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

Naroa Amezaga*
College of Computing, Illinois Institute of Technology
namezagavelez@hawk.iit.edu

Jeremy Hajek
College of Computing, Illinois Institute of Technology
hajek@iit.edu

ABSTRACT

Artificial Intelligence and especially 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 [1]. But currently, there is a fundamental common drawback: unnatural, robotic and impersonal synthesized voices [2].

So, what happens when the robotic computer voice no longer sounds like a computer, but sounds like you? That's where Voice Cloning technology comes into play, which allows one to generate an artificial speech that resembles a targeted human voice. This new practice offers many benefits, but with its development, the generation of fake voices and videos, known as "deepfakes", has risen, causing a loss of trust and greater fear towards technology [3].

In this way, the objective of this paper is to analyze the availability of voice deepfake technologies, its ease of construction and its impact for good and evil. We chose to focus on the educational field by implementing a "deepfake professor" via a survey of readily available voice deepfake technologies. The goal is then to demonstrate the potential capabilities for good and for evil that need to be considered with this technology, so we also conduct an analysis about the misuse, the current regulation, and the future of it.

The results of the case study show that it is possible to clone someone's voice with a standard laptop, with no need of high-performance computing resources and 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 that future generations will have to deal with. There is no doubt that to understand the dynamics and impact of voice cloning and to reach more solid conclusions, future research is needed.

*https://github.com/naroaamezaga/VOICE-QUERY-ASSISTANT-with-SV2TTS.git

SIGITE '22, September 21–24, 2022, Chicago, IL, USA

ACM ISBN 978-1-4503-9391-1/22/09.
https://doi.org/10.1145/3537674.3554742

CCS CONCEPTS

• **Computing methodologies** → Artificial intelligence; Natural language processing.; • **Human-centered computing** → Human computer interaction (HCI); Interaction paradigms; Natural language interfaces.; • **Applied computing** → Education; Interactive learning environments.; • **Social and professional topics.**;

KEYWORDS

Speech-To-Text, SQL query, Text-To-Speech, Speech Synthesis, Voice Cloning, Deepfake, Virtual Assistant, SV2TTS, Speech Recognition

ACM Reference Format:
Naroa Amezaga and Jeremy Hajek. 2022. Availability of Voice Deepfake Technology and its Impact for Good and Evil. In *The 23rd Annual Conference on Information Technology Education (SIGITE '22), September 21–24, 2022, Chicago, IL, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3537674.3554742

1 INTRODUCTION

With the recent 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 [4]. 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 [1]. This technology consists of converting text into artificial human speech and has been utilized to enhance a wide range of applications, like chatbots or virtual assistants [5].

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 one to generate someone's cloned natural-sounding audio samples based on a reference voice [6]. However, the advancement of such technologies has led to the development of techniques for manipulation of video and audio, known as a "deepfake" [7]. In the past, the creation of this type of content was available just to specialists, but nowadays this has changed, which causes a reluctance towards using this technology [3]. Therefore, an assessment about the current availability of voice deepfake technology is needed, along with an analysis about its impact for good and evil.

The structure of this paper goes as follows: we begin with a review of the background and state of the art in section 2 for a better understanding of the topic. We then present the project goal and research questions in section 3. Section 4 describes the chosen tools and the methodology of the whole process. We later demonstrate the results of the work in section 5, and we conclude with a discussion and conclusion in section 6 and section 7, respectively.

2 BACKGROUND

2.1 Speech Synthesis – Text-To-Speech

Speech Synthesis or Text-to-Speech (TTS) is the computer-generated simulation of human speech and refers to the artificial transformation of text to audio. The goal of a good TTS system is to have a computer do it, considering the naturalness and expressiveness of the voice [5]. 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 [5].

A computer system used for this purpose is called a speech synthesizer and can be implemented in software or hardware [9]. The quality of a speech synthesizer is judged by its similarity to the naturalness of a human voice (naturalness) and by its ability to be understood (intelligibility) [9].

2.1.1 History. Before electronic signal processing was invented, speech researchers tried to build mechanical machines to create human speech. In St. Petersburg 1779, the scientist Christian Kratzenstein, explained the differences and built models of the five long vowels [10]. This was followed by von Kempelen of Vienna, in 1791, who added models of the tongue and lips, enabling the production of consonants as well as vowels [9].

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 [10]. In the 1930s, Bell Laboratories developed the VOCODER, an electronic speech analyzer and synthesizer [9], then refined into the VODER. After this, the scientific world became more interested in speech synthesis since it was finally shown that intelligible speech could be produced artificially [10].

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", which was used in the climactic scene of Arthur C. Clarke's screenplay *2001: A Space Odyssey*. [9]

In late 1970's and early 1980's, a considerable number of commercial speech synthesis products were introduced and many computer operating systems have included speech synthesizers since then [10]. 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 [9]. The second operating system with advanced speech synthesis capabilities was AmigaOS, introduced in 1985, with both male and female voices [9]. Speech systems were first available on Microsoft-based operating systems in Windows 95 and Windows 98 [9].

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 [11]. 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. It takes advantage of a set of audios of the original voice to train a model capable of generating new audios that sound alike [12].

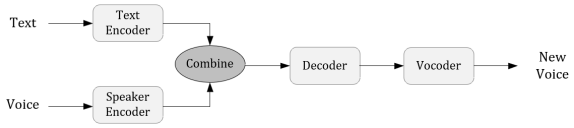

Figure 1: Voice Cloning technology architecture.

Nowadays, due to its interesting and varied applications, voice cloning is being increasingly demanded by the market [12]. Indeed, it helps automate and personalize many tasks carried out in several types of applications, using specific or favorite voices to develop customized and fully personalized conversational assistants [12]. While voice cloning technology has developed, audio deepfakes have come hand in hand with it, becoming each time more and more popular. 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 [7]. So, audio deepfakes might mean you can no longer trust your ears.

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 to be read and a sample of the voice we want to clone [5]. This is shown in Figure 1.

One of the biggest advancements in voice cloning has been the reduction of how much raw data is needed to clone a voice. Not long ago, developers needed enormous quantities of recorded voice audios 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, generating competent voices from just minutes of reference content. [12]

2.3 Current Technologies

There are several applications and technology examples that use Speech Synthesis and/or Voice Cloning nowadays. IBM became the first to introduce a voice assistant with its Shoebox device in 1961. Then, Microsoft's text-based virtual assistant, Clippy, showed how natural language in text could be used for interactive feedback.

After that, 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 [13], whose popularity played a crucial role in developing other skills, like The Blackboard Alexa skill. This allows users to request and receive information about homework and assignments via an Alexa-device, rather than having to log into Blackboard to look the information up [14].

Cerence Inc. introduced My Car, My Voice, in December 2019, a revolutionary product that lets people create custom voices for their in-car assistants. 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. [15]

Table 1: Summary of used equipment and tools

Equipment/Tool	Selection
Headset with microphone	Logitech H390 USB Computer Headset
Programming language	Python 3.7 [16]
Speech-To-Text	SpeechRecognition (Google Speech Recognition) [17]
Database Management System	SQLite [18]
Text-To-Speech with Voice Cloning	SV2TTS [19] [20]

3 RESEARCH QUESTIONS AND STUDY GOALS

Within this umbrella, the primary research question of this project was: "Is Voice Deepfake Technology available for anyone?", meaning "Can anyone have access to cloning a voice?" and immediately, we came up with the second one: "What impact does this technology have on society?"

To answer the first research question, our aim was to check if anyone can have access to clone a voice and how easy is that nowadays. To this effect, we thought of focusing on the educational field, since technology in education has been proven to be of great importance, for example, to boost the outcome of students [8]. So, the main objective of the project is to build a voice query assistant that allows students to request and receive information appearing in the syllabus of a course (professor's contact information, assignment dates, grading scheme. . .), instead of having to log into the learning management system (e.g., Blackboard) or ask the professor during a lecture. And to enhance user experience and check the easiness of cloning a voice, we came up with the idea of adding the cloned voice of the corresponding professor when answering the requests.

When it comes to the second research question, a discussion and actual point of view about voice cloning and deepfakes will be assessed, where misuse, up-to-date regulation and future will be studied to analyze the impact for good and evil of the mentioned technology.

4 METHODOLOGY

4.1 Summary of used equipment and tools

Table 1 presents the summary of the equipment and tools selected to build the voice query assistant.

4.2 Design Process

This section describes the process to complete the work, which has been divided into four steps. To this end, a Python tool has been built, and as mentioned previously, the aim is to simulate a voice query assistant that is able to answer questions about course information using the proper professor's cloned voice.

The architecture of the designed tool is shown in Figure 2. First, the student's question about a specific course is captured by a microphone and translated into text using Speech-To-Text technology. The answer to the question is obtained by performing an SQL query in a database in which the details about the student's courses are saved. Then, the required information, which is in plain text, is converted into an audio response using Text-To-Speech technology and Voice Cloning. This second feature allows one to reproduce

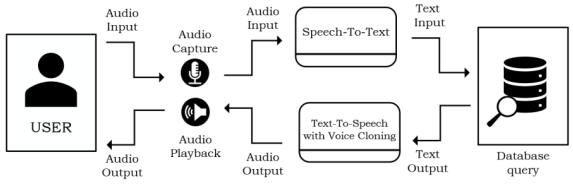

Figure 2: Voice query assistant architecture.

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.

4.2.1 Database definition. The first step is to define the database that will be used during the process. To unify the format of the syllabi, a syllabus template has been created. The information appearing in each syllabus is then passed to a common CSV type file with an automatic script. To know which information is needed, it is necessary to define the possible questions that the user can ask. In our case, we designed a 15-question bank.

Each row of this CSV file will represent a course and each column will define the details about it (semester, professor, classroom, assignment dates, final exam date. . .). Once the CSV file is created, it will be possible to import it into an SQLite table and store it in an SQLite database. Once this is done, we will have our database defined and ready to use.

4.2.2 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 into human readable text. Therefore, the main goal of this step is to capture audio from the microphone and transcribe it into text. We used the Python library named SpeechRecognition with Google Speech Recognition API [17] and PyAudio [21] for our microphone.

So, first, when the program is executed, the user can ask one of the fifteen questions from the question bank. The microphone will listen for ten seconds and will then stop recording. Speech will be converted from physical sound to an electrical signal by means of the microphone. Once the recording is finished, the recognizer will try to recognize the words that appear in the recorded audio by breaking it down into individual sounds and analyzing each of them. It will then try to find the most probable word fit and will transcribe those sounds into text. Finally, after doing the transcription, the recognizer will return a text string containing the sentence that has been recorded.

4.2.3 Database Query. The third step is to perform an SQL query on our database to look for an answer depending on the question that was recorded and transcribed on the previous stage. In our case, since there are fifteen possible questions, there will be fifteen possible answers for each course. To select the corresponding column for each question keywords will be used. Hence, whenever a keyword is detected in a question, a column will be assigned to that request. Then, the complete answer will be built and ready to be reproduced with the corresponding professor's cloned voice in the next step.

4.2.4 Text-To-Speech with Voice Cloning. 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. In addition, apart from just converting the text into speech, we also applied voice cloning technology to clone specific professors' voices and enhance user experience.

The technology that has been used to perform this step is Speaker Verification to Text to Speech Synthesis (SV2TTS) [19]. The open-source SV2TTS tool [20], based on a recent research study [22], allows one to clone a voice only from few seconds of original speech and without needing to retrain the model. This is much more data efficient, orders of magnitude faster and less computationally expensive than training a separate model for each speaker. [19]

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 (low-dimensional and meaningful representation of the voice of a speaker). The synthesizer, conditioned by the embedding, gets a text as an input, and outputs a log-mel spectrogram (a deterministic, non-invertible lossy function). Finally, the vocoder generates the speech waveform. The quality of this can only be as good as the reference audio. [19]

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 the SV2TTS framework, we will need as input a reference audio file of the professor's original speech and the string sentence from the previous step, containing the answer to the user request derived from the SQL query. The output will be an audio containing the input sentence reproduced with the cloned voice of the reference audio.

To test the designed tool, we used real information about four different courses, 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 (Chicago, US) who have given their consent. So, the cloned answers for all fifteen questions for each of those four voices were captured for further analysis.

5 RESULTS

5.1 Project Results

Regarding the similarity of the cloned voice to the real one, a similarity measurement has been applied, based on *Resemblyzer* [23], a python package to analyze and compare voices with deep learning. 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. An optimal model is expected to output high similarity values (>0.70) when the real and cloned voices are compared and lower values elsewhere. Different similarity matrices have been calculated to draw some conclusions (an example is shown in Table 2).

There is a large discrepancy when it comes to the duration of the reference audio, since it is said to be determining on the similarity of the generated voice with respect to the true one. So, we compared the results depending on the duration of the reference audio with six different lengths: 5 seconds, 5 minutes, 15 minutes, 30 minutes, 45 minutes and 1 hour.

When a voice is cloned for the first time, an embedding is created based on the reference audio. The time it takes to create such embedding (a delay due to the SV2TTS speaker encoder) depends on the duration of the original audio, being directly proportional, the longer the reference audio, the more time it takes to create an embedding (Table 3).

The major conclusions we have drawn from the results are the following:

- 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. At a glance, it seems male voices present better results than female ones. This may be due to the reference audio quality, which impacts the quality of the result.
- Note in Table 3 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.
- 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) [19].
- 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 [19], and therefore, the speaker encoder captures some form of prosody in the embedding of such speaker.
- Punctuation is not supported by the SV2TTS model [19], 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 DISCUSSION

6.1 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 [11]. As the voice cloning algorithms are getting better, it is becoming more and more difficult to discern what is real and what is not [11], which leads to several issues related to:

Table 2: Similarity values

		Cloned Voices			
		Female 1	Female 2	Male 1	Male 2
Original Voices	Female 1	**0.77**	0.66	0.54	0.58
	Female 2	0.71	**0.89**	0.52	0.58
	Male 1	0.53	0.59	**0.97**	0.62
	Male 2	0.55	0.60	0.58	**0.83**

Table 3: Embedding creation time

Reference audio length	Embedding creation time	Reference audio length	Embedding creation time
5 sec	2 sec	30 min	2 min 25 sec
5 min	25 sec	45 min	4 min 15 sec
15 min	1 min 15 sec	1 hour	6 min 10 sec

- Trust. Things that people never uttered could be pushed on the internet in a planned manner for political gains or to create unrest in society [24]. Some deepfakes have been used for a myriad of purposes such as bullying, revenge pornography, video manipulation, and extortion which definitely harm individuals' reputations [25]. Barack Obama or Mark Zuckerberg, for instance, have been the target of deepfake videos [24].
- Scamming. It will become easier for scammers to perform phishing and spoofing attacks [11]. Indeed, audio deepfakes have already been used to clone voices and defraud people [24]. In 2019, a company in the U.K. claimed it was tricked by an audio deepfake phone call into wiring money to criminals [7].

6.1.1 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 [26]. Apart from causing financial damages, it is also often used for psychological and emotional harm.

Another ethical problem could be the privacy of personal information. When deep-learning algorithms are applied, private information and data is used. Unavoidably, every platform can suffer from a privacy violation, which could potentially lead to personal information being accessed 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 some concerns about the morality of the power of resurrection [27].

6.1.2 Regulation. Legislation and regulation about voice cloning and deepfakes are both critical pieces of this technological puzzle. Voice is a very personal aspect of who we are and often a unique identifier. But do we have rights in our own voice?

The starting point would be protection under copyright. A song, an advertisement or a movie may be copyrightable, and voice may get protected if it is a part of the tangible medium [28]. However, copyright protection is not available specifically for voice per se [28]. 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 [29].

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 [30]. 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. States diverge on whether the right survives posthumously and, if so, for how long [30].

Yet as deepfakes become more realistic and accessible, concern about the potential harm they pose has increased [31]. 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. Starting in July 2019, Virginia, Texas, and California, regulated the distribution of non-consensual deepfake pornography and election related deepfakes [32]. On December 20, 2019, the U.S. Congress signed the nation's first federal law related to deepfakes which is part of the National Defense Authorization Act [25]. Experience from previous years 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.

6.2 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 [7]. 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 [11]. However, there are 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. We need to be upfront about how we're going to protect consumers who are going to be victimized in ways by criminals." [33] Therefore, transparency and an abundance of caution will be the keys to grappling with voice cloning technologies in the years to come [33].

7 CONCLUSION

This project provides a voice query assistant including a voice cloning feature that answers requests about information extracted from a syllabus of a course. Apart from that, it also presents an analysis about the misuse, regulation, and future of this technology.

After achieving the proposed goal, 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, together with eLearning, continue to gain steam due to the COVID-19 pandemic and the market is increasingly propelling demand for high quality and excellent user experience emerging technologies for both students and teachers [8].

Apart from that, we believe that even more powerful forms of voice cloning will become available in a near future. Therefore, this type of application will continue to develop in a massive way and cloned voices may soon be indistinguishable from the original ones. 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 could lead to misuse, for instance, to trick people. This induces people to think voice cloning could be a threat to their privacy and be exploited to perpetuate scams. Consequently, it is essential to raise awareness about the technology and its usage. In the future, an accurate and updated regulation will be needed for both encouraging voice cloning and managing associated risks. 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?

REFERENCES

[1] Ning, Y., He, S., Wu, Z., Xing, C. & Zhang, L. (2019). A Review of Deep Learning Based Speech Synthesis. Applied Sciences, 9(19), p.4050. https://doi.org/10.3390/app9194050

[2] Mori, M. (1970). The Uncanny Valley. Energy, 7(4), 33—35.

[3] Gottfried, J. (2019, June 14). About Three-Quarters of Americans Favor Steps to Restrict Altered Videos and Images. Pew Research Center. https://www.pewresearch.org/fact-tank/2019/06/14/about-three-quarters-of-americans-favor-steps-to-restrict-altered-videos-and-images/

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

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

[6] 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

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

[8] Nafea, I. T. (2018). Machine Learning in Educational Technology. https://doi.org/10.5772/intechopen.72906

[9] Speech synthesis. (n.d.). https://www.cs.mcgill.ca/~rwest/wikispeedia/wpcd/wp/s/Speech_synthesis.htm

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

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

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

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

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

[15] 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/cerenceintroduces-my-car-my-voice-new-voice-clone-solution/.

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

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

[18] 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-vspostgresql-a-comparison-of-relational-database-management-systems.

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

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

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

[22] 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

[23] Resemble-AI. (2020). Resemblyzer. GitHub. https://github.com/resembleai/Resemblyzer.

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

[25] 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.

[26] The United States Department of Justice. (2020, November 16). Identity Theft. https://www.justice.gov/criminal-fraud/identity-theft/identity-theft-and-identity-fraud.

[27] 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

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

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

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

[31] 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-usecases-safeguards-ai-machine-learning/.

[32] 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-deepfakes42346/.

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