computer sings the same song while it is being put to sleep by astronaut Dave Bowman. [8]

When it comes to the first complete text-to-speech system for English, this was developed by Noriko Umeda in the Electrotehnical Laboratory in Japan in 1968. The speech was quite intelligible but monotonous and far away from the quality of present systems [9]. A decade later, in 1979, the MITalk (text-to-speech system developed at M.I.T by Allen, Hunnicutt and Klatt) was demonstrated, which the technology used in it, forms the basis for many synthesis systems today [9].

Meanwhile, in 1976, Kurzweil introduced the first reading aid with optical scanner [8]. These machines for the blind were capable to read quite well the written text. However, the system was far too expensive for average customers, but were used in libraries and service centers for visually impaired people [9].

In late 1970's and early 1980's, considerably amount of commercial text-to-speech and speech synthesis products were introduced and many computer operating systems have included speech synthesizers since then [9]. Some examples can be seen in the following section.

2.1.2 Computer operating systems with speech synthesis

Apple

The first speech system integrated into an operating system was Apple Computer's MacInTalk in 1984, presented during the introduction of the Macintosh in which the computer announced itself to the world. In 1990, Apple hired many researchers in the field of speech recognition and invested a lot of money in it. During most of the early 1990s, Apple voices were synthetic; however, more recently, Apple has added sample-based voices. [8]

AmigaOS

The second operating system with advanced speech synthesis capabilities was AmigaOS, introduced in 1985. It featured a complete system of voice emulation, with both male and female voices and was licensed by Commodore International from a third-party software house. The user could redirect console output since AmigaOS considered speech synthesis a virtual hardware device. [8]

Microsoft Windows

Speech systems were first available on Microsoft-based operating systems like Windows 95 and Windows 98. Windows XP featured a speech synthesis program called Narrator, a text-to-speech utility for people with visual handicaps directly available to users. Some third-party programs can perform various text-to-speech tasks, such as, reading text aloud from a website, document or email account. Microsoft Speech Server, for instance, is a complete package for voice synthesis and recognition for commercial applications such as call centers. [8]

GNU/Linux

Several systems operate on GNU/Linux and include open-source programs such as, *gnuspeech*, which uses articulatory synthesis from the Free Software Foundation, or the Festival Speech Synthesis System, which uses diphone-based synthesis. [8]

2.1.3 Architecture

A Text-To-Speech system is composed of a front-end and a back-end, which receive text as input and convert it into speech output.

- **Front-end**: it has two major tasks.
    1. It converts raw text containing symbols (e.g., numbers and abbreviations) into the equivalent of written-out words. This process is often called text normalization, or tokenization. [8]
    2. It assigns phonetic transcriptions to each word (text-to-phoneme), and divides and marks the text into prosodic units, like phrases and sentences. Phonetic transcriptions and prosody information together make up the symbolic linguistic representation that is output by the front-end. [8]
- **Back-end**: converts the symbolic linguistic representation into sound. The back-end is often referred to as the synthesizer. [8]

The following figure (extracted from [8]) represents the architecture of a Text-To-Speech system.

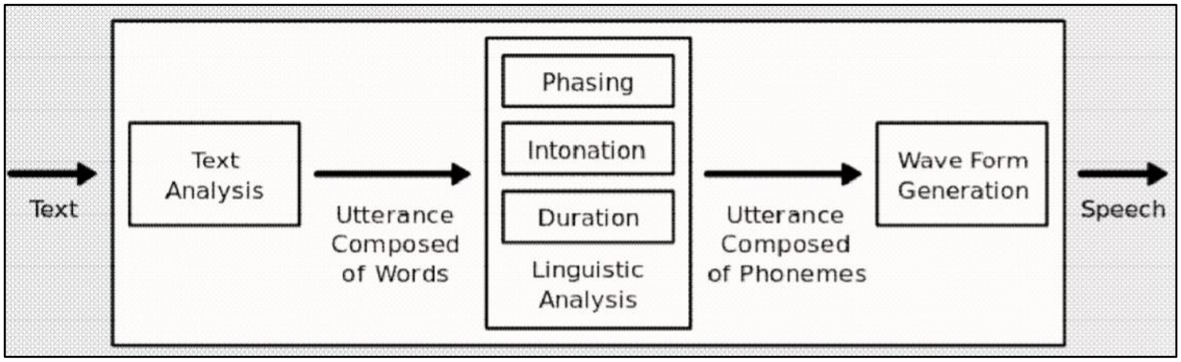


Figure 1: Speech Synthesis architecture

2.2 Voice Cloning – Audio Deepfake

One of the limitations of speech synthesis and voice assistants is that they normally tend to sound very unnatural and robotic [10]. But, as time has progressed, the advancements in the field of artificial intelligence and machine learning have led to what we call voice cloning.

Voice cloning aims to generate synthetic voices very similar to an original voice. Based on deep-learning techniques, this technology takes advantage of a set of audios of the original voice in order to train a model capable of generating new audios that sound alike [11]. Basically, voice cloning consists of capturing the voice of a speaker to perform text-to speech on arbitrary inputs [12].

Nowadays, due to its interesting and varied applications, voice cloning is being increasingly demanded by the market [11]. Indeed, this technology helps automate and personalize many tasks carried out in several types of applications and domains, using specific or favorite voices to develop customized and fully personalized conversational assistants [11].

While voice cloning technology has developed, audio deepfakes have come hand in hand with it, becoming each time more and more popular. But what is an audio deepfake?

The term “deepfake,” which has its origin on a Reddit thread in 2017, is used to describe the recreation of a human’s appearance or voice through artificial intelligence [4]. So, audio deepfakes might mean you can no longer trust your ears. An audio deepfake is when a “cloned” voice that is potentially indistinguishable from the real person’s is used to produce synthetic audio [4]. Indeed, a study shows that people usually only guess if an audio deepfake is real or fake with about 57% accuracy [4].

2.2.1 Applications

Currently, there is an enormous demand for voice cloning due to the benefits that this technology can offer in several applications.

- **Dubbing**. For actors, it will become easier for them to dub their movies in different languages [10]. Being able to go back and change the dialogue in a video or movie without the need for a reshoot and creating entire videos just by selecting from a menu of presenters and entering the script [13].
- **Healthcare**. People who have lost their voice can use this technology to communicate by voice replacement. Stephen Hawking, for instance, used a robotic synthesized voice after losing his own in 1985. In 2008, synthetic voice company, CereProc, gave late film critic, Roger Ebert, his voice back after cancer took it away by processing a large library of voice recordings. [4]
- **eLearning**. As a result of the nation-wide lockdown to contain the COVID-19 outbreak, online learning is stepping up among students. Providers of virtual classes and informative video content can significantly benefit from voice cloning to produce interactive content with minimum operational costs. It can significantly transform the way teachers impart knowledge to students in the form of professionally recorded lectures, complex topics, and other educational materials. [1]
- **Customer Experience**. The technology would open new opportunities for a range of businesses by being able to personalize voice-controlled interactions to enhance customer experience. For example, adding a familiar voice to healthcare services for comforting the patients or boosting customer engagement with audible product descriptions using famous voices. [1]
- **Gaming**. In the past, speech was the one component in a game that was impossible to create on-demand. Now, though, studios have the potential to clone an actor’s voice and use text-to-speech engines so characters can say anything in real time. [4]
- **Advertising**. Last year, a UK-based charity used deepfake technology to create a video of David Beckham delivering an anti-malaria message in nine languages. In addition, an advertising company created corporate training videos that used artificial intelligence to create a presenter that could speak the recipient’s language and address them by name. [13]
- **Professional training**. Voice cloning technology can be used to create AI avatars for use in training videos. Indeed, exercising with recorded videos has been getting more attention from the corporate world during the COVID-19 pandemic. [13]

2.2.2 How it works

Once understood the concept of voice cloning, it is also useful to explain how it works. Thanks to artificial intelligence —specifically, deep-learning algorithms— it has been possible to match recorded speech to text to understand the component phonemes that make up someone's voice. Then, the resulting linguistic building blocks are used to approximate words that have not been spoken/heard. [11]

Not long ago, developers needed enormous quantities of recorded voice data to get passable results. Then, a few years ago, scientists developed generative adversarial networks (GANs), which could, for the first time, extrapolate and make predictions based on existing data. One of the biggest advancements in voice cloning has been the reduction in how much raw data is needed to clone a voice. In the past, systems needed dozens or even hundreds of hours of audio. Now, however, competent voices can be generated from just minutes of content. [11]

In order for a computer to be able to read out-loud with any voice, it needs to somehow understand two things: what it's reading and how it reads it. Thus, the voice cloning system needs to have two inputs: the text we want to be read and a sample of the voice which we want to read the text. [7]

From a technical view, the system architecture is shown in Figure 2 (extracted from [7]), which is broken down into the following [7]:

1. Given a small audio sample of the voice we wish to use, encode the voice waveform into a fixed dimensional vector representation.

2. Given a piece of text, also encode it into a vector representation.

3. Combine the two vectors of speech and text.

4. Decode them into a spectrogram.

5. Use a Vocoder to transform the spectrogram into an audio waveform that we can listen to.

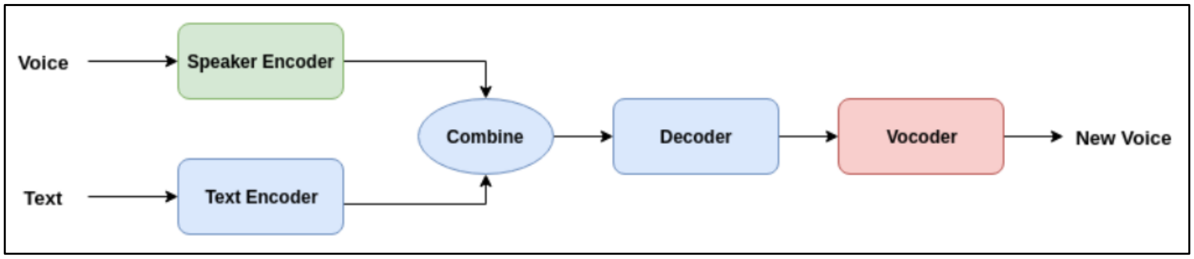


Figure 2: Voice Cloning technology architecture

The quality in terms of naturalness and similarity of the final output will depend on the input audio sample, which is sometimes of low-quality or has been recorded in noisy locations. The worse the sound quality, the harder it is to get a natural cloned voice which is similar to the original one.

2.3 Technologies Nowadays

There are several applications and technology examples that use Speech Synthesis and/or Voice Cloning nowadays.

2.3.1 Voice Assistants: Siri, Cortana, Alexa and much more

It all began with what it is called the Origin period. IBM became the first to introduce a voice assistant with its Shoebox device in 1961. While very primitive, it did understand 16 words and 9 digits. This laid the groundwork for the pre-modern era when voice and virtual assistants became available to consumers for the first time. Microsoft's text-based virtual assistant, Clippy, showed how natural language in text could be tracked, interpreted, and used as a basis for interactive feedback. While Clippy was never popular, it did teach some lessons about what not to do when it comes to building virtual assistants. Namely, that virtual assistants should only be heard when explicitly called for. [14]

Then, in the Modern Era of voice assistants, smartphones and voice interaction collided. Siri was the first voice assistant to reach a wide audience and others, like Google Now and Microsoft's Cortana, soon followed. In 2014, Amazon introduced the Alexa voice assistant and Echo smart speaker. [14]

2.3.2 Alexa Blackboard Skill

This new integration is the first skill officially released by Blackboard. The Blackboard Alexa skill allows Blackboard Learn users to request and receive information about homework and assignments in the Blackboard Learn application via an Alexa-device, such as Amazon's Echo, rather than having to log into Blackboard Learn to look the information up. [15]

The user's voice command will be translated by Amazon into a machine-digestible command that is sent to Blackboard. Blackboard returns the requested information securely to Amazon which in turn converts the response into an audio response that the Alexa-device will securely provide to the Blackboard user. [15]

This Blackboard Alexa Skill only handles what homework and assignments are due in a user's classes and it uses Alexa's voice to reproduce the answer [15]. Therefore, an added feature would be to be able to request for more information and having the professor's voice cloned, for example.

2.3.3 Cerence's Voice Clone Technology

Cerence Inc. introduced My Car, My Voice, in December 2019, a revolutionary product that lets people create custom voices for their in-car assistants [16].

Cerence’s voice clone technology is a game-changing innovation for the world of in-car voice assistants, which typically come with a set of pre-determined voice options. Now, with this new Cerence innovation, people can quickly and easily create a carbon copy of their own voice or that of a family member or friend to be the persona of the voice assistant in their cars that can be used to give directions, read messages, and provide updates. Not only does this create a more human-like experience in the car, but also enhances safety – when the car is delivering notifications or information, the voice of a loved one can generate a more attentive or urgent response from a driver than a generic voice. [16]

So, after analyzing the context about Speech Synthesis and Voice Cloning, it is clear there is a huge market related to these technologies and that several new applications have developed making use of these advancements. Nevertheless, there is still a chance of improvement, what we are looking to achieve with the ongoing project.

3. Goal and Benefits

3.1 Project Goal

The main objective of the project is to build a voice query assistant that allows students to request and receive information that appears in the syllabus of a course, such as, professor's contact information, assignment dates, grading scheme, the schedule… instead of having to log into the learning management system (e.g., Blackboard) and open the syllabus or ask the professor during a lecture.

First, the student's voice question is captured by a microphone and translated into text using Speech-To-Text technology. The answer to the question is obtained by performing a query in a database in which the details about the student's courses are saved. Then, the required information is converted into an audio response using Text-To-Speech technology and Voice Cloning. This second feature allows to reproduce the answer with the professor's voice instead of a synthetic/robotic voice. In that way, the interaction becomes more authentic and trustworthy, creating an enhanced experience that mimics a natural human-to-human interaction.

Figure 3 shows the architecture of the tool that we aim to build:

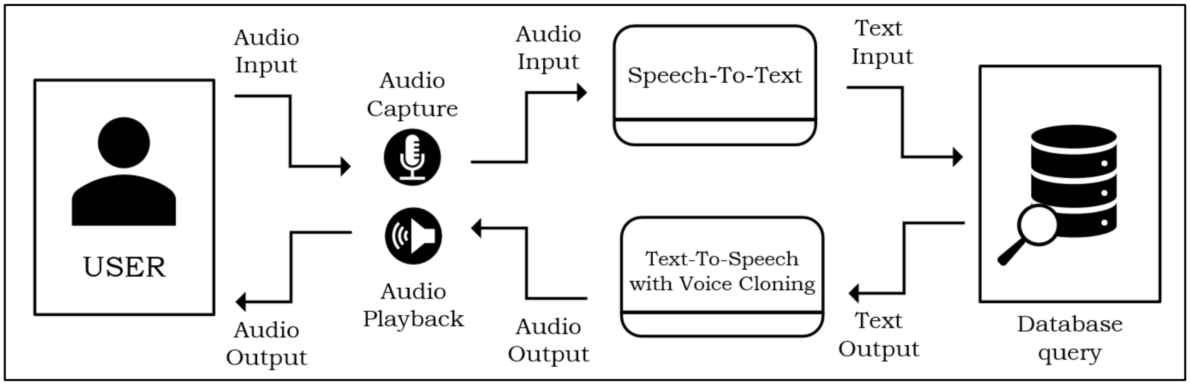


Figure 3: Voice query assistant architecture

Moreover, a qualitative research study will be designed in order to complete the research about the emerging technology of Speech Synthesis and Voice Cloning, where several types of questions will be included and the reaction of college students and professors to a video demo of the developed tool will be analyzed.

Finally, a discussion and actual point of view about Voice Cloning and Deepfakes will be assessed where technology misuse, safeguard and mitigation strategies and regulation will be studied, so as to analyze the impact for good and evil of the mentioned technology.

3.2 Project Benefits

This work can bring several benefits in different areas, the most significant being social, scientific, technical, and economic benefits.

3.2.1 Social Benefits

The project offers various social benefits. When it comes to the education focused voice query assistant that we intend to build:

- Efficiency for the student; they do not have to log into the learning management platform (e.g., Blackboard), look in the syllabus or ask the professor during a lecture.
- Efficiency for the professor; a lot of times, students ask the same questions even from one week to another within a term. So, having a virtual assistant which is capable to answer this type of questions automatically, helps reduce the workload of teachers, without needing to repeat themselves multiple times. That way, there would be more time in the lectures to engage in deeper discussions with the students.
- Some students can be afraid to ask questions in front of the other students. This allows students to ask questions without fearing judgment for the quantity or content of their inquiries.
- The possibility to get immediate feedback on your questions at any time of the day. If a student needs to ask their teacher a question in the middle of the night, the virtual assistant can answer them within seconds. This is a vast improvement over the traditional way of waiting to talk to the teacher after class or waiting to receive an answer by email.
- Thanks to the Voice Cloning technology, the answer is reproduced using the professor's voice, which makes the interaction more authentic and less robotic, creating a more natural and enhanced human-to-human experience.

Regarding the subsequent design of the qualitative study:

- It provides a potential understanding of people's attitude and opinion about the topic, which can be beneficial towards developing future tools and systems. This will be relevant for the development of a marketable product based on our tool in the future and to draw more solid conclusions about the impact and needed regulation of voice cloning in the social landscape.
- It allows to collect genuine ideas from different socioeconomic demographics, which can be applied while designing future frameworks. Indeed, as qualitative research processes are usually open-ended, there is no "right" or "wrong" answer, which makes data collection much easier.

3.2.2 Scientific Benefits

As far as scientific benefits are concerned, the main benefit is the research about Speech Synthesis and Voice Cloning technology and its use for good (in this case in education) and evil. The results obtained in this project can help to carry out future research about the mentioned technology and its usage. In addition, the conclusions collected from the qualitative study can also be useful to get to know the opinion of the society and to develop other suitable projects based on that.

Illinois Institute of Technology carries out various technology related research projects, which are focused on different topics; therefore, this project can contribute to the field of Speech Synthesis technology research and enrich the knowledge of the IIT community.

3.2.3 Technical Benefits

The main technical benefits of this work are the materials, resources and tools used throughout the whole process.

First, the developed Python code during the project may be useful for other similar projects in the future. Moreover, the syllabus template created to test the overall process can also be reused in another research work. In fact, this document can be modified and redesigned for an office environment document for example.

Finally, the research study can also be useful in the following years in order to examine the evolution of what people think about Speech Synthesis and Voice Cloning and compare the results.

3.2.4 Economic Benefits

In terms of economic benefits, this project is not mainly focused on having an economic income at the moment; it emphasizes mostly on the benefits described above. Nevertheless, if the research goes forward and the results continue being favorable in the future, it will be a major step in the research field and will perhaps bring economic benefits in the long run.

4. Specifications and Requirements

4.1 Planning

4.1.1 Workgroup

The present project has been developed by Naroa Amezaga Vélez, who is an international student that has graduated in Telecommunication Engineering in 2019 from the University of The Basque Country in Bilbao (Spain), and is currently pursuing a Master of Information Technology and Management at Illinois Institute of Technology.

The project supervision, advising and correction has been carried out by Jeremy Hajek, director of the Smart Tech: Embedded Systems Lab and Industry Associate Professor of Information Technology and Management at Illinois Institute of Technology.

Table 1 shows a summary of the workgroup of the project:

FULL NAME	**OCCUPATION**	**ROLE IN THE PROJECT**
Jeremy Hajek	Director of Smart Tech: Embedded Systems Lab at Illinois Institute of Technology	Project supervision, advising, and correction
Naroa Amezaga Velez	Master of Information Technology & Management student	Project development

Table 1: Project workgroup

4.1.2 Milestones

A milestone is a reference point that marks a significant event or a branching decision point within a project. The current project consists of nine milestones which can be seen on the coming table along with the date that each of them was completed:

MILESTONE	**DESCRIPTION**	**DATE**
M1	Start of the project	01/15/2021
M2	Project definition	01/30/2021
M3	STT mechanism developed	03/27/2021
M4	Syllabus template and database created	02/22/2021
M5	Database query performed	04/30/2021
M6	TTS and Voice Cloning mechanism developed	05/26/2021
M7	Qualitative research design carried out	07/01/2021
M8	Documentation finished	07/16/2021
M9	Project delivered	07/28/2021

Table 2: Project milestones

4.2 Risk Analysis

The purpose of this section is to examine the risks that may arise during the work. In fact, as with any project, there are some risks that could lead to its delay or cancellation. Since the project is completed, it can be said that the risks have been avoided or mitigated, but for that it was necessary to anticipate these first.

When assessing risks, one must take into account, on the one hand, the probability of a risk occurring and, on the other hand, the effect or impact it may have on the project. The following points describe each risk and outline contingency measures to avoid them. In addition, a summary of the risk analysis is provided at the end of the section.

4.2.1 Data loss

Although the probability of the risk occurring is quite low, it would have a significant impact on the project. In fact, it is dangerous to lose data (database, developed code and tools…) during work, either due to a computer breakdown, electrical outrage or any other factor.

To avoid this, it is enough to store the useful data in different locations instead of in just one place, making copies on a regular basis. Moreover, the most important information is shared with the project director so that the information can be stored in multiple devices.

4.2.2 Delays

The probability of delays during the development of a project can be medium, since due to several reasons, the proposed objectives may not be met at the right time. However, the impact of this is quite low. In fact, in this case, the working team is small, and usually a delay will only affect oneself, so he/she can make up for the time.

In order to avoid this risk, factors that could lead to delays should be anticipated and the time required to develop each section of the project should be accurately measured.

4.2.3 Issues with developed tools

There may be some issues with the tools that are developed during the process, such as a code not working as intended, having syntactic errors, or not properly fulfilling the goal to be accomplished. Although the probability of occurrence of this type of risk is medium, its effect would be large.

As a precautionary measure, each tool should be checked one by one and little by little in order to check its correct functionality and verify that all requirements are met. This way, it will be possible to correct the errors before spreading them out.

4.2.4 Leaving of a team member

As with the rest of the projects, it should be borne in mind that all of a sudden, a teammate may take leave due to certain reasons (illness, accident, maternity…), not being able to carry out the work. This risk would have a quite high impact. Also, adding another member to the project would cause a huge delay. However, the probability of happening is low.

With regard to contingency measures, it would be advisable to think of someone that could take the role of research project director, since it is the only one involved apart from the student. In case something happens, it would be very important to adapt to a new project planning as quickly and efficiently as possible.

4.2.5 Inefficient processing

Several problems can lead if the processing of the developed tools is not efficient, if the database is too large and burdensome, or if the programming is not done in an optimized way, for example. This can generate considerable delays during the process. The probability of occurrence is medium, and the impact is quite low, as it is possible not to waste too much time by leaving the code running while doing something else.

To avoid this, the development of tools should be programmed as efficiently as possible. When processing requires time, the project director can be asked for help or more capable equipment can be used.

4.2.6 Risk analysis summary

In order to summarize the risk analysis, all risks are compiled in a table, as well as the probability of happening of each of them and the impact they would have if they occurred. In addition, Figure 4 shows a matrix presenting the relationship between the probability and impact of different types of risks described, where different levels of probability can be seen on the horizontal axis and different levels of impact on the vertical one.

RISK ANALYSIS		
RISK	**PROBABILITY**	**IMPACT**
R1	Quite low	High
R2	Medium	Quite low
R3	Medium	High
R4	Low	Quite high
R5	Medium	Quite low

Table 3: Risk analysis summary

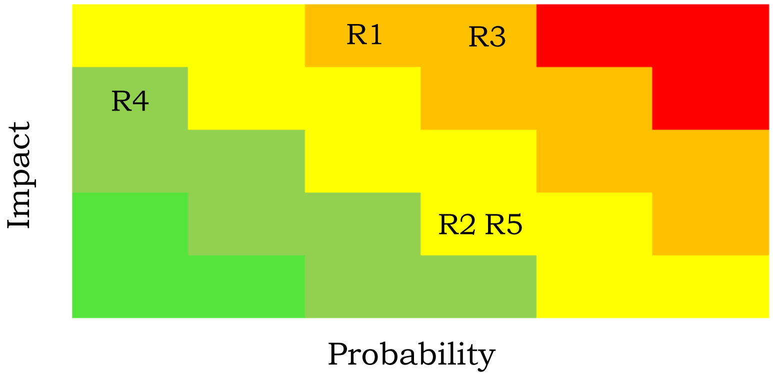


Figure 4: Risk probability/impact matrix

4.3 Alternative Analysis and Selection

In this section, an alternative analysis and subsequent selection of different tools and systems will be described. In the following section a summary of the selected equipment and tools will be outlined.

4.3.1 Programming language

When it comes to implementing software developments, different programming languages can be found, each of which, depending on the purpose to be achieved and the tools to be used, has its advantages and disadvantages.

Nowadays, object-oriented programming has become the main tool to develop any type of implementation. Therefore, it has been decided to carry out the project using this type of programming. The main object-oriented programming languages are C++, Java and Python. GUI (Graphic User Interface) implementations can be done with any of the three, and models for implementing communication between computers and databases are highly developed. Nevertheless, each of them has its own advantages and disadvantages [17]. The following table shows a comparison between C++, Java and Python:

TOPIC	**C++**	**JAVA**	**PYTHON**
Memory Allocation	Manual allocation Manual deallocation	Manual allocation Automatic garbage collection	Automatic allocation Automatic deallocation
Code Length	Longer lines of code as compared to Python	About 40% more code in comparison to C++	3-5 times shorter than Java programs
Typing System	Statically typed, both code and compiler	Statically typed, both code and compiler	Dynamically typed, interpreter is strongly typed
Compiling	Compiled to machine code	Compiled to byte code, then interpreted by the JVM	Interpreted by the Python interpreter
Execution Speed	Faster than Java and Python	Slower than C++ and faster than Python	Slower than C++ and Java
Portability	Platform dependent	Platform independent	Platform independent
Syntax Complexity	Defining blocks using {, ending statements using ;	Defining blocks using {, ending statements using ;	Using indentation only

Table 4: Programming language comparison

After analyzing all the features, we have chosen **Python** mainly due to its easiness to learn, short code length and platform independency.

4.3.2 Speech-To-Text

When it comes to Speech-To-Text (STT) systems, these are quite well stablished at this time, as they have incredibly progressed over the past years [18]. Therefore, several private and open-source options can be found, from which we chose to use a library for Python called *SpeechRecognition* due to its flexibility and ease of use. It performs speech recognition with support for several engines and APIs, online and offline [19]. In order to use the microphone, *PyAudio* is required, which provides Python bindings for *PortAudio*, the cross-platform audio I/O library [20]. The headset that has been used during the project is Logitech H390 USB Computer Headset. As mentioned, *SpeechRecognition* library supports several APIs, such as, Microsoft Bing Speech, Google Speech Recognition, IBM Speech to Text or CMU Sphinx.

To begin to understand the technology and its usefulness, we have selected two APIs, one that works just online (Google Speech Recognition) and another one that can work offline (CMU Sphinx) and performed a comparison so as to pick one for our case study. For that, fifteen sentences, which appear in section 5.1, were recorded from the microphone. Once the audio was converted into text, the Word Accuracy (WA) and the Word Error Rate (WER) were calculated by comparing the obtained result with the original text [21]. These were calculated according to the following equations [21]:

$$WA = (N - D - S) / N$$

$$WER = (I + D + S) / N$$

where N represents the number of words in original the sentence, I the number of words inserted, D the number of words deleted and S the number of words substituted.

The following tables show the results for each of the options and a summary of them. CW is the number of correct words and EW the number of error words.

SENTENCE	**GOOGLE SPEECH RECOGNITION**							
	N	**I**	**D**	**S**	**CW**	**EW**	**WA**	**WER**
1	7	0	0	0	7	0	100,00%	0,00%
2	8	0	0	0	8	0	100,00%	0,00%
3	8	0	0	1	7	1	87,50%	12,50%
4	8	0	0	0	8	0	100,00%	0,00%
5	8	0	0	0	8	0	100,00%	0,00%
6	8	0	0	0	8	0	100,00%	0,00%
7	8	0	0	1	7	1	87,50%	12,50%
8	8	0	0	1	7	1	87,50%	12,50%
9	7	0	0	0	7	0	100,00%	0,00%
10	7	0	0	0	7	0	100,00%	0,00%
11	9	0	0	1	8	1	88,89%	11,11%
12	9	1	0	2	7	2	77,78%	33,33%
13	10	1	0	2	9	2	80,00%	30,00%
14	10	0	0	1	9	1	90,00%	10,00%
15	11	0	0	0	11	0	100,00%	0,00%

Table 5: Results for Google Speech Recognition

SENTENCE	**CMU SPHINX**							
	N	**I**	**D**	**S**	**CW**	**EW**	**WA**	**WER**
1	7	1	0	2	5	2	71,43%	42,86%
2	8	1	0	3	5	3	62,50%	50,00%
3	8	0	0	3	5	3	62,50%	37,50%
4	8	0	0	3	5	3	62,50%	37,50%
5	8	1	0	4	4	4	50,00%	62,50%
6	8	0	0	5	3	5	37,50%	62,50%
7	8	0	0	3	5	3	62,50%	37,50%
8	8	1	0	4	4	4	50,00%	62,50%
9	7	0	0	4	3	4	42,86%	57,14%
10	7	0	0	4	3	4	42,86%	57,14%
11	9	1	1	3	6	3	55,56%	55,56%
12	9	0	0	6	3	6	33,33%	66,67%
13	10	1	0	6	4	6	40,00%	70,00%
14	10	2	0	3	7	3	70,00%	50,00%
15	11	1	0	4	7	4	63,64%	45,45%

Table 6: Results for CMU Sphinx

	GOOGLE SPEECH RECOGNITION		**CMU SPHINX**	
	WA	**WER**	**WA**	**WER**
Mean	93.28%	8.13%	53.81%	52.99%

Table 7: Result summary

So, as we can see on Table 7, we obtained 8.13% WER for Google Speech Recognition and 52.99% WER for CMU Sphinx.

Therefore, **Google Speech Recognition** was chosen to carry out the project, since it can be stated that the acoustic modeling and language model of Google Speech Recognition is superior even though it needs to work online.

4.3.3 Database Management System

A Database Management System is a software that communicates with the database itself, applications, and user interfaces to obtain data and parse it. There are two types:

- **Relational**. This type is also called SQL since a Structured Query Language is their core. The data appears as tables of rows and columns with clear dependencies and strict structure [22].
- **Non-Relational**. This type is also called NoSQL as instead of being limited to a table structure, they are considered document-oriented where non-structured data such as articles, photos, videos, and others are collected [22].

Due to their simplicity, ease of use, scalability and query capability relational database management systems have become the single most popular approach nowadays [22]. So, we will focus our selection on that type of system, taking into account three of them: SQLite, MySQL and PostgreSQL.

- **SQLite.** It is file-based, self-contained and fully open-source, known for its portability, reliability, and strong performance even in low-memory environments [23]. One of its advantages is that SQLite library is very lightweight and also user-friendly since it does not run as a server process [24]. In addition, it is great for developing and testing, as it offers an in-memory mode which can be used to run tests quickly without the overhead of actual database operations [23].
- **MySQL.** It is the most popular one of all the large-scale database servers [23] since it is a feature-rich product that powers many of the world's largest websites and applications, including Twitter, Facebook, Netflix, and Spotify [24]. Its installation is easy, and it provides a lot of security features, apart from being scalable, really powerful and speedy [23].
- **PostgreSQL.** It is known for being the most advanced open-source and community-driven relational database in the world [23]. Its main goal is to be highly extensible and standards compliant [24]. It is the best choice when reliability and data integrity are an absolute necessity or when the database is required to perform complex, custom procedures [23].

After analyzing the three database management systems, we chose to use **SQLite** due to its simplicity and greatness for testing. Indeed, for the development of the current research project, there is no need of outstanding security features, large databases or complex operations.

4.3.4 Text-To-Speech with Voice Cloning

Since Text-To-Speech (TTS) systems have applications in a wide range of scenarios, huge advancements have been done in that field. In recent years, voice cloning has been introduced for enhanced experiences in the framework of deep learning applied to text-to-speech technology.

Nowadays, we can find several open-source software options for TTS, but most of them sound robotic and lack of naturalness. Therefore, we decided to implement Voice Cloning, which tries to clone someone's voice synthetically. In regard to this technology, there are various options already available:

- Descript Overdub [1]
- Resemble AI [2]
- Lovo [3]
- ReSpeecher [4]

Nevertheless, all of the mentioned software are not open-source at the current time of the writing of this paper and their pricing is out of our budget. Hence, we have decided to use **Speaker Verification to Text to Speech Synthesis (SV2TTS)** [7], a three-stage open-source implementation [25] that allows to clone a voice from only a few seconds of reference speech. It is the first open-source implementation of a framework based on a recent research study [26] that shared remarkably natural-sounding outcomes but provided no implementation.

The three stages of the framework are as follows:

1. A speaker encoder that derives an embedding from the source audio of the speaker to be cloned. The embedding is a meaningful representation of the voice of a speaker, such that similar voices are close in latent space. This model is based on the GE2E loss described in [27].
2. A synthesizer that, depending on the embedding of the speaker, generates a spectrogram from the input text. This second model is based on Tacotron 2 [28] without WaveNet.
3. A vocoder that infers an audio waveform from the spectrogram generated by the synthesizer. This final model is based on WaveNet [29].

More information about SV2TTS will be given in section 5.4.

[1] For more information: https://www.descript.com/overdub

[2] For more information: https://www.resemble.ai/cloned

[3] For more information: https://www.lovo.ai/custom

[4] For more information: https://www.respeecher.com/product

4.4 Summary of Equipment and Tools Used

EQUIPMENT / TOOL	**SELECTION**
Headset with microphone	Logitech H390 USB Computer Headset
Programming language	Python 3.7
Speech-To-Text	SpeechRecognition (Google Speech Recognition)
Database Management System	SQLite
Text-To-Speak with Voice Cloning	SV2TTS

Table 8: Summary of used equipment and tools

5. Methodology

This section describes the process followed to complete the work. To this end, a Python tool has been built, and the details of which will be explained next. The aim of this, is to simulate a voice query assistant that is able to answer using the proper professor's voice to fifteen questions about course information. When executed, the user starts by asking one of the fifteen possible questions about a specific course. The microphone captures the question and converts it into text.

Once the conversion is done, an SQL query is performed to find the appropriate answer to the question. All the information about the courses is saved in a database.

Finally, the answer, which is in plain text, is converted into spoken language. In addition, the answer is reproduced by a cloned voice of the professor of that course. To test the tool, information about four courses has been used, and the voices of their corresponding professors: two female and two male voices that belong to four current or former professors from Illinois Institute of Technology who have given their consent.

Figure 5 shows the same architecture of section 3.1.1. but with the specific technologies that have been selected after the alternative analysis in section 4.3:

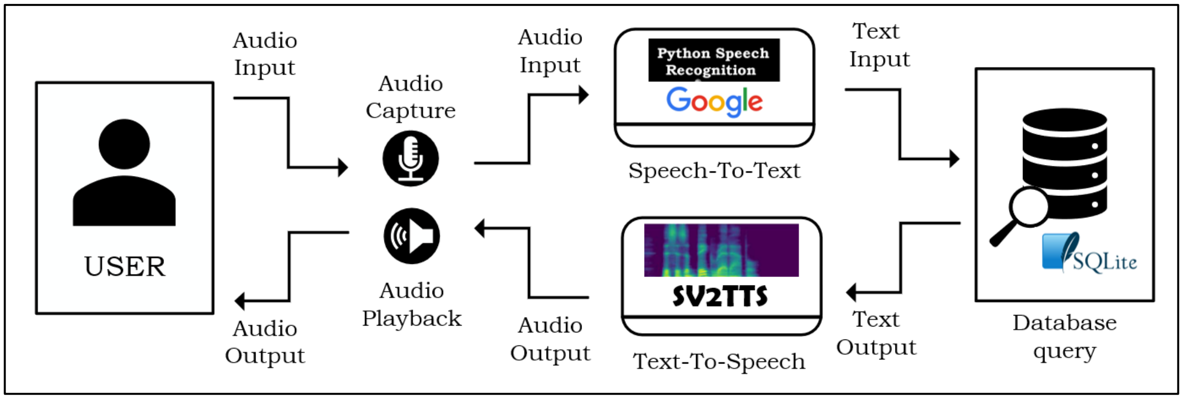


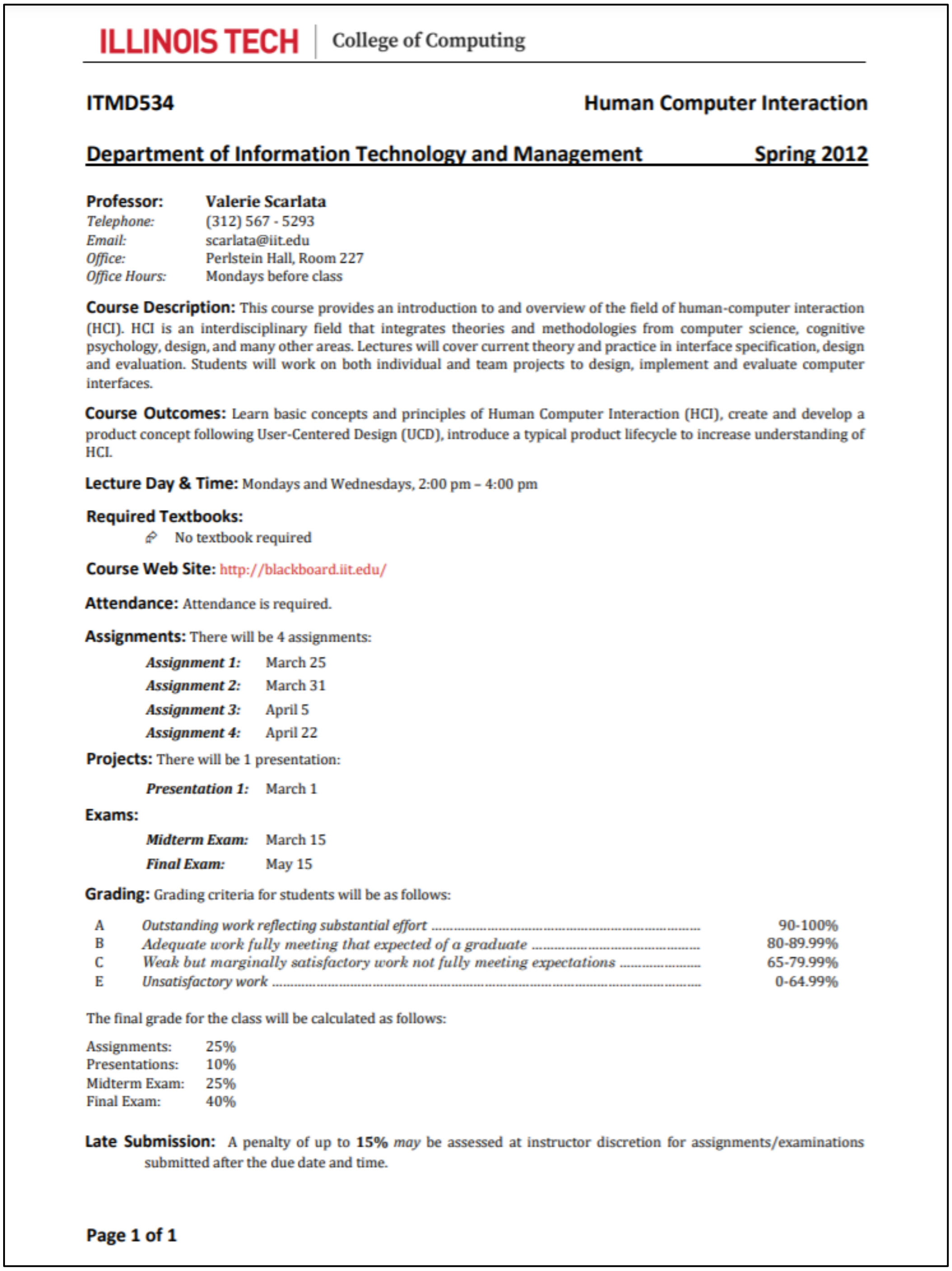
Figure 5: Voice query assistant architecture

So, the methodology of the project has been divided into five steps, which will be explained in detail on the next pages:

1. Database Definition
2. Speech-To-Text Implementation
3. Database Query
4. Text-To-Speech with Voice Cloning Implementation
5. Qualitative Research Study Design

5.1. Database Definition

The first step is to define the database that will be used during the process. To simulate a real educational environment, a syllabus template has been created, which will allow to save the information in the database. An example can be seen in the following figure:



ILLINOIS TECH | College of Computing

ITMD534 **Human Computer Interaction**

Department of Information Technology and Management **Spring 2012**

Professor:	**Valerie Scarlata**
Telephone:	(312) 567 - 5293
Email:	scarlata@iit.edu
Office:	Perlstein Hall, Room 227
Office Hours:	Mondays before class

Course Description: This course provides an introduction to and overview of the field of human-computer interaction (HCI). HCI is an interdisciplinary field that integrates theories and methodologies from computer science, cognitive psychology, design, and many other areas. Lectures will cover current theory and practice in interface specification, design and evaluation. Students will work on both individual and team projects to design, implement and evaluate computer interfaces.

Course Outcomes: Learn basic concepts and principles of Human Computer Interaction (HCI), create and develop a product concept following User-Centered Design (UCD), introduce a typical product lifecycle to increase understanding of HCI.

Lecture Day & Time: Mondays and Wednesdays, 2:00 pm – 4:00 pm

Required Textbooks:

- No textbook required

Course Web Site: http://blackboard.iit.edu/

Attendance: Attendance is required.

Assignments: There will be 4 assignments:

Assignment 1:	March 25
Assignment 2:	March 31
Assignment 3:	April 5
Assignment 4:	April 22

Projects: There will be 1 presentation:

Presentation 1:	March 1

Exams:

Midterm Exam:	March 15
Final Exam:	May 15

Grading: Grading criteria for students will be as follows:

A	*Outstanding work reflecting substantial effort*	90-100%
B	*Adequate work fully meeting that expected of a graduate*	80-89.99%
C	*Weak but marginally satisfactory work not fully meeting expectations*	65-79.99%
E	*Unsatisfactory work*	0-64.99%

The final grade for the class will be calculated as follows:

Assignments:	25%
Presentations:	10%
Midterm Exam:	25%
Final Exam:	40%

Late Submission: A penalty of up to **15%** *may* be assessed at instructor discretion for assignments/examinations submitted after the due date and time.

Page 1 of 1

Figure 6: Syllabus example

The information appearing in each syllabus is then passed to a common CSV type file with an automatic script in order to gather the information from all the syllabus together and define the database file. So as to know which information is needed, it is necessary to define the possible question bank that the user can ask. This consists of the following fifteen questions:

1. **Which is *course_name* professor's telephone?**
   Example: Which is Human-Computer Interaction professor's telephone?
2. **Which is *course_name* professor's office address?**
   Example: Which is Human-Computer Interaction professor's office address?
3. **What are *course_name* professor's office hours?**
   Example: What are Human-Computer Interaction professor's office hours?
4. **Which is the schedule for *course_name*?**
   Example: Which is the schedule for Human-Computer Interaction?
5. **Which is the textbook for *course_name*?**
   Example: Which is the textbook for Human-Computer Interaction?
6. **Is the attendance required in *course_name*?**
   Example: Is the attendance required in Human-Computer Interaction?
7. **When is *course_name number* assignment due?**
   Example: When is Human-Computer Interaction first assignment due?
8. **When is *course_name number* presentation due?**
   Example: When is Human-Computer Interaction first presentation due?
9. **When is *course_name* midterm exam?**
   Example: When is Human-Computer Interaction midterm exam?
10. **When is *course_name* final exam?**
    Example: When is Human-Computer Interaction final exam?
11. **What is the assignment in *course_name*?**
    Example: What is the assignment grading in Human-Computer Interaction?
12. **What is the presentation grading in *course_name*?**
    Example: What is the presentation grading in Human-Computer Interaction?
13. **What is the midterm exam grading in *course_name*?**
    Example: What is the midterm exam grading in Human-Computer Interaction?
14. **What is the final exam grading in *course_name*?**
    Example: What is the final exam grading in Human-Computer Interaction?
15. **What is the penalty for late submission in *course_name*?**
    Example: What is the penalty for late submission in Human-Computer Interaction?

So, as you can see in the examples, *course_name* needs to be substituted by the name of a course that is saved in the database and *number* by an ordinal number (e.g., first, second, third…) corresponding to the assignment or presentation.

Once the questions are defined, the required information can be extracted from the syllabus and stored in the CSV file.

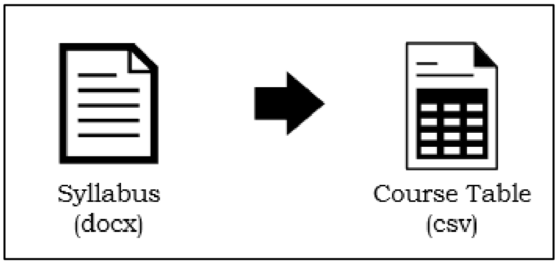


Figure 7: Syllabus to CSV file conversion

Each row of this CSV file will represent a course and each of those will include the following information organized in columns:

COLUMN NAMES
course_code
course_name
semester
year
first_name
last_name
telephone
email
office_address
office_hours
schedule
textbook
attendance
assig_dates
present_dates
midterm_date
final_date
assig_grade
present_grade
midterm_grade
final_grade
late_submission

Table 9: Column names

So, our database will be consisted of a table of 4 rows (4 courses) and 22 columns. Once the CSV file is created, it will be possible to import it into an SQLite table and store it in an SQLite database (.db file).

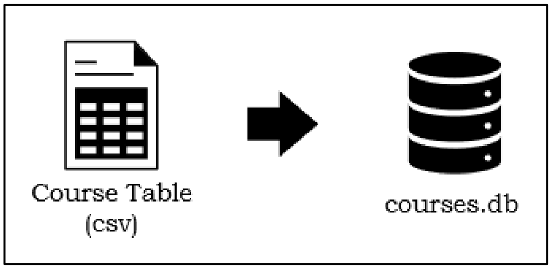


Figure 8: CSV file to database

This is done by executing the following command lines in the sqlite3 prompt where “courses_table.csv” is the CSV file, “courses_table” is the SQLite table and “courses.db” the SQLite database.

- Open SQLite database:

```
sqlite> .open courses.db
```

Listing 1: Opening database

- Set the mode to CSV to instruct the command-line shell program to interpret the input file as a CSV file:

```
sqlite> .mode csv
```

Listing 2: Activating CSV mode

- Import the data from the CSV file into the SQLite table:

```
sqlite> .import courses_table.csv courses_table
```

Listing 3: Importing CSV file into SQLite table

Once this is done, we will have our database defined and ready to use.

5.2. Speech-To-Text Implementation

The next step is to implement the Speech-To-Text system. A STT system refers to the ability of a computer software to identify words and phrases in spoken language and convert them to human readable text.

Therefore, the main goal of this step is to capture audio from the microphone and transcribe it into text (Figure 9). As mentioned in section 4.3.2, we have used the Python library named *SpeechRecognition* with *Google Speech Recognition* API and *PyAudio* in order to use our microphone (Logitech H390 USB Computer Headset).

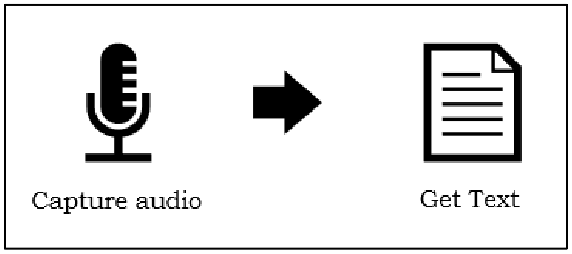


Figure 9: Speech-To-Text

So, first of all, when the program is executed, a welcoming sentence is reproduced: “Welcome to your virtual assistant, how can I help you?”. After that, the user can ask one of the fifteen questions appeared on section 5.1. The microphone will listen for ten seconds (this time is defined at the command line, and therefore it is variable) and will then stop recording. Speech will be converted from physical sound to an electrical signal with the microphone. During this process, several parameters can be adjusted to get a better-quality audio, such as:

- **Energy threshold:** minimum audio energy to consider for recording.
- **Ambient noise:** calibrates the energy threshold with the ambient energy level.
- **Pause threshold:** seconds of non-speaking audio before a phrase is considered complete.

Regarding of how Speech-To-Text is works, this is done by following a Hidden Markov Model (HMM) in most cases [30]. In a typical HMM, the speech signal is divided into 10-millisecond fragments, and the power spectrum of each of them is mapped to a vector of real numbers [30]. So, the final output of the HMM is a sequence of these vectors [30].

Then, the vectors are matched to one or more phonemes (fundamental units of speech) and a special algorithm is used to determine the most likely word or words that produce the sequence of phonemes [30].

In order to reduce computational requirements, voice activity detectors (VADs) are used to keep just the portions that are likely to contain speech and therefore, prevent from wasting time analyzing unnecessary parts of the signal [30].

So, once our recording is finished, the recognizer will try to recognize the words that appear in the recorded audio by breaking down the audio into individual sounds and analyzing each sound. It will then try to find the most probable word fit in that language and will transcribe those sounds into text. Indeed, this library offers several languages and dialects to select from and obtain better results. We have selected US-English for the process.

Finally, after transcribing the speech into text, the recognizer will return a text string containing the sentence that has been recorded.

5.3. Database Query

The third step is to perform an SQL query on our database to look for an answer depending on the text containing the question that was recorded and transcribed on the previous stage.

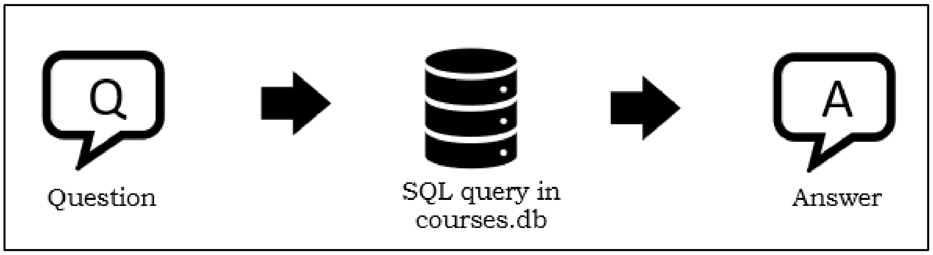


Figure 10: SQL Database Query

In SQL language, the SELECT and FROM statements are used to select data from a table in a database and the WHERE clause is used to filter records and extract only those records that fulfill a specified condition.

```
SELECT column1, column2,... FROM table_name WHERE condition;
```

Listing 4: SELECT, FROM and WHERE statement syntax

In our case, since there are fifteen possible questions (Section 5.1), there will be fifteen possible answers for each course, and therefore, fifteen columns will provide us that information. To select the corresponding column for each question keywords will be used. So, whenever a keyword or keywords are detected in a question, a column will be assigned to that request. Table 10 shows the corresponding keyword(s) and column containing the information for each of the questions:

QUESTION	**KEYWORD(S)**	**COLUMN NAME**
1	telephone	telephone
2	address	office_address
3	hours	office_hours
4	schedule	schedule
5	textbook	textbook
6	attendance	attendance
7	*number* & assignment & due	assig_dates
8	*number* & presentation & due	present_dates
9	midterm	midterm_date
10	final	final_date
11	assignment & grading	assig_grade
12	presentation & grading	present_grade
13	midterm & grading	midterm_grade
14	final & grading	final_grade
15	late	late_submission

Table 10: Question, keywords, and column mapping

In order to select the specific course, the column named *course_name* will be used, which will be the primary key. The primary key uniquely identifies each record in a table.

As an example, for question 1 the SQL query statement will be as follows:

Question: Which is Human-Computer Interaction professor's telephone?

- *Keyword*: telephone → *Column*: telephone
- *Course name*: Human-Computer Interaction

SQL Query:

```
SELECT telephone FROM courses_table WHERE course_name = Human Computer
Interaction;
```

Listing 5: SQL Query example

So, after completing the SQL query statement, the database needs to be connected to SQLite and then, the SQL query statement can be executed in order to get the specific answer. To connect to the database, we will use:

```
conn = sqlite3.connect("courses.db")
```

Listing 6: SQL database connection

Finally, once the database connection is successfully done and the SQL query is executed, the complete answer sentence will be built and ready to be reproduced with the corresponding professor's voice in the next step.

As an example, for question 1 the SQL query statement answer, and complete answer will be as follows:

Question: Which is Human-Computer Interaction professor's telephone?

SQL Query Answer: (312) 567 – 5293

Complete Answer: Human-Computer Interaction professor's telephone is (312) 567 – 5293.

The fifteen complete answers for syllabus in Figure 6 are:

1. **Which is Human-Computer Interaction professor's telephone?**
   Human Computer Interaction professor's telephone is (312) 567 – 5293.
2. **Which is Human-Computer Interaction professor's office address?**
   Human Computer Interaction professor's office address is Perlstein Hall, Room 227.
3. **What are Human-Computer Interaction professor's office hours?**
   Human Computer Interaction professor's office hours are Mondays before class.
4. **Which is the schedule for Human-Computer Interaction?**
   The schedule for Human Computer Interaction is Mondays and Wednesdays 2 – 4 pm.
5. **Which is the textbook for Human-Computer Interaction?**
   There is no textbook required in this course.
6. **Is the attendance required in Human-Computer Interaction?**
   The attendance in Human Computer Interaction is required.
7. **When is Human-Computer Interaction first assignment due?**
   Human Computer Interaction first assignment is due March 25.
8. **When is Human-Computer Interaction first presentation due?**
   Human Computer Interaction first presentation is due March 1.
9. **When is Human-Computer Interaction midterm exam?**
   Human Computer Interaction midterm exam is March 15.
10. **When is Human-Computer Interaction final exam?**
    Human Computer Interaction final exam is May 15.
11. **What is the assignment grading in Human-Computer Interaction?**
    The assignment grading in Human Computer Interaction is 25%
12. **What is the presentation grading in Human-Computer Interaction?**
    The presentation grading in Human Computer Interaction is 10%.

13. **What is the midterm exam grading in Human-Computer Interaction?**
    The midterm exam grading in Human Computer Interaction is 25%.
14. **What is the final exam grading in Human-Computer Interaction?**
    The final exam grading in Human Computer Interaction is 40%.
15. **What is the penalty for late submission in Human-Computer Interaction?**
    The penalty for late submission in Human Computer Interaction is 15%.

5.4. Text-To-Speech with Voice Cloning Implementation

The objective of this step is to convert the answer that was derived from the database query into speech. The aim is to try to maximize naturalness, intelligibility, and similarity to a human voice. Hence, apart from just converting the text into speech, we also applied voice cloning technology to get to clone specific professors' voices and enhance the experience of users.

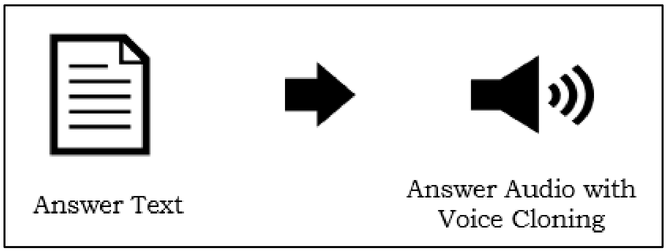


Figure 11: Text-To-Speech with Voice Cloning

So, as pointed out in previous sections, four courses and their corresponding teachers' voices have been selected.

COURSE	**PROFESSOR VOICE**
Human-Computer Interaction	Female 1
Musics of the World	Female 2
Advanced Informatics	Male 2
Internet Technologies and Web Design	Male 2

Table 11: Courses and corresponding professors

The technology that has been used is Speaker Verification to Text to Speech Synthesis (SV2TTS) [12], as mentioned in section 4.3.4.

Usually, in order to clone a voice, a deep neural network is trained using a collection of several hours of recorded speech from the speaker. And when a new voice wants to be cloned, a new dataset and a retrained model are required, making it highly expensive. [12]

So, the open-source SV2TTS tool [25], based on a recent research study [26], allows to clone a voice only from few seconds of speech and without needing to retrain the model.

Indeed, the interest lies in creating a fixed model able to incorporate new voices with little data instead of a complete training of a single-speaker model. The objective is to condition a model trained to generalize to new speakers on an embedding (low-dimensional and meaningful representation of the voice of a speaker) of the voice to clone. This is much more data efficient, orders of magnitude faster and less computationally expensive than training a separate model for each speaker. [12]

The complete SV2TTS framework is a three-stage pipeline that consists of a speaker encoder, a synthesizer, and a vocoder. First, the speaker encoder is fed a reference source audio of the speaker to clone, and it generates an embedding, which is used to condition the synthesizer. The synthesizer gets a text as an input, which is processed as a phoneme sequence, and outputs a log-mel spectrogram. A log-mel spectrogram is a deterministic, non-invertible (lossy) function that extracts speech features from a waveform, so as to handle speech in a more tractable fashion in machine learning. Finally, the vocoder generates the speech waveform considering the output of the synthesizer. The quality of the generated audio can only be as good as that of the reference audio. [12]

These three stages are trained separately, but the synthesizer needs to have embeddings from a trained speaker encoder and the vocoder log-mel spectrograms from a trained synthesizer. Apart from that, the SV2TTS includes an improvement that allows to use two different datasets when training (one for the speaker encoder and the other one for the synthesizer and the vocoder). These avoids the problem of a single dataset needing to meet the requirements for all three models. [12]

The whole SV2TTS framework is shown in Figure 12 and each of the three parts will be described next:

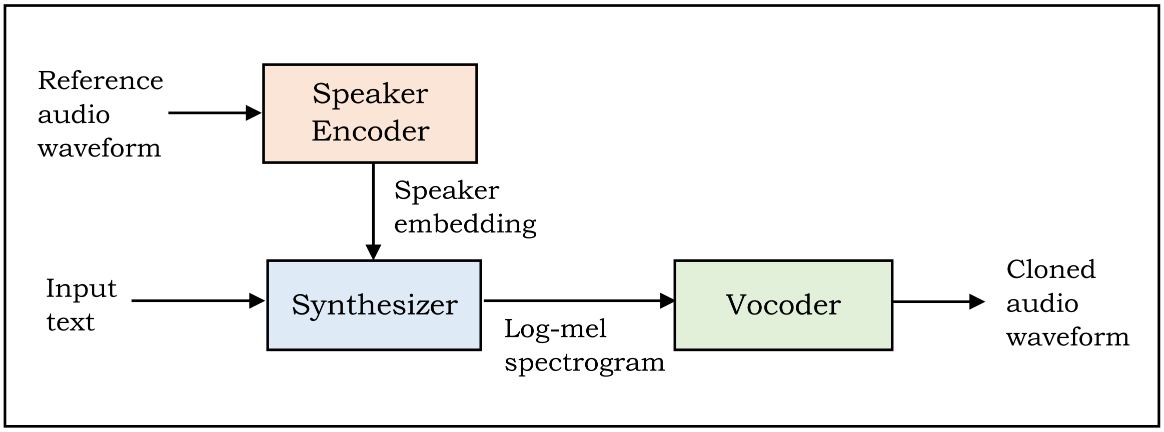


Figure 12: SV2TTS Framework

Speaker encoder

The speaker encoder is responsible for deriving an embedding from a reference source audio of a speaker that wants to be cloned. This model and its training process are described over several papers ([27] and [31]) and it is fundamentally based on the Generalized End-To-End loss (GE2E) approach [27]. SV2TTS reproduced this model with a PyTorch implementation of their own.

As the name of the whole tool implies, the speaker encoder is trained on a speaker verification task. Speaker verification is a typical application of biometrics where the identity of a person is verified through their voice. So, a template is created for a person by deriving her/his speaker embedding from the reference source audio, which is a meaningful representation of her/his voice. The GE2E loss simulates this process to optimize the model. [12]

The input to the model is a reference audio source which is consisted of 40-channels log-mel spectrograms with a 25ms window width and a 10ms step. The output is an embedding that consists of a vector of 256 elements. [12]

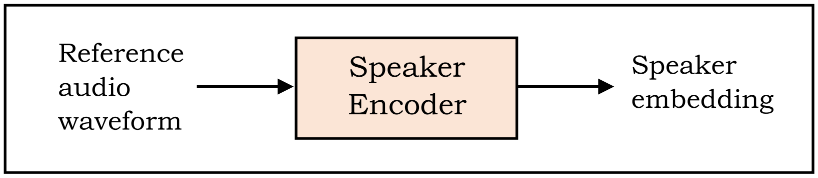


Figure 13: SV2TTS Speaker Encoder

For the speaker encoder, it is sought to have a model that is able to capture as many characteristics as possible of the reference human voice. Therefore, at training time, the authors used a dataset with a large corpus of different speakers. Indeed, the model computes the embeddings of M utterances of fixed duration from N speakers when training. The fixed duration of these sentences is of 1.6 seconds, which are partial utterances sampled from the longer complete utterances in the dataset. Therefore, at inference time a sentence is split in segments of 1.6 seconds overlapping by 50%, and the encoder forwards each segment individually. The resulting outputs are averaged and then normalized to produce the speaker embedding. [12]

Synthesizer

The synthesizer is in charge of predicting a log-mel spectrogram having text as an input, getting conditioned by the embedding created by the speaker encoder. The Tensorflow implementation [32] of Tacotron 2 is used, stripping Wavenet and implementing some modifications.

Tacotron is a recurrent sequence-to-sequence model that features an ecoder-decoder structure to predict a mel spectrogram from text [12].

The synthesizer targets log-mel spectrograms that are computed from a 50ms window with a 12.5ms step and have 80 channels. These are generated from an input text and taking into account the speaker embedding derived from the speaker encoder. [12]

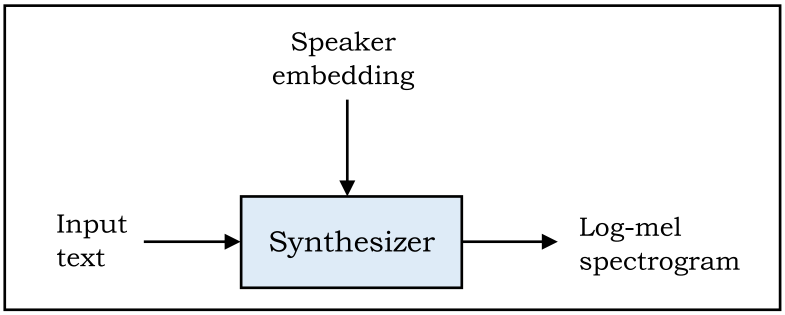


Figure 14: SV2TTS Synthesizer

The input texts are not processed for pronunciation, but there are a few cleaning procedures:

- Replacing abbreviations and numbers by their complete textual form.
- Forcing all characters to ASCII.
- Normalizing whitespaces.
- Making all characters lowercase.

For the training process of the synthesizer, a dataset that contains transcripts is required. The author used utterances not shorter than 1.6 seconds (the duration of partial utterances used for training the encoder), and not longer than 11.25 seconds in training time. [12]

When generating sentences at inference time, in order to synthesize the spectrogram in multiple parts, the input text has to be split, which gives the advantage of fast execution [12].

Vocoder

The vocoder infers an audio waveform from the log-mel spectrogram generated by the synthesizer. In SV2TTS the vocoder is based on the WaveNet approach [29], which even though it has been at the heart of deep learning since its release, it is also known for being the slowest practical architecture at inference time. In spite of that, WaveNet remains the vocoder in SV2TTS since speed is not the main concern and because Google's own WaveNet implementation with several improvements already generates good results. This improved model is called WaveRNN [33]. So, the vocoder model used in SV2TTS is an open source PyTorch implementation [34] that is based on WaveRNN but presents quite a few different design choices. [12]

The inputs to the model are the log-mel spectrogram generated by the synthesizer, and the output the waveform of the cloned audio.

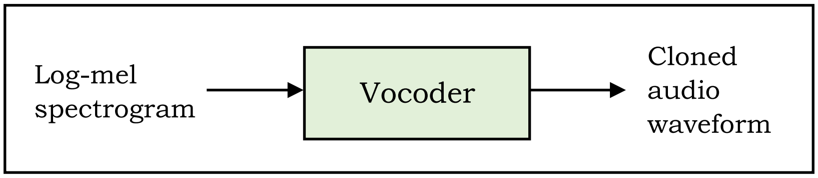


Figure 15: SV2TTS Vocoder

For the training process of the vocoder, as well as for the synthesizer, the same dataset that contains transcripts is required. At training time, the model predicts fixed-size waveform segments, and the generation of the whole utterance waveform is done in parallel over all segments. A small part of the end of a segment is repeated at the beginning of the following one, in order to preserve some context between segments. This process is called folding. Then, the overlapping sections of consecutive segments are merged by a cross-fade to retrieve the unfolded tensor. [12]

This can be seen in Figure 16:

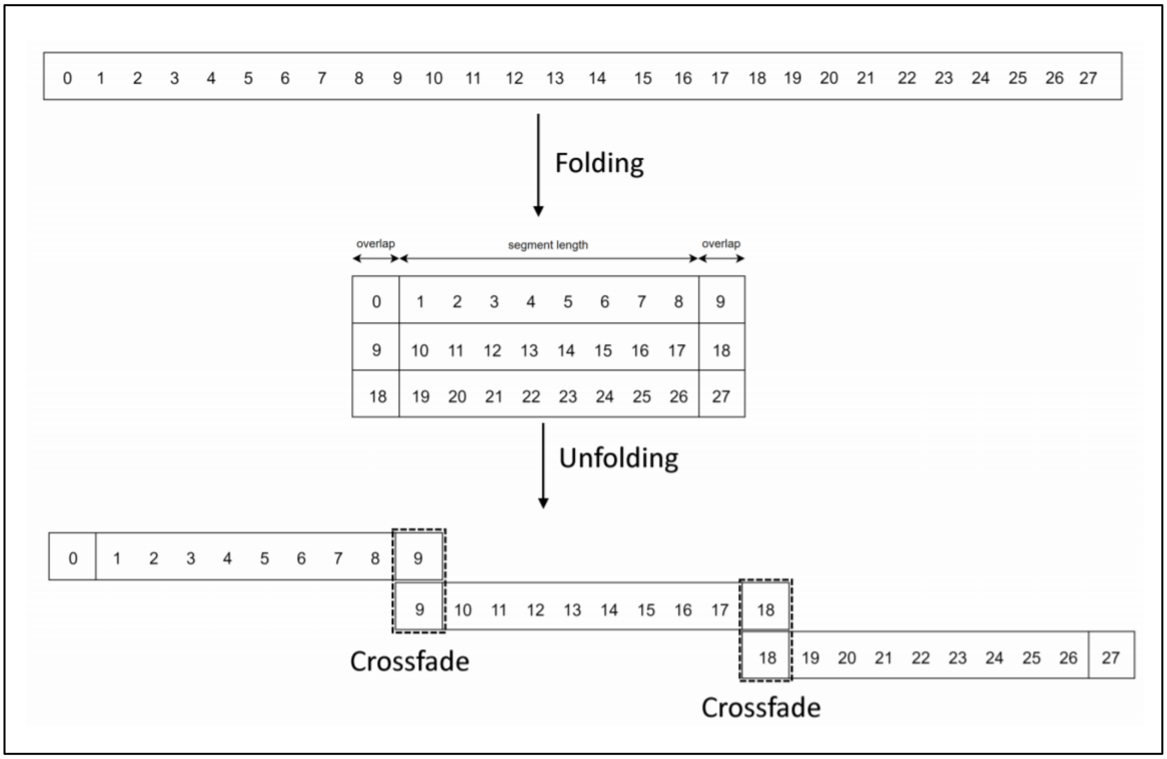


Figure 16: Folding and unfolding process

On the topic of speed, the vocoder usually runs below real-time and the inference speed is highly dependent of the number of folds, and therefore, on the duration of the utterances [12].

Once the SV2TTS framework has been explained, our specific use case can be described. As mentioned, the purpose of this step is to have the answer from the previous step reproduced in the corresponding professor's voice. So, following SV2TTS, we will need as an input:

- A reference audio file of the professor's speech in WAV format.
- A string sentence, that in our case, will contain the answer to the user's request that has been derived from the SQL query. Following the example in the previous section, the string would be: "Human-Computer Interaction professor's telephone is (312) 567 – 5293."

The output will be:

- A WAV format audio file containing the input sentence reproduced with the voice in the reference audio, which is reproduces out loud for the user. Since the SV2TTS tool's vocoder outputs a waveform, this will need to be converted into a padded array in order to be saved in a WAV format audio file using its sample rate.

This is shown in Figure 17:

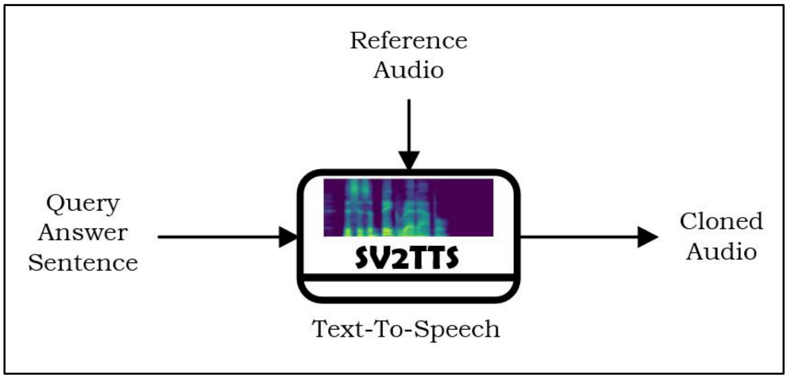


Figure 17: SV2TTS Application

When it comes to the duration of the reference audio, there is a large discrepancy among the different methods, ranging from half an hour per speaker to only a few seconds. This factor is said to be determining of the similarity of the generated voice with respect to the true voice of the speaker. Therefore, we will analyze the results depending on the duration of the reference source audio. We have chosen six different lengths: 5 seconds, 5 minutes, 15 minutes, 30 minutes, 45 minutes and 1 hour.

Although the SV2TTS tool claims it is enough with a 5 second length audio, we think it is interesting to observe if results change depending on the duration of the source.

When a voice is cloned for the first time, an embedding is created based on the reference audio. The time that it takes to create such embedding depends on the duration of the reference audio, being directly proportional, the longer the reference audio, the more time it takes to create an embedding. These results will also be shown in section 6.

Once the embedding for a particular voice is created, this can be saved and used for the following sentences that want to be cloned using the same voice, which reduces considerably the time of computation.

Regarding the similarity of the cloned voice to the real one, a similarity measurement has been applied, based on *Resemblyzer* [35], a python package to analyze and compare voices with deep learning (results in section 6).

Finally, after following these four steps, we have achieved our goal of building a tool that reproduces the answer to the user's question about information of a course using the corresponding professor's voice. In order to test it, and get to know its value for users, we have designed a qualitative research study that will be explained next.

5.5. Qualitative Research Study Design

Our goal in this last step is to design a qualitative research study about the possible use of Voice Cloning technology for good purposes, in the educational field for instance. Qualitative research involves collecting and analyzing non-numerical data (e.g., text, video, or audio) to understand concepts, opinions, or experiences [36]. Therefore, different groups of people will need to be involved on the research: college students and professors. The main idea is to interview them in order to collect their opinion and reaction to a video demo about the developed Voice Cloning Assistant technology and to a live interaction experience with it. This will also include some survey type questions.

So, the design of the research study will be done in steps:

1. Recruit suitable participants for the research. Get in contact with college students, college student's parents and professors. Try to recruit, for instance, more or less 10 people.
2. Think of the location (building and room) where the research will be performed. The location will need a screen.
3. Get all the material prepared. In this case, we will need the video demo, the tool for the live interaction experience and the questions.
4. Establish a recording method. It will be preferable to videorecord the whole process since facial expressions and body language can provide valuable information. If it is not possible, audio record the interview. Regardless of the method, the permission from each participant needs to be obtained.

5. Due to the ongoing COVID-19 pandemic, we need to follow the University safety committee guidelines, which include some prescreen questions that are to be asked, on the phone prior to face-to-face meeting, of the participants [37].

    - In the past two weeks have you come in close contact with someone who has tested positive for COVID-19?
    - Have you or someone who lives with you traveled out of state/country in the last 30 days?
    - Do you live in a high-populated area/apartment complex that is exposed to COVID-19?
    - In the past two weeks, have you experienced any of the following symptoms: fever or chills, cough, shortness of breath or difficulty breathing, fatigue, muscle or body aches, headache, new loss of taste or smell, sore throat, congestion or runny nose, nausea or vomiting, diarrhea?

    In general, participants that answer yes to any of the questions should not be invited to come to campus to participate until two weeks have passed since the date of the triggering event.

6. One participant will be scheduled per defined time period and therefore, the interview will be done one by one. Upon arrival [37]:

    - Assess body temperature.
    - Re-ask questions related COVID-19 symptoms.
    - Participants need to be wearing mask; researchers need to provide masks to those participants that do not have masks.

7. Sit down with the participant in a safe space where they feel comfortable and relaxed. Start introducing yourself and explaining the format of the interview: first, a survey with some questions about the familiarity with Speech Synthesis and Voice Cloning technology will be conducted, and consecutively some concepts will be explained in order to understand the goal of the tool we have built and the purpose of the present research study.

    Then, the participant will watch a video demo about the tool we have built and will participate in a live interaction with it. Finally, the interview will end with some final questions.

8. Start videorecording and conduct the initial survey:

    - Demographic questions: age, gender, nationality, current employment status and highest degree or level of education.
    - On a scale from 0 (Nothing) – 5 (Very much), how much do you know about Artificial Intelligence?

- On a scale from 0 (Nothing) – 5 (Very much), how much do you support or oppose the development of Artificial Intelligence?
- Have you heard about Speech Synthesis before?
- Have you heard about Voice Cloning before?

9. Clarify briefly what Artificial Intelligence, Speech Synthesis and Voice Cloning are. Then, explain the main goal of the developed project and the purpose of the research.
    - "We have built a voice query assistant to obtain course information that uses voice cloning technology to clone professors' voices."
    - "We want to discover your opinion about this emerging technology and your reaction to this tool and to hearing to cloned familiar/unfamiliar voices."
10. Show the video demo on the screen and videorecord the reaction of the participant while watching it.
11. In order to have the participant interact with the tool, provide the possible 15 questions that can be asked. Start the execution of the tool on the screen and invite the participant to ask one of the fifteen questions. Wait for the tool to reproduce the answer and videorecord the reaction of the participant while hearing it.
12. Continue videorecording and perform a subjective evaluation to report the Mean Opinion Score (MOS). For that, participants will be asked to rate the naturalness of the cloned voices and their similarity with the real ones on a scale from 1 to 5.
13. Continue videorecording and ask the final questions:
    - Do you think using Voice Cloning is ethical?
    - Do you think Voice Cloning can be used for good (education, research, healthcare, cybersecurity…) as we have done in this example?
    - Do you think Voice Cloning can be problematic in the future?
    - If yes, how can Voice Cloning be problematic?
    - Do you think there should be any regulation on Voice Cloning?
    - Do you know if there is any regulation on Voice Cloning in your country?
    - How do you think Voice Cloning can be regulated?
    - Would you let your voice be cloned for good (education, research, healthcare, cybersecurity…) as we have done in this example?
    - How would you feel about getting your voice cloned without your consent?
14. Thank the participant for coming.

15. Once all the interviews are finished, analyze the collected data and draw conclusions from it by answering the following questions:
    - How do *adults* interpret *Artificial Intelligence*?
    - How do *adults* interpret *Voice Cloning technology*?
    - How do *adults* see the future of *Voice Cloning technology*?
    - How do *adults* see the regulation of *Voice Cloning technology*? How would they regulate it?
    - What do *students* think about *using Voice Cloning in education*?
    - What do *professors* think about *using Voice Cloning in education*?
    - How do *students* feel about *getting their own voice cloned* for good?
    - How do *professors* feel about *getting their own voice cloned* for good?
    - How do *students* feel about *getting their own voice cloned* without their consent?
    - How do *professors* feel about *getting their own voice cloned* without their consent?
    - How does *hearing an unreal but familiar voice* impact *students*?
    - How much of a threat is *getting a voice cloned so easily* for *professors*?

6. Interpretation of the Results

During this section, the results that have been obtained from the development of the project will be analyzed and compared, and some possible and future improvements will be described also.

6.1. Project results

In order to test the tool that we have built, the cloned answers for all fifteen questions for each of the four courses have been captured. As pointed out in section 5.4, this process has been repeated for different reference source audio durations. We have chosen six different lengths: 5 seconds, 5 minutes, 15 minutes, 30 minutes, 45 minutes and 1 hour. That way, the comparison of female and male voices will be assessed apart from checking the impact of changing the reference audio duration.

When creating an embedding for a new voice for the first time, there is a delay that is due to the SV2TTS speaker encoder, which is directly proportional to the duration of the reference audio. In other words, the longer the reference audio, the more time it takes to create an embedding. These results can be seen in Table 12 and Figure 18:

REFERENCE AUDIO LENGTH	**EMBEDDING CREATION TIME**
5 sec	2 sec
5 min	25 sec
15 min	1 min 15 sec
30 min	2 min 25 sec
45 min	4 min 15 sec
1 hour	6 min 10 sec

Table 12: Embedding creation time

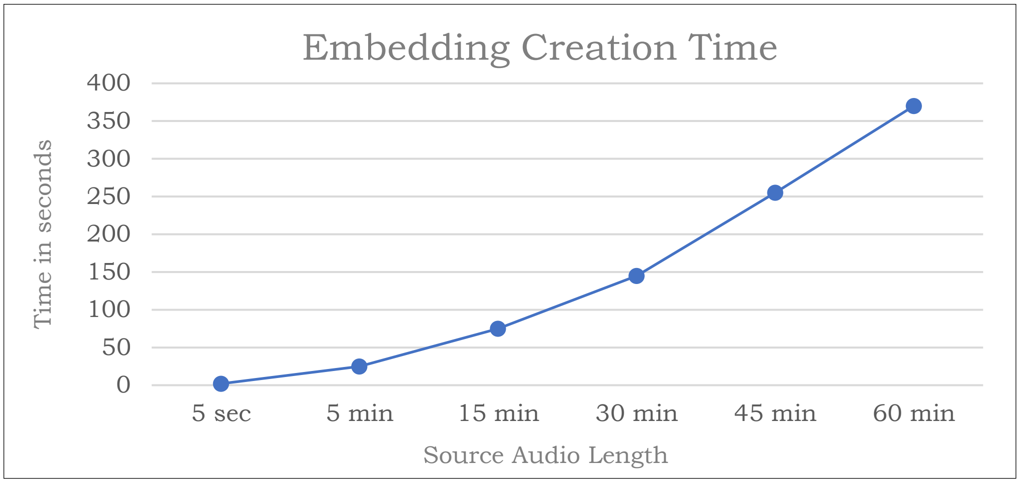


Figure 18: Embedding creation time

Note that for a reference audio of 5 seconds, the embedding creation time is just of 2 seconds, while for a reference audio of an hour, the delay increases significantly. Apart from that, the computational resources (CPU, Memory, Disk…) that are used also increment substantially when the duration is higher, which can be problematic. So, this will need to be considered when choosing the duration of the reference audio and think about if it is worth taking into account the quality of the cloned voice, which will be analyzed afterwards.

Once the embedding of a voice is created, this can be saved and used for the following sentences that want to be cloned using the same voice. In these cases, when generating the answer audio, there is also a little delay that is due to the SV2TTS vocoder. Nevertheless, this delay is not related to the duration of the reference audio, and it is the same for the six of them. The computational resources do not increase either depending on the sentence.

This delay, as stated in section 5.4, is associated with the number of folds in the folding process of the vocoder. Indeed, it depends on the length of each of the fifteen possible answers. Since all the answers contain a single sentence and the number of words does not vary much, the generation time variation from one answer to another is also insignificant, with a variation of just five seconds.

REFERENCE AUDIO LENGTH	**ANSWER GENERATION TIME**
All reference audios	25-30 sec

Table 13: Answer generation time

When it comes the similarity of the cloned voice to the real one, we have used *Resemblyzer* [35], a python package to analyze and compare voices. This package offers a voice similarity metric that compares different audio files and gets a value from 0 to 1 on how similar they sound.

This is done by calculating a similarity matrix, which is the result of a two-by-two comparison that applies the cosine similarity by computing the dot product of two vectors that correspond to the two files that is comparing. An optimal model is expected to output high similarity values (represented in green in Figure 19) when the real and cloned voices are compared and lower values elsewhere (represented in red in Figure 19).

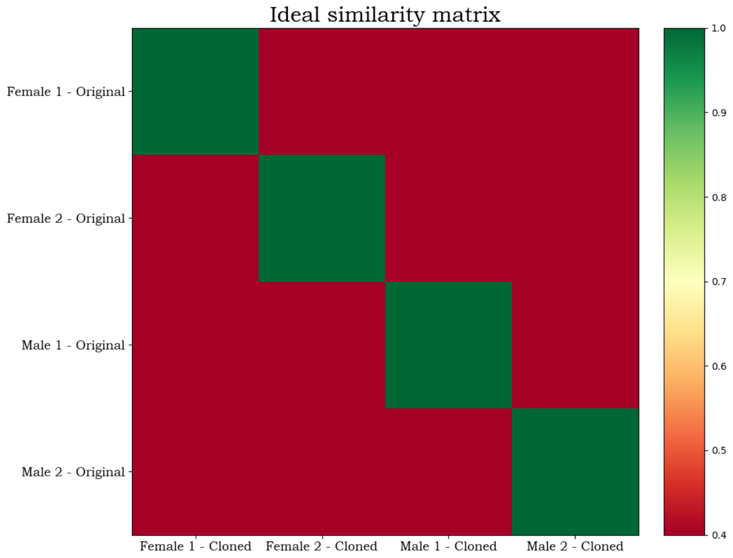


Figure 19: Ideal similarity matrix

Different similarity matrices have been calculated to draw some conclusion. For each of the four voices, we have computed two matrices:

- A 1x15 matrix that compares the real voice with the fifteen possible sentences reproduced with the cloned voice. We have chosen to use the duration of 5 seconds for all fifteen sentences and four voices.
- A 1x6 matrix that compares the real voice with six cloned voices that have been built using the six durations (5 seconds, 5 minutes, 15 minutes, 30 minutes, 45 minutes and 1 hour) of the reference audio of the speaker. We have chosen the sentence that best sounds, considering naturalness and intelligibility, among the fifteen possible answers for each speaker. We have taken the previous matrix results as a reference to make the pick.

These matrices are shown in the following figures:

Female 1

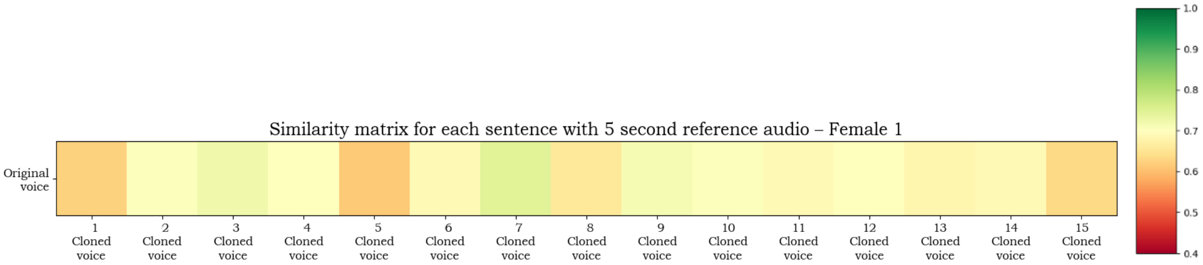


Figure 20: Similarity matrix 1 – Female 1

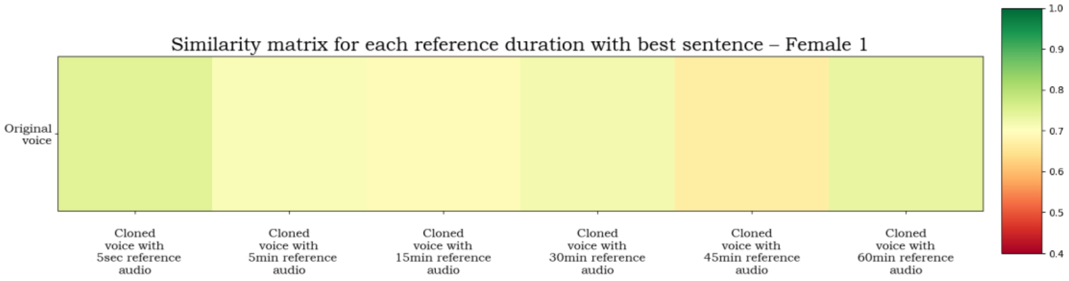


Figure 21: Similarity matrix 2 – Female 1

Female 2

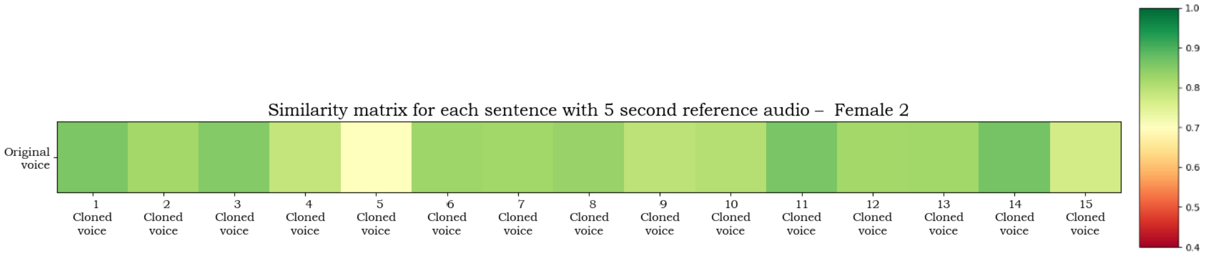


Figure 22: Similarity matrix 1 – Female 2

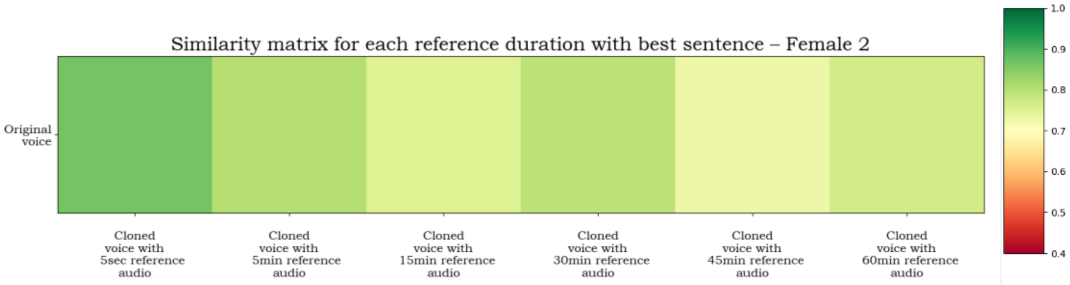


Figure 23: Similarity matrix 2 – Female 2

Male 1

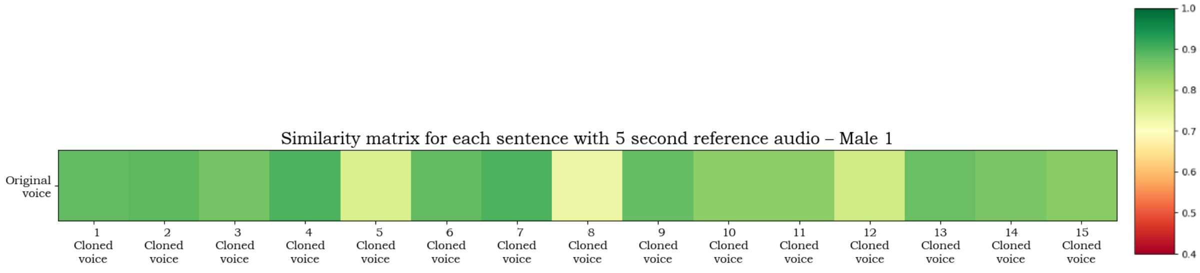


Figure 24: Similarity matrix 1 – Male 1

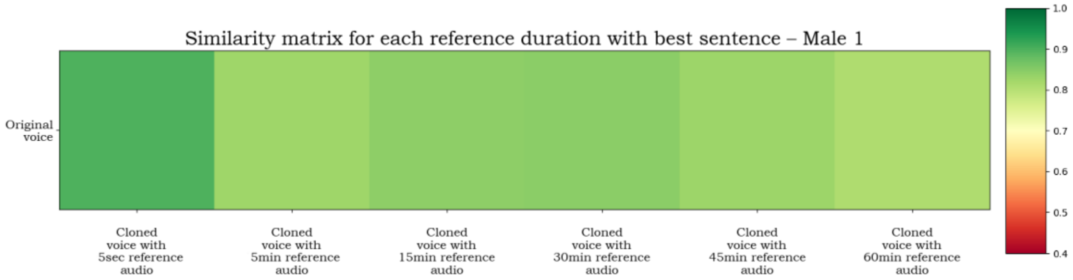


Figure 25: Similarity matrix 2 – Male 1

Male 2

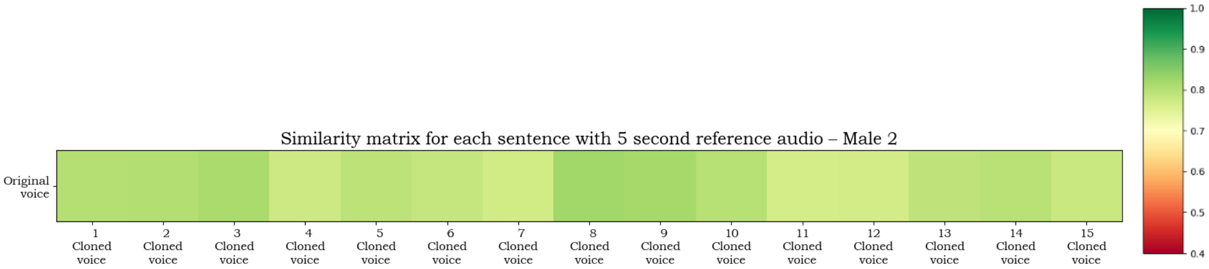


Figure 26: Similarity matrix 1 – Male 2

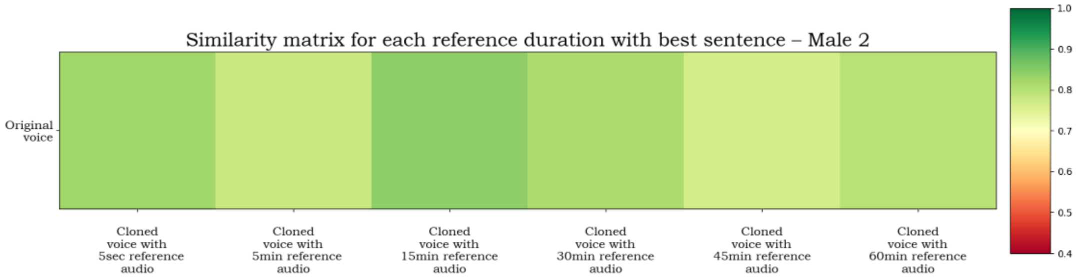


Figure 27: Similarity matrix 2 – Male 2

Apart from those, by combining all four voices, a 4x4 similarity matrix has been computed, which compares the real and cloned voices using the 5 second reference audio in all cases. This is shown in Figure 28:

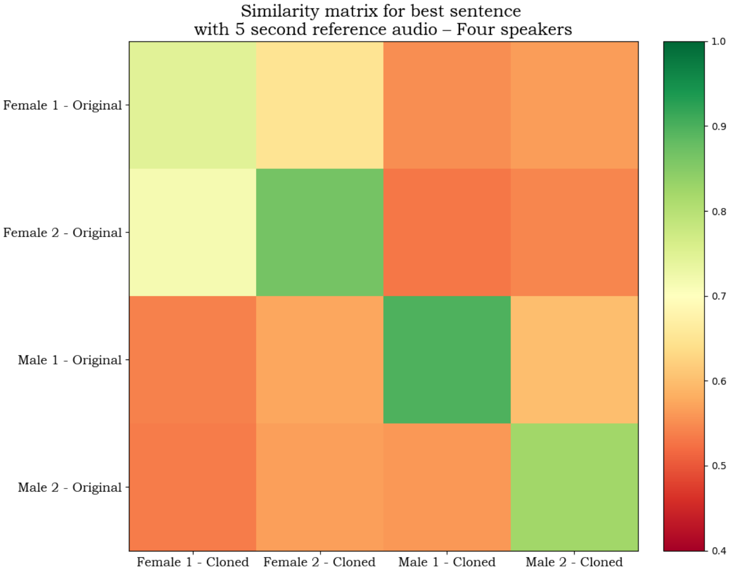


Figure 28: Similarity matrix– Four speakers

As it can be seen in the previous figures, the best results are obtained with the Male 1 voice, followed by the voices of Female 2 and Male 2, being the worst cloned voice Female 1. This can be because of the reference audio quality.

Figure 28 shows that the tool gets favorable results since it approximates to the ideal similarity matrix, having the best values in the diagonal. We can also observe that the female voices are quite similar between them but have nothing to do with the male ones.

As a final test, the *Resemblyzer* [35] package provides a speaker diarization video demo, which recognizes who is talking when with a reference audio per speaker. This has been modified to our case and a diarization video has been created using our four voices. In order to show it in this paper, we have captured screenshots of the video when the tool recognizes each of the cloned voices and matches them with the real ones.

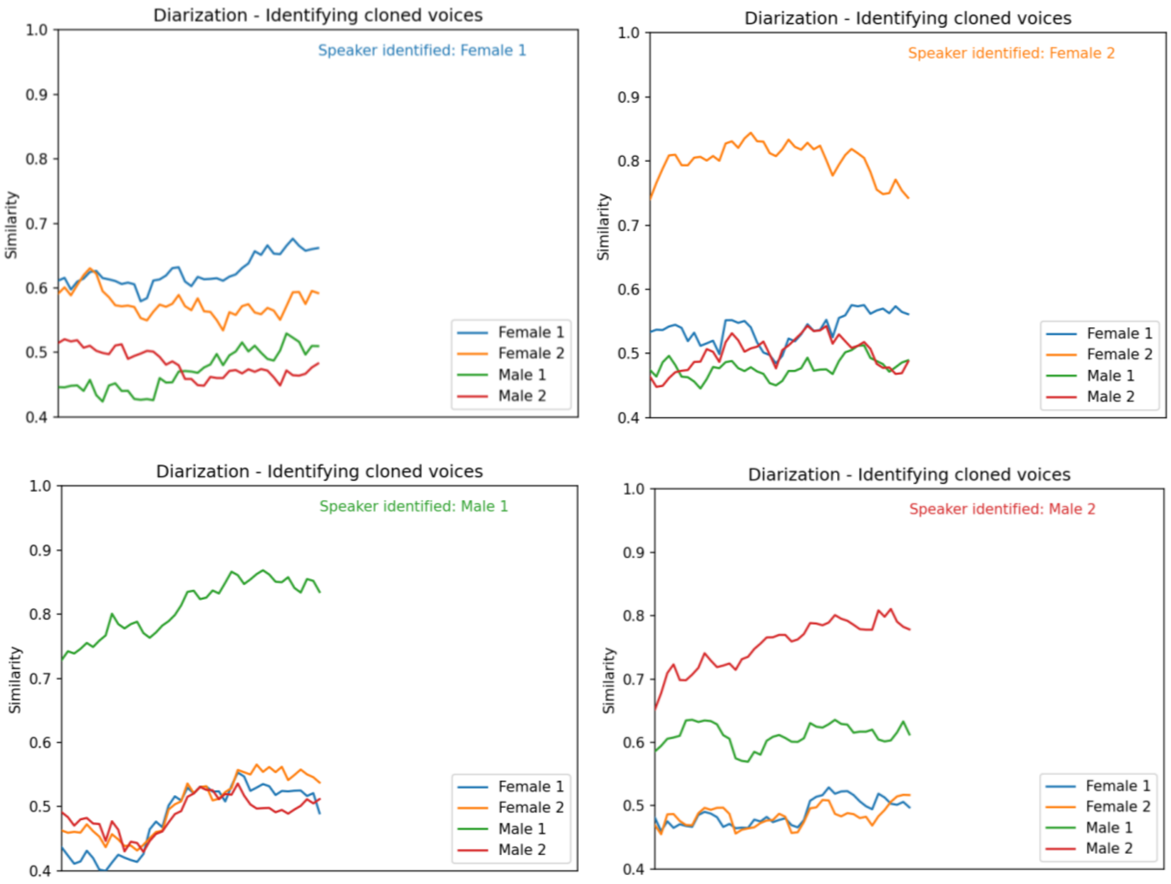


Figure 29: Diarization

Looking at the figures above, we have reached to some conclusions:

- We find the voice cloning ability of the framework to be reasonably satisfactory. In general, the similarity values are quite good, since when a similarity value is higher than 0.70, the cloned voice is able to be identified and matched to the real voice in a confident manner.
- At a glance, it seems male voices present better results than female voices. This can be due to the quality of the reference audio, which impacts the quality of the result. Therefore, further research is needed to confirm this statement.
- As already stated previously, when the duration of the reference audio increments, there is an increase in the delay and in the computational resources. Observing the results, we decided this is not worth, since the similarity when using a reference audio of minimum duration (5 seconds) is the same or even higher.

- Depending on the length of each of the fifteen answers, sentences that are too short are stretched out with long pauses, and for too long ones, the voice is sometimes rushed. This may be due to the limits that are imposed on the duration of the sentences in the training dataset (1.6 s – 11.25 s) [12], which has been mentioned in section 5.4.
- The prosody can be sometimes unnatural and a little robotic, with pauses at unexpected locations in the sentence, or the lack of pauses where they are expected. This can be the result of having reference audios of speakers talking slowly [12], and therefore, the speaker encoder captures some form of prosody in the embedding of such speaker.
- Punctuation is not supported by the SV2TTS model [12], so it is discarded. In our case, this cannot be perceived for periods, as each answer contains just one sentence, but it is noticeable for commas, which are ignored.

6.2. Improvements

Once the results were analyzed and we have pointed out some conclusions, we thought of some possible improvements. Some of them have been carried out during the project, and others will be suggested for future research.

- When it comes to gender comparison results, in order to confirm that male voice features provide better results than female ones, more extensive research is needed. This could be done by collecting better quality reference audio samples and testing the tool in more cases.
- Regarding the duration of the reference audio, so as to improve the results, we thought of stripping out the silent sections of them and compare the outcomes. So, with an audio editor, we trimmed the unvoiced parts from five out of six reference audios (the 5 second one does not contain any silent chunk) and kept the parts where speech is detected.

  Then, the same procedure was applied, and the fifteen cloned answers were obtained for the five different reference audio durations and for each of the four speakers. When conducting the comparison, in 52% of the cases the results improved with respect to the case in which we used reference audios that contained silent parts. This could be a first step of a future research case topic.

  As an example, we have picked the 30-minute original reference audio and the same one without silent sections for Male 2. The next figure shows the comparison of the results obtained with both. In this case, the outcomes improve when silences are stripped out.

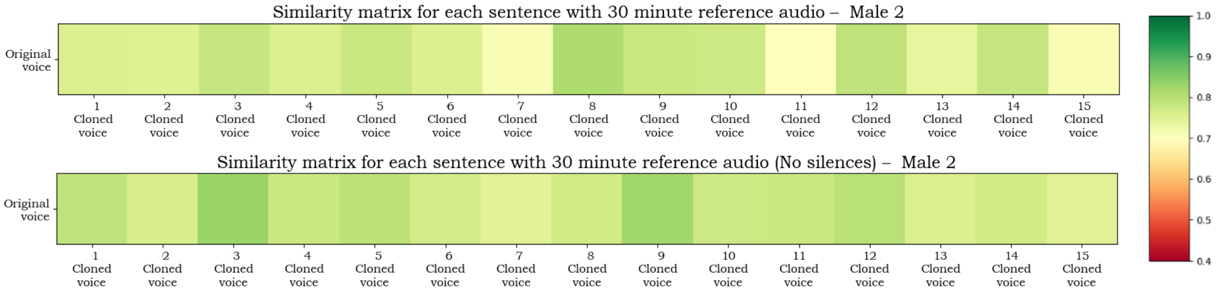


Figure 30: Comparison when stripping silences - Male 2

- In order to have an impact on the natural flow of the spoken response and get closer to the way in which people talk in everyday life, we came up with the idea of using subject pronouns, which replace the subject noun phrase in the answer. For example:

  Question: Which is Human-Computer Interaction professor's telephone?

  Complete Answer: Human-Computer Interaction professor's telephone is (312) 567 – 5293.

  Answer with subject pronoun: It is (312) 567 – 5293.

  We have analyzed the results of shortening the answers by using subject pronouns, but in some cases, these are less satisfactory than what we expected. Although in certain sentences the results are acceptable and they could even be used rather than the whole answers, some others give a synthetic and unnatural impression. As stated in the fourth bullet point of the previous section, sentences that are too short are stretched out with long pauses.

- For the purpose of refining the punctuation and prosody issue mentioned in the last two bullet points of the prior section, the author of SV2TTS suggests inserting line breaks between parts that should be synthesized individually [12]. This is an interesting point considered for future research since it has not been assessed throughout the current project.

- To try to improve the performance speed of the process, we thought to switch to an in-memory database. In fact, SQLite offers a configuration in which the database is stored in RAM memory instead of in a single ordinary disk file. This way, a new database is created purely in memory, and it ceases to exist as soon as the database connection is closed. The major advantage of in-memory databases is that their internal optimization algorithms are simpler and execute fewer CPU instructions, thereby facilitating faster response times than disk-optimized databases.

So, instead of using the command line specified in Listing 6, we will use the following one:

```
conn = sqlite3.connect(":memory:")
```

Listing 7: SQL database in-memory connection

After testing the tool with this configuration, we conclude that the speed performance improvement is not noticeable, since the timing does not barely vary from using an ordinary disk database.

7. Market Analysis

During the forecast period of 2021-2026, voice cloning market is expected to grow at 17.2% CAGR (Compound Annual Growth Rate) [38]. In fact, enterprises are focusing each time more on enhancing their customer experience by introducing a familiar voice on the products and services they provide. Proof of that are the technology providers that are developing cutting-edge technologies for creating efficient voice cloning solutions. On that way, these businesses can form significant long-term relationships with customers by providing them a much better customer experience.

The market scope of voice cloning pursues the adoption of different voice cloning solutions and services used by several end-user verticals such as IT & telecommunication, media & entertainment, banking & insurance industry, educational sector, healthcare…

When it comes to this project, it may be a first step towards building a marketable product in the future that applies voice cloning for enhancing user experience in the educational field. For this reason, we thought that sizing the market is a necessary task for any startup's business planning. The starting point for estimating market size is to understand the problem you solve for customers and the potential value your product generates for them. In this case, considering the purpose of the tool we have built and that at the time of the COVID-19 pandemic the requirement of digital education platforms is increasing, our main focus would be to target Learning Management System (LMS) vendors.

A LMS is an online system or software that is used to plan, execute, and assess specific learning process. In other words, it is the software used in eLearning programs and which helps in administration, documentation, tracking and recording [39]. Its main purpose is to enhance the learning process and maintain online collaboration over the internet.

The first introduction of the LMS was in late 1990s and have evolved over time, constantly morphing to fit the current demands and trends [40]. A robust LMS is needed in almost every industry, but the most obvious is the education sector, but it is also required in corporate world, such as, IT and telecom, healthcare, manufacturing, government, or software development, among others [41]. The following graphic (extracted from [41]) shows the LMS usage by industry:

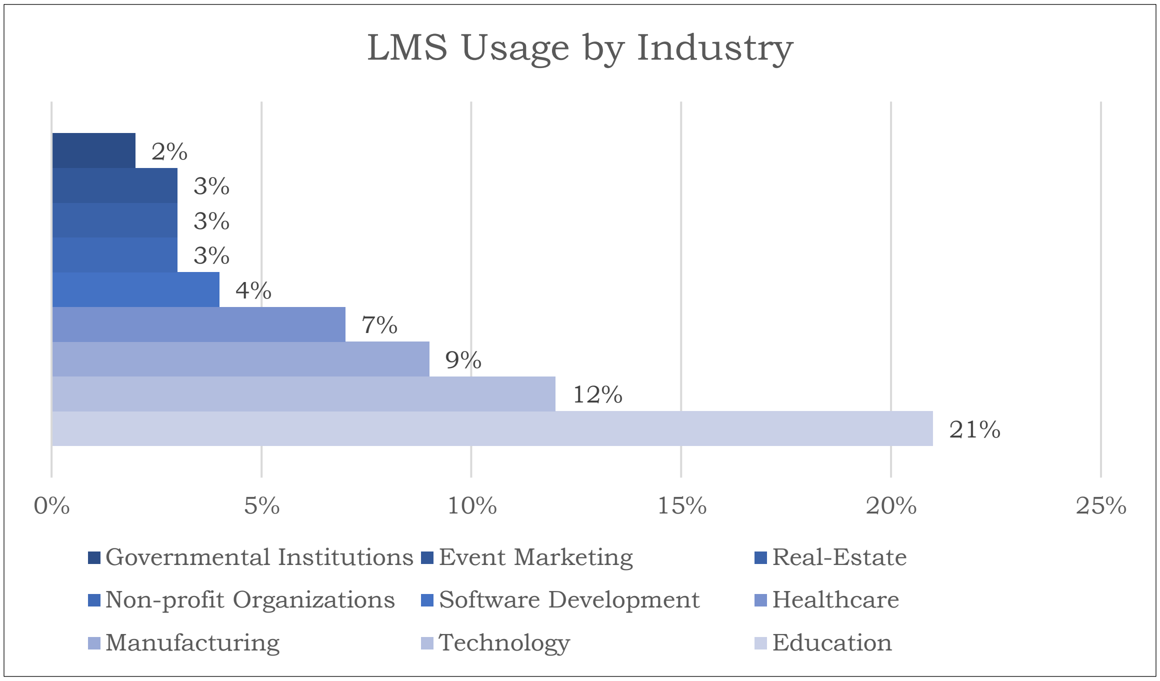


Figure 31: LMS usage by industry

eLearning has experienced rapid growth in recent years and is expected to grow even more due to the COVID-19 ongoing pandemic. In consequence, LMS technologies are gaining traction in the eLearning market as the demand for distance learning and quality education increases on a global scale. With LMS platforms enabling the delivery, tracking, and management of eLearning content, it comes as no surprise that the market is seeing positive growth today.

As far back as 30 years ago, there were around 15 LMS vendors on the market; today the LMS market size is about 700, which is expected to be among the biggest in the entire technology industry in the immediate future, becoming more exciting and intense in terms of competition and upcoming technological breakthroughs. The LMS market is expected to grow from $9.2 billion in 2018 to $22.4 billion in 2023 at a CAGR of 19.6% [42]. Therefore, the market will be probably assured in the coming years.

Looking at the broad LMS market, the idea would be to start by getting into the educational sector by creating a demo based on this project. When it comes to LMSs who dominate the market, Canvas and Blackboard come across in US and Canada higher education. We present the following data (extracted from [43]) by institutions, with market share as a percentage of the total number of institutions using each LMS as a primary system. With the current data, Canvas leads with 32% of US & Canadian higher education institutions, followed by Blackboard at 23%, Moodle at 22%, and D2L at 13%. [43]

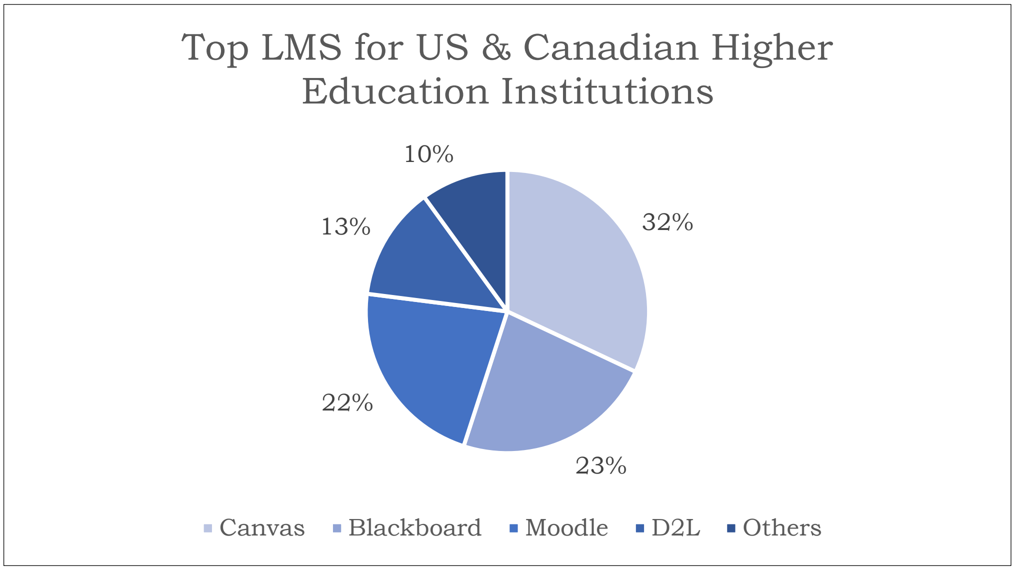


Figure 32: Top LMS for US & Canadian higher education institutions

Primary and secondary education teachers have proven most receptive to "LMS-lite" platforms such as Google Classroom, Schoology, Edmodo, and Quizlet. Each of these platforms can be used to track and assign online work associated with blended learning, which sometimes correlates with improved student learning outcomes. These platforms are less feature-rich than the LMSs that dominate higher education. Instead, these services tend to be better tailored to end users (teachers, students, parents) in terms of both cost and ease of use.

But despite this increasingly market growth, there is still room for improvement. Of all the issues people have with their LMS, the most cited include poor user experience; that is where our project comes into play. By using voice cloning technology, we aim to create a smoother, more realistic, and convincible experience for the users.

So, trying to target these education focused LMSs would be a good starting point in an early stage, since as seen in Figure 31, the education sector comprises the 21% of LMS usage. In the future, an expansion to the office environment could also be considered, as technology companies are the second industry in LMS usage.

8. Discussion

8.1 Voice Cloning Technology Misuse

Although the concept of voice cloning is fascinating and has many benefits, we cannot deny the fact that this technology is susceptible to misuse. In the past few years, we have seen how Deepfakes have been used to spread misinformation and to create questionable content [10].

As the voice cloning algorithms are getting better, it is becoming more and more difficult to discern what is real and what is not [10]. Indeed, according to research, the human brain does not register significant differences between real and artificial voices [10]. In fact, it is harder for our brains to distinguish fake voices than to detect fake images [10]. So, this leads to using the mentioned technology inappropriately creating several issues related to:

- **Trust**. It is not hard to think of reasons to be terrified of a technology that can potentially make anyone appear to be saying or doing anything [13]. Things that people never uttered could be pushed on the internet in a planned manner for political gains or to create unrest in society, for example. Some deepfakes have been used for a myriad of purposes such as bullying, revenge pornography, video manipulation, audio manipulation, and extortion which definitely harm individuals' reputations [44].

  In 2017, researchers at the University of Washington shined a light on the potential pitfalls of this type of technology when they released a paper describing how they had created a fake video of President Barack Obama [13]. Google's chief executive, Mark Zuckerberg, has also been the target of a deepfake video that appeared to show him credit a secretive organization for the success of the social network [13]. YouTube channel Vocal Synthesis features well-known people saying things they never said, like George W. Bush reading "In Da Club" by 50 Cent [4].

  Elsewhere on YouTube, you can hear a group of ex-Presidents, including Obama, Clinton, and Reagan, rapping [4]. The music and background sounds help disguise some of the obvious roboticness, but the potential is obvious.

- **Scamming**. Financial scams are another area of concern [13]. It will become easier for scammers to perform phishing and spoofing attacks [10]. Indeed, audio deepfakes have already been used to clone voices and convince people they are talking to someone trusted and defraud them [13]. In 2019, a company in the U.K. claimed it was tricked by an audio deepfake phone call into wiring money to criminals [4]. Last year, scammers used a deepfake of a tech company's CEOs voice to try and convince an employee to transfer money to the scammer's account [13].

8.1.1 Ethical implications

Apart from being various legal problems that arise with the development of such technology, there can also be ethical implications involved. One of the biggest problems that comes with the use of Deepfakes is identity theft. Identity theft consists of someone wrongfully obtaining and using another person's personal data in some way that involves fraud or deception [45]. Apart from causing financial damages, identity theft is also often used for psychological and emotional harm, making it more difficult to provide a remedy.

Another ethical problem could be the privacy of personal information and data that is provided when this technology is used. When creating Deepfakes or cloned voices, deep-learning algorithms are applied, where the more information the algorithm receives, the better the provided results are [46]. Therefore, the risk of having a data breach increases, and unavoidably, every platform can suffer from a privacy violation, which could potentially lead to personal information being accesses by people that are not consented to. Moreover, deceased people’s privacy also comes into question, since in the past few years, the image of those who have left us have been used for commercial purposes, for example. This has opened up some concerns about the morality of the power of resurrection [47].

8.1.2 Safeguards and mitigation

As mentioned earlier, the human brain is not able to distinguish between real and artificial voices. So, the first step toward safeguarding will be raising the awareness that this technology exists. Algorithms that can differentiate real voices from artificial voices should be developed [10]. Indeed, detection solutions for cloned voices are starting to emerge. For instance, Resemble released an artificial intelligence tool that detects deepfakes by deriving high-level representations of voice samples and predicting if they are real or not [48]. Companies like ID R&D have also published algorithms that look at the different features of a voice (frequency, tone, prosody…) in order to determine if it is coming from a human vocal tract or from a reproduction device [49].

In regard to people’s capability to recognize deepfakes, a study by MIT Media Lab and Max Planck Institute researchers showed that people improve in their ability to detect fakes when they get feedback [50]. Therefore, scientists believe that it is possible that consumers will learn to catch deepfakes on their own after repeated exposure [50].

Deepfake prevention has also been taken into account by several social media networks. Twitter, for example, announced plans to implement policy around media that has been altered and that it will delete any media that it is threatening someone’s public safety or leads to serious harm [50]. Facebook has also said it would remove anything that has been modified in ways that “aren’t apparent to the average person.” [50]

8.1.3 Regulation

Legislation and regulation about voice cloning and deepfakes are both critical pieces of this technological puzzle.

Voice Protection

Voice is a very personal aspect of who we are and often a unique identifier. Indeed, it can also be very valuable, with some voice actors being incredibly memorable for their characters [51]. But do we have rights in our own voice?

The starting point would be protection under copyright. Copyright is "*a type of intellectual property that protects original works of authorship as soon as an author fixes the work in a tangible form of expression*" [52]. Therefore, a song, an advertisement or a movie may be copyrightable, and voice may get protected if it is a part of the tangible medium [53]. However, copyright protection is not available specifically for voice per se [53]. Indeed, in the United States, the main indicators to apply for a copyright appear to be originality and creativity, although the extent of them is not clearly defined [52]. So, copyright protects digital models when they include elements of creativity (color, animation, lighting…), but it does not protect the basic models themselves, such as, someone's voice itself [54].

When it comes to trademarks, a voice cannot be trademarked as it does not fall under the United States Patent and Trademark Office's criteria for material that can be trademarked. A trademark is "*any word, name, symbol, or design, or any combination thereof, used in commerce to identify and distinguish the goods of one manufacturer or seller from those of another and to indicate the source of the goods*" [55]. It would be impossible to place restrictions on the voices of competitors since a voice is considered to be a distinct feature of an individual, which they have no control over [56].

Even though the law on copyright or trademarks may not protect someone's voice, law recognizes the right of publicity in most states of the US [57]. Publicity rights, also known as, personal rights, protect against unauthorized commercial use of one's name, image, voice, signature, likeness, or other personal identifying traits that are unique to someone [58]. This gives an individual the exclusive right to license the use of their identity for commercial promotion [59]. States diverge on whether the right survives posthumously and, if so, for how long [57]. So, it is possible to take legal actions by claiming that the right of publicity is being violated if someone's voice is used without any consent, for example, making statements that are out of their personality [58].

A good example of this is the Midler v. Ford Motor Co. (1988) case. Singer Bette Midler sought against Ford Motor Company regarding a series of commercials where Midler's distinctive voice was used without her authorization [51]. Ford hired a singer and asked her to sound as much like Midler as possible. Although the Court of Appeal stated that "*A voice is not copyrightable. The sounds are not 'fixed'*", they did afford her rights under common law for appropriation of identity and right of publicity [51] since her recognizable voice was used for another's commercial gain. They highlighted that "*A voice is as distinctive and personal as a face. The human voice is one of the most palpable ways identity is manifested.*" [58]

Spain's constitution also offers extraordinarily high protection in the realm of personality rights [60]. It guarantees three fundamental rights: the right to protect someone's image, the right to privacy and the right to honor [61]. This prohibits the dissemination of the name, voice, or image of a person without their consent. This includes their use for advertising, commercial or financial gain [61]. However, it does permit actions in which there is a "predominant and relevant historical, scientific, or cultural interest" [60].

Apart from that, there is a regulation in the European Union (General Data Protection Regulation) that covers all matters related to personal image or voice rights and their use, which provides penalties for those who use someone's image or voice for advertising purposes without their permission [62].

Deepfake Legislation

Yet as deepfakes become more realistic and accessible, concern about the potential harm they pose has increased [50]. Hence, there is a need for action to be taken to address the issues created by this type of technology.

In the US, though the available legal remedies concerning deepfakes are still in their early stages, both state and federal legislators have already enacted laws specifically aimed at deepfakes.

In July 2019, Virginia became the first state in the nation to impose criminal penalties on the distribution of nonconsensual deepfake pornography, making it punishable by up to a year in jail and a fine of $2,500 [63].

Two months later, Texas prohibited the creation and distribution of deepfake videos intended to harm candidates for public office or influence elections, punishable by up to a year in jail and a fine of $4,000 [63].

In October 2019, California enacted two laws, allowing victims of nonconsensual deepfake pornography to sue for damages and give candidates the ability to sue individuals or organizations that distribute election-related deepfakes without warning labels near Election Day [63].

On December 20, 2019, the U.S. Congress signed the nation's first federal law related to "deepfakes." The deepfake legislation is part of the National Defense Authorization Act for Fiscal Year 2020 (NDAA). The NDAA requires a comprehensive report on the foreign weaponization of deepfakes and the government to notify Congress of foreign deepfake-disinformation activities targeting US elections. It also establishes a "Deepfakes Prize" competition to encourage the research or commercialization of deepfake-detection technologies [44].

Intelligence, Law Enforcement, and other Governmental agencies are taking active roles against the threat of misinformation and are spending growing budgets on tools that can detect fake news and deepfake media [64].

In Spain, in the contrary, deepfakes are not specifically regulated, since crimes are never classified by the used technology, but considering the intention with which are committed, and the damaged legal good [65]. Indeed, the European Parliament has drawn up three different regulations: one that regulates the ethics of this technology, another one focused on civil liability for the damages it causes and the last one related to intellectual and industrial property rights [66].

Therefore, it is not the same to use cloned voices to create an ad with permissions or to humiliate and send false messages [65]. Depending on the video, we could be facing a crime of insult, slander or against moral integrity [65].

Past years' experience shows that legislation in this area is changing rapidly as new and emerging deepfake-related threats to national security, individuals, and businesses are arising. Indeed, since artificial intelligence technology continues to evolve day by day, the law must progress as well.

8.2 Future

It is undeniable that this technology will get better with time. Systems will need less audio samples to create a model, and there will be faster process to build the model in real time. Smarter Artificial Intelligence technologies will learn how to ensure more convincing human-like speech [4].

Consequently, many people will each time be more skeptical if humans should even try creating such models and some researchers have refrained from sharing their findings publicly [10].

However, there is people who think quite the opposite and defend the benefits of such technology, supporting what U.S. Department of Justice attorney Mona Sedky says: "*Just like the internet can be weaponized against people, it doesn't mean we shouldn't have the internet. It just means that these are things that we need to be thinking about, and … that will make it harder to weaponize against people. We need to be upfront about how we're going to protect consumers who are definitely going to be victimized in ways by criminals.*" [48]

Therefore, transparency and an abundance of caution will be the keys to grappling with voice cloning technologies in the years to come [48].

So, unquestionably, a huge debate about voice cloning and deepfake technology is booming lately; the future appears to be uncertain, as to how this technology will be used. Dystopia or utopia?

9. Conclusion

This project provides a voice query assistant including a voice cloning feature that is able to answer to requests about information extracted from a syllabus of a course. Apart from that, it also presents a starting point of a qualitative research study about the impact of the developed framework on students and professors by means of a step-by-step design. So, future investigations are necessary to complete the conclusions that can be drawn from this case study.

After achieving the proposed goal of building the tool, we can state that the results are satisfactory and along with some improvements, it may be a kickoff towards a marketable product in the following years. In fact, Learning Management Systems and eLearning continue to gain steam and the market is increasingly propelling demand for high quality and excellent user experience emerging technologies.

Apart from that, we believe that even more powerful forms of voice cloning will become available in a near future. Therefore, this type of applications will continue to develop in a massive way and cloned voices may soon be indistinguishable from the original voice. Indeed, one of the biggest innovations has been the overall reduction in how much raw data is needed to create a voice. In the past, hundreds of hours of audio were required to get passable results; now, however, being this project a proof of it, cloned voices can be generated from just a few seconds of reference audio.

The increasing sophistication of voice cloning not only has clear commercial potential, but also raises growing concerns that it could be misused, for instance, to trick people, by means of voice deepfakes. This leads to people thinking voice cloning could be a threat to their privacy and fear that it could be exploited to perpetuate scams, such as, identity theft. Consequently, it is essential to raise awareness about the technology and its usage and in the future, an accurate and updated regulation will be needed for both encouraging voice cloning and managing associated risks.

10. Bibliography

[1] AI Oodles. (2020, July 7). *Copy That: Realistic Voice Cloning with Artificial Intelligence.* https://artificialintelligence.oodles.io/blogs/voice-cloning-with-artificial-intelligence/.

[2] Seif, G. (2021, February 14). *You can now speak using someone else's voice with Deep Learning*. Medium. https://towardsdatascience.com/you-can-now-speak-using-someone-elses-voice-with-deep-learning-8be24368fa2b.

[3] Napolitano, D. (2020). The Cultural Origins of Voice Cloning. *Proceedings of the Eighth Conference on Computation, Communication, Aesthetics.* https://www.researchgate.net/publication/342924151_The_Cultural_Origins_of_Voice_Cloning

[4] Johnson, D. (2020, July 31). *Audio Deepfakes: Can Anyone Tell If They're Fake?* https://www.howtogeek.com/682865/audio-deepfakes-can-anyone-tell-if-they-are-fake/.

[5] Stimpson, R. (2020, February 5). *"Don't Believe Everything You Hear" Has Taken on a New Meaning.* The Bull &amp; Bear. http://bullandbearmcgill.com/dont-believe-everything-you-hear-has-taken-on-a-new-meaning/.

[6] Nafea, I. T. (2018). Machine Learning in Educational Technology. *Machine Learning - Advanced Techniques and Emerging Applications.* https://doi.org/10.5772/intechopen.72906

[7] Seif, G. (2021, February 14). *You can now speak using someone else's voice with Deep Learning*. Medium. https://towardsdatascience.com/you-can-now-speak-using-someone-elses-voice-with-deep-learning-8be24368fa2b.

[8] *Speech synthesis*. (n.d.). https://www.cs.mcgill.ca/~rwest/wikispeedia/wpcd/wp/s/Speech_synthesis.htm#:~:text=The%20first%20computer%2Dbased%20speech,the%20history%20of%20Bell%20Labs.

[9] *History and Development of Speech Synthesis*. (n.d.). http://research.spa.aalto.fi/publications/theses/lemmetty_mst/chap2.html.

[10] Saini, M. (2020, February 6). *Voice Cloning Using Deep Learning*. Medium. https://medium.com/the-research-nest/voice-cloning-using-deep-learning-166f1b8d8595.

[11] Vicomtech. (2021, February 24). *Voice Cloning*. Speech and Language Solutions. https://www.speechandlanguagesolutions.com/voice-cloning/.

[12] Jemine, C. (2019, June 25). *Master thesis: Real-Time Voice Cloning.* https://matheo.uliege.be/handle/2268.2/6801.

[13] Springwise. (2020, October 8). *Pros and Cons: Deepfake technology and AI avatars.* https://www.springwise.com/pros-cons/deepfake-technology-ai-avatars.

[14] Mutchler, A. (2021, March 26). *Voice Assistant Timeline: A Short History of the Voice Revolution.* Voicebot.ai. https://voicebot.ai/2017/07/14/timeline-voice-assistants-short-history-voice-revolution/.

[15] Blackboard. (n.d.). *Alexa Education Skill Integration.* https://help.blackboard.com/Learn/Administrator/SaaS/Integrations/Alexa_Education_Skill.

[16] Cerence. (2019, December 30). *Cerence Introduces My Car, My Voice – New Voice Clone Solution to Personalize the In-Car Voice Assistant.* https://www.cerence.com/news-releases/news-release-details/cerence-introduces-my-car-my-voice-new-voice-clone-solution/.

[17] C++ vs Java vs Python. (n.d.). https://www.tutorialspoint.com/cplusplus-vs-java-vs-python.

[18] Isaacs, D. (2010). *A comparison of the network speech recognition and distributed speech recognition systems and their effect on speech enabling mobile devices.* Speech Technology and Research Group, University of Cape Town. https://open.uct.ac.za/handle/11427/11232

[19] PyPI. (2017, December 5). *SpeechRecognition.* https://pypi.org/project/SpeechRecognition/.

[20] *PyAudio: PortAudio v19 Python Bindings. (*n.d.). http://people.csail.mit.edu/hubert/pyaudio/#downloads.

[21] Këpuska, V. & Bohouta, G. (2017). Comparing Speech Recognition Systems (Microsoft API, Google API and CMU Sphinx). *International Journal of Engineering Research and Applications,* 07(03), 20–24. https://doi.org/10.9790/9622-0703022024

[22] AltexSoft. (2019, October 15). *Comparing Database Management Systems: MySQL, PostgreSQL, MSSQL Server, MongoDB, Elasticsearch and others.* https://www.altexsoft.com/blog/business/comparing-database-management-systems-mysql-postgresql-mssql-server-mongodb-elasticsearch-and-others/.

[23] Drake, M. (2019, March 19). *SQLite vs MySQL vs PostgreSQL: A Comparison of Relational Database Management Systems*. DigitalOcean. https://www.digitalocean.com/community/tutorials/sqlite-vs-mysql-vs-postgresql-a-comparison-of-relational-database-management-systems.

[24] Wang, Y. (2017, August 10). *SQLite vs MySQL vs PostgreSQL.* Medium. https://medium.com/@yangforbig/sqlite-vs-mysql-vs-postgresql-a-comparison-of-relational-database-management-systems-afd5afd6566

[25] Jemine, C. (2019, June 25). *Real-Time-Voice-Cloning.* GitHub. https://github.com/CorentinJ/Real-Time-Voice-Cloning.

[26] Jia, Y., Zhang, Y., Weiss, R. J., Wang, Q., Shen, J., Ren, F., Chen, Z., Nguyen, P., Pang, R., Lopez Moreno, I. & Wu, Y. (2018). Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis. *CoRR, abs/1806.04558.* https://arxiv.org/pdf/1806.04558.pdf

[27] Wan, L., Wang, Q., Papir, A., & Lopez Moreno, I. (2018). Generalized End-to-End Loss for Speaker Verification. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* https://doi.org/10.1109/icassp.2018.8462665

[28] Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerrv-Ryan, R., Saurous, R. A., Agiomvrgiannakis, Y., & Wu, Y. (2018). Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* https://doi.org/10.1109/icassp.2018.8461368

[29] Van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., & Kavukcuoglu, K. (2016). Wavenet: A generative model for raw audio. *CoRR, abs/1609.03499.* http://arxiv.org/abs/1609.03499.

[30] Amos, D. (2021, July 13). *The Ultimate Guide To Speech Recognition With Python.* https://realpython.com/python-speech-recognition/#how-speech-recognition-works-an-overview.

[31] Heigold, G., Loper Moreno, I., Bengio, S. & Shazeer, N. (2015). End-to-end text dependent speaker verification. *CoRR, abs/1509.08062.* http://arxiv.org/abs/1509.08062.

[32] Rayhane. (2019, January 26). *Tacotron-2*. GitHub. https://github.com/Rayhane-mamah/Tacotron-2.

[33] Kalchbrenner, N., Elsen, E., Simonyan, K., Noury, S., Casagrande, N., Lockhart, E., Stimberg, F., van den Oord, A., Dieleman, S. & Kavukcuoglu, K. (2018). Efficient neural audio synthesis. https://arxiv.org/abs/1802.08435

[34] MIT. (2019). *WaveRNN*. GitHub. https://github.com/fatchord/WaveRNN.

[35] Resemble-AI. (2020). *Resemblyzer*. GitHub. https://github.com/resemble-ai/Resemblyzer.

[36] Bhandari, P. (2020, July 30). *What is Qualitative Research?.* Scribbr. https://www.scribbr.com/methodology/qualitative-research/#:~:text=Qualitative%20research%20involves%20collecting%20and,concepts%2C%20opinions%2C%20or%20experiences.&amp;text=Qualitative%20research%20is%20commonly%20used,health%20sciences%2C%20history%2C%20etc.

[37] Illinois Institute of Technology. (2020). *Protocol for Bringing Human Subjects to IIT (Illinois Tech) for Research Studies.* https://research.iit.edu/sites/research/files/elements/OSRP/pdfs/IRB%20and%20Human%20Subjects%20COVID.pdf

[38] MarketWatch. (2021, April 21). *Voice Cloning Market Size, Growth, Trends and Segments Analysis Report and Forecast to 2025*. https://www.marketwatch.com/press-release/voice-cloning-market-size-growth-trends-and-segments-analysis-report-and-forecast-to-2025-2021-04-21.

[39] Sharma, A. (2021, May 12). *Discovering Learning Management Systems: Basic Functions and Benefits.* eLearning Industry. https://elearningindustry.com/discovering-learning-management-systems-basic-functions-benefits.

[40] *The Evolution of the LMS: From Management to Learning.* Research Library | The Learning Guild. (n.d.). https://www.learningguild.com/insights/137/the-evolution-of-the-lms-from-management-to-learning/.

[41] Ferriman, J. (2020, January 14). *LMS Industry Snapshot*. LearnDash. https://www.learndash.com/lms-industry-snapshot/.

[42] Bouchrika, I. (2020, September 2). *51 LMS Statistics: 2019/2020 Data, Trends & Predictions*. Guide 2 Research. https://www.guide2research.com/research/lms-statistics.

[43] Hill, P. (2021, February 4). *State of Higher Ed LMS Market for US and Canada: Year-End 2020 Edition.* PhilOnEdTech. https://philonedtech.com/state-of-higher-ed-lms-market-for-us-and-canada-year-end-2020-edition/#comments.

[44] Vazquez, L. (n.d.). *RECOMMENDATIONS FOR REGULATION OF DEEPFAKES IN THE U.S.: DEEPFAKE LAWS SHOULD PROTECT EVERYONE NOT ONLY PUBLIC FIGURES.* https://www.ebglaw.com/content/uploads/2021/04/Reif-Fellowship-2021-Essay-2-Recommendation-for-Deepfake-Law.pdf.

[45] The United States Department of Justice. (2020, November 16). *Identity Theft.* https://www.justice.gov/criminal-fraud/identity-theft/identity-theft-and-identity-fraud#:~:text=What%20Are%20Identity%20Theft%20and,deception%2C%20typically%20for%20economic%20gain.

[46] Dang, L., Hassan, S., Im, S., Lee, J., Lee, S., & Moon, H. (2018). Deep Learning Based Computer-Generated Face Identification Using Convolutional Neural Network. *Applied Sciences*, 8(12), 2610. https://doi.org/10.3390/app8122610

[47] Savin-Baden, M., & Burden, D. (2018). Digital Immortality and Virtual Humans. *Postdigital Science and Education*, 1(1), 87–103. https://doi.org/10.1007/s42438-018-0007-6

[48] Gadney, G. (2021, July 8). *Clone Synthetic AI Voices with Neural Text to Speech.* Resemble AI. https://www.resemble.ai/.

[49] ID R&D. (n.d.). *Voice Anti-Spoofing: ID R&D Voice Spoof Detection.* https://www.idrnd.ai/voice-anti-spoofing/.

[50] Wiggers, K. (2020, January 30). *Voice cloning experts cover crime, positive use cases, and safeguards*. VentureBeat. https://venturebeat.com/2020/01/29/ftc-voice-cloning-seminar-crime-use-cases-safeguards-ai-machine-learning/.

[51] Ihalainen, J. (1970, January 1). *That Sounds Good - Do You Have IP Rights in Your Own Voice?* https://www.ipiustitia.com/2018/01/that-sounds-good-do-you-have-ip-rights.html.

[52] Office, U. S. C. (n.d.). *What is Copyright?* U.S. Copyright Office. https://www.copyright.gov/what-is-copyright/.

[53] Gupta, A. (2010, December 27). *When celebrities seek copyrights*. The Financial Express. https://www.financialexpress.com/archive/when-celebrities-seek-copyrights/729569/.

[54] Newell, B. (2010, June 22). "*Independent Creation and Originality in the Age of Imitated Reality: A Comparative Analysis of Copyright and Database Protection for Digital Models of Real People*". Brigham Young University International Law & Management. 6 (2): 93–126. https://digitalcommons.law.byu.edu/cgi/viewcontent.cgi?article=1078&context=ilmr.

[55] Legal Information Institute. (n.d.). *Trademark.* https://www.law.cornell.edu/wex/trademark.

[56] *Can You Trademark a Voice? Secure Your Trademark.* (2020, March 16). https://secureyourtrademark.com/can-you-trademark/trademark-a-voice/.

[57] International Trademark Association. (n.d.). *Right of Publicity.* https://www.inta.org/topics/right-of-publicity/.

[58] Attenborough, I. (2020, October 13). *Voices, Copyrighting and Deepfakes.* IPWatchdog.com | Patents & Patent Law. https://www.ipwatchdog.com/2020/10/14/voices-copyrighting-deepfakes/id=126232/.

[59] Legal Information Institute. (n.d.). *Publicity.* https://www.law.cornell.edu/wex/publicity.

[60] Schulman, A. (2020, December 16). *Image Rights, Personality Rights and the Right of Publicity in The US and the EU.* Jayaram Law: Jayaram Law Group, Ltd. https://www.jayaramlaw.com/blog/2019/07/image-rights-personality-rights-and-the-right-of-publicity-in-the-us-and-the-eu/.

[61] *Derecho a la privacidad en España. Guía 2021*. (2021, July 14). https://ayudaleyprotecciondatos.es/2018/11/13/derecho-privacidad-espana/

[62] *REGLAMENTO (UE) 2016/679 DEL PARLAMENTO EUROPEO Y DEL CONSEJO.* (2016, April 27). https://www.boe.es/doue/2016/119/L00001-00088.pdf.

[63] Chipman, J., Ferraro, M., & Preston, S. (2019, December 24). *First Federal Legislation on Deepfakes Signed into Law*. JD Supra. https://www.jdsupra.com/legalnews/first-federal-legislation-on-deepfakes-42346/.

[64] Wood, L. (2021, March 11). *Global Counter Misinformation (DeepFake & Fake News) Solutions Market 2020-2026: Global Security Concerns & Assessment.* Cision. https://www.prnewswire.com/news-releases/global-counter-misinformation-deepfake--fake-news-solutions-market-2020-2026-global-security-concerns--assessment-301245886.html.

[65] Merino, M. (2019, July 3). *En Estados Unidos están empezando a legislar contra los 'deepfakes', y así está la normativa al respecto en España*. Xataka. https://www.xataka.com/inteligencia-artificial/estados-unidos-estan-empezando-a-legislar-deepfakes-asi-esta-normativa-al-respecto-espana.

[66] Pachón, M., & Aguiar, A. R. (2021, March 2). *Voces clonadas y máquinas que conversan*. Business Insider España. https://www.businessinsider.es/desafios-eticos-legales-ia-confundiremos-humanos-818639.