Radboud Universiteit

AI voice deepfakes: The implications of increasingly sophisticated speech synthesis software – An online experiment

Adrian Koch (s1064965)
International Business Communication
Radboud University

Bachelor Thesis
LET-CIWB351-IBC-2022-SCRSEM2-V
14.07.2023

Contents

Abstract

AI speech synthesis has seen rapid development in recent years, allowing for the creation of highly realistic and convincing synthetic voices. While this technology has many promising applications, including in the fields of virtual assistants, gaming, and entertainment, it also raises important questions about the genuineness of synthetic speech and the potential risks of its misuse.

In this research paper, the genuineness of AI speech synthesis was investigate by conducting a perceptual online experiment to evaluate how well synthetic voices mimic the qualities of human speech, such as naturalness, expressiveness, and tone. Participants will be presented with both real and AI synthesized speech recordings of Barack Obama, Donald Trump and Joe Biden whereupon they will rate how authentic they perceived the given audio clip). Furthermore, the potential risks of AI speech synthesis were discussed, including the creation of voice deepfakes for malicious purposes such as impersonation, fraud, or disinformation campaigns. We highlight the importance of developing safeguards and regulations to prevent the misuse of this technology, such as authentication mechanisms and ethical guidelines.

Overall, this research sheds light on the current state of AI speech synthesis, its potential implications, and the need for responsible development and deployment of this technology.

Introduction and theoretical framework

Artificial intelligence (AI) has become increasingly prevalent and visible in our daily lives. In recent years there has been considerable technological progress regarding neural networks. Generally speaking, the main disciplines are language models and natural language processing (NLP), image generation, as well as text-to-speech (TTS) synthesis. New tools such as ChatGPT (text synthesis), DallE (image synthesis) – both provided by OpenAI – or Prime Voice AI (voice synthesis) – provided by ElevenLabs Inc. – are quickly growing in popularity. This progress is accompanied by a variety of changes, some of which are slow and steady and others which are disruptive. Crafts such as writing, speaking and designing can be facilitated immensely through different combinations of neural network models which are the foundation of cutting-edge AI tools (Lee et al., 2022). Besides assisting humans in a variety of tasks, these types of software tools are set to disrupt creative processes of all kinds as it can be difficult to expose products of some AI tools as generated by AI, especially if the output is altered further. The fact that

AI is synthesizing entirely new products from its training data is raising doubts about originality which gives rise to ethical questions about proving the authenticity of creativity.

Deepfakes are a type of synthetic media that are created using deep learning techniques such as artificial neural networks. These videos or images are created by superimposing or replacing an existing image or video with another image or video, often resulting in a highly realistic and convincing portrayal of a person or event that did not actually occur as specified by Johnson and Diakopoulos (2021). Deepfakes can be used for a variety of purposes which appear to be of rather malicious nature such as creating fake news, spreading propaganda, manipulating public opinion, blackmailing individuals, or committing fraud. Given that context, its use for entertainment purposes seems trifling. Due to their potential negative consequences, many companies and researchers are working on developing technology to detect and prevent deepfakes (Mirsky & Lee, 2021).

Until recently, AI speech synthesis was not imposing major threats in regards of deception and fraud. Early deepfake videos which were accompanied by audio were oftentimes still voiced by impressionists or comedians (Thomas, 2020). Synthesized speech was still sounding choppy and artificial, simply put robotic, but as progress is continuous the circumstances are changing. Particularly, if paired with deep fake videos, AI synthesized voices can imitate well-known personalities authentically. This undoubtedly has the potential to be used as a matter of disruption. At first this started out with videos which were made for entertainment purposes like the alleged public service announcement by Obama (BuzzFeedVideo, 2018). However, this can also be used in more serious matters as recently demonstrated in a deepfake video message of Volodymyr Zelenskyy surrendering in the war of Ukraine (e.g. Schneider, 2022; Wakefield, 2022). After ElevenLabs released a beta version of their Prime Voice AI, synthesized speech is sounding alarmingly realistic. The imitation of voices of politicians, actors, etc. is almost flawless as there is usually enough high quality sample data available. Even emotional states can be imitated and are audible in the synthesized voice.

This brings to mind numerous cybersecurity risks that come with deepfake technology like (e.g. gather personally identifiable information (PII), scamming money, gaining access to company systems, espionage, etc.). The above mentioned cases might just be harbingers of waves of disinformation and scams which could hit society in the future. As the technology is constantly evolving and its availability increasing, potential cases of mass scale fraud do not seem far-fetched. The aforementioned tech company ElevenLabs, pioneer in the field of AI speech software, admitted that the speech software appeared in voice cloning misuse cases following a report that it had been used to create

deepfake audio versions of Emma Watson and Joe Rogan which was trending on social media (Hern & Milmo, 2023). At present, cases involving deepfake voice imitations of Donald Trump, Barack Obama and Joseph Biden seemingly playing online video games together while using jargon one would never expect to be uttered by them (MalachyWho, 2023).

Presumably, the majority of people who are up to date on news and political events on a global scale know that nowadays you cannot only trust your eyes anymore when it comes to political announcements and the like. As stated in Groh et al. (2022), “The realism heuristic predicts “people are more likely to trust audiovisual modality [relative to text] because its content has a higher resemblance to the real world.” This raises the question how the relation changes if the visual part is left out. Will there be a change in how easily people can be convinced? Will it be easier or will it be harder to evade raising suspicion? Up to this point in time only few empirical studies on the impact of deepfakes were conducted (Hancock & Bailenson, 2021), let alone focusing only on voice deepfakes. While reliable deepfake detectors are just being developed and not yet available for commercial use, the ordinary audience is still vulnerable especially if they are not aware of how advanced the latest AI-based deep learning models are in terms of their ability to synthesize and clone any person’s voice requiring just a few seconds of audio as training data. Existing research investigating measures to mitigate the threat only includes detection through machine learning. In any case, current datasets to train said detection tools are still biased and insufficient (Wang et al., 2020). Furthermore, adversarial attacks have to make spoof detection more resilient and are expected which will leave such tools decreasing in reliability if the datasets fail to adapt in time. However, while these measures would be most efficient with audio-video material, synthesized speech recordings will maintain to be more inconclusive if presented exclusively.

On one hand, younger members of society could be expected to be generally more suspicious towards the congruence of speech with the context of a speech recording (e.g. of a politician), and also quicker to verify a piece of content. On the other hand, members of older generations might contemplate less on or be less aware of the capabilities of AI tools (Brashier & Schacter, 2020). More precisely, one would assume that members of Gen Z are more suspicious to deepfakes than members of Gen X as they might be more familiar with AI and the pitfalls of the newest technological developments. As the amount of genuine sounding TTS tools is increasing it will be crucial to regulate their use better to prevent any fraudulent activities (Lorenzo-Trueba et al., 2018). Following previous research by Lorenzo-Trueba et al. (2018) and Halpern et al. (2020), we

will refer to AI synthesized voice imitations as “spoof” and real authentic speech as “genuine”. The findings about spoof detection will facilitate to develop a valid methodology in order to investigate the risks connected with state-of-the art AI powered TTS synthesis. One will be soon not be able to rely on their senses anymore as, even if you are aware that the possibility to exploit someones voice exists, it just has to be good enough and catch you in an inattentive moment (Howard, 2020).

Evidently, since technology which enables users to create voice clones – as demonstrated by MattVidPro AI (2023) – has become increasingly sophisticated in recent years, with the potential to create highly realistic and convincing voice deepfakes, the range of possibilities for abusing this technology is growing. It was already found that in the majority of the cases “seeing is believing” (Groh et al., 2022), but is hearing also believing? Would the same effect be witnessed under restricted circumstances of only having audio content available but no video content to perceive the stimulus? Are people still able to distinguish between AI synthesized (spoof) speech and actual (genuine) speech and if so how? Consequently, the importance of spreading awareness and testing for the potential harm it could cause is growing likewise.

Due to the deepfakes emerging from contributions by many different actors it is unclear who needs to be held accountable in case a situation takes severe turns (Johnson & Diakopoulos, 2021). First and foremost, technical experts are the ones who have to act upon their ethical responsibility to control the consequences of their inventions. It is in the interest of tech companies developing such AI tools to be proactive and ensure their products are not used for malicious purposes. ElevenLabs, which provides “the most realistic and versatile AI speech software, ever” is in favor of minimizing potential risks as stated in the “Ethical AI statement” on the product website (Eleven Labs, 10.03.23). Thus, the research question is derived aiming to investigate how easy it remains to distinguish between genuine speech and AI generated spoof speech.

RQ1: Are humans still able to distinguishing between AI synthesized voice clones and real recordings of the person whose voice was cloned?

Based on the theory it can be expected that a considerable amount of subjects fails to distinguish between genuine and spoof speech clips with certainty. I expect the majority of people to make the right guess but it will remain a guess in most cases as the resemblance of the AI synthesized output is too refined for a human to be able to unambiguously deduct whether a speech recording is real or fake only using their sensory perception.

Methodology

The goal of this study is to first evaluate the degree of how deceptive state-of-the-art speech synthesis software is and then, based on the outcome, conclude implications for its future development. To investigate the hypothesis, a perceptual online experiment was conducted meaning that participants ability to distinguish between real speech and fake speech will be put to test while they are only allowed to process the surface features of a stimulus. In that way, the evaluation method is set to resemble a real life encounter of potential voice deepfakes on the news or the internet with the goal of collecting quantitative data in order to investigate the severity of the deceptive factor of spoof speech.

Participants

Participants of this online-experiment were randomly sampled by distributing a link to the questionnaire through social media platforms and personal contacts of the researcher. In common, the questionnaire reached a sample size of 50 respondents between 19 and 71 years old to take part in the experiment. 24 of which were male, 25 of which were female, and one subject preferred not to state their gender. The average age of the sample group is almost 30 years (M=29,66; SD=12,23) which makes it evident that people of young age are represented more than people of middle age and old age.

Findings by Brashier and Schacter (2020) show that age is an important factor when it comes to dealing with technology and the media. Thus, we expected the older a person the less pronounced their ability to debunk deepfakes. Based on Brunjes (2023) the generations were devided as follows: Gen Z: age 18 to 26; Millennials: age 27 to 42; Gen X: age 43 to 58; Boomers: age 59 to 77. Accordingly, concerning the generational distribution the sample group is strongly skewed towards Gen Z (N=32) making up more than half of participants while Millennials (N=12), Gen X (N=3) and Babyboomers (N=3) are underrepresented. Due to the sample sizes of Gen X and Babyboomers being too small, no inferences can be made concerning these generations.

In terms of native language, German speakers (N=27) were strongly overrepresented followed by Dutch speakers (N=6), English speakers (N=4) and French speakers (N=2). The following languages were only represented by one subject: Spanish, Finnish, Polish, Turkish, Greek, Romanian, Bulgarian, Cantonese, Standard Chinese, Vietnamese and Tamil.

All respondents were required to have a baseline level of English proficiency starting at A1 and were asked to state their estimated proficiency level: A1 (N=1); B1 (N=3); B2 (N=9); C1 (N=22); C2 (N=15).

Lastly, participants were asked to indicate their highest education. While a large proportion of the sample group obtained a Bachelor's degree (N=22), the remaining subjects answered that they finished high school (N=15), obtained a Master's degree (N=9) or completed an apprenticeship (N=4). Hence, the vast majority of participants attended academic education.

Materials

Resembling previous research carried out with audiovisual content (Groh et al., 2022), the audio clips were manually compiled from speeches of Barack Obama, Donald Trump or Joseph Biden as there is a sufficient amount of recordings which serves as useful training data to be fed into the speech synthesis software. The stimuli consists of 12 English audio clips in total, 6 of which were spoof (AI synthesized voice clones), and 6 of which were genuine (extracts of real speech recordings). To break it down, we get 2 genuine recordings and 2 spoof recordings per speaker.

Each audio clip was of a duration of approximately 30 seconds. The goal was to ensure a high quality of audio output. For this purpose, only flawless extracts of presidential speech recordings (amount of noise kept at a minimum) served as samples to synthesize the spoof audio clips (see Appendix).

The genuine recordings were sourced from the website of the Miller Center which "[…] is a nonpartisan affiliate of the University of Virginia that specializes in presidential scholarship, public policy, and political history […]" (U.S. Presidents | Miller Center, 15.05.23), while the spoof recordings were synthesized by ElevenLabs' PrimeVoice AI. PrimeVoice AI offers two additional settings which alter the the speech output, namely "Stability" which was set to 70%, and "Clarity + Similarity Enhancement" which was set to 50% (see Appendix).

ChatGPT 3.0 was employed in the process of writing the transcripts of the spoof audio clips (see Appendix) to create realistic speech content and recreate patterns of word use that is appropriate in presidential speeches. This also concerns distinctive utterances (e.g. Obama: "but let me be clear"; Trump: "tremendous") that can be associated with a given speaker so as to male a voice clone sound more authentic and and familiar in terms of style. Overall, it was attempted to recreate the voices of the given

speakers as accurately as possible with low effort using resources which are accessible to the public.

Design

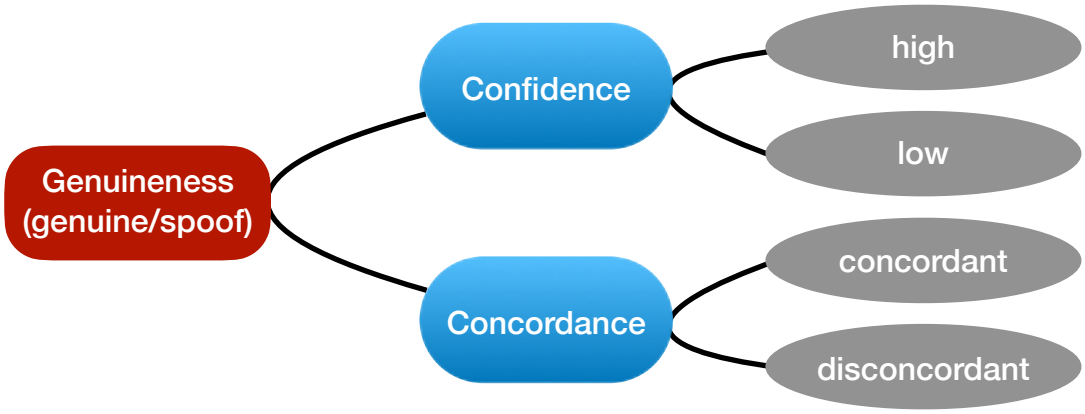
The research was conducted using a unifactorial within-subjects experimental design evaluating Genuineness to ensure that every participant would be exposed to the same audio clips of each speaker as well as to both levels of the independent variable, namely genuine and spoof audio clips.

Instruments

Through comparing the quality of spoofed audio content to genuine audio content, the aim was to get an impression of the misleading nature of the high-quality output of AI TTS software. In comparison to work done by Van Heuven and Van Bezooijen (1995), the following variables were chosen to be evaluated as they were found to account for the quality of the audio output most efficiently considering that there was no video material accompanying the audio clips and that only recordings with minimum noise were used:

The independent nominal variable 'Genuineness' with the levels genuine and spoof is operationalized through the dependent ordinal variables 'Confidence a given audio clip is fabricated' (not confident at all - very confident) and 'Concordance of an audio clip with the general perception of a speaker's speech style' (very disconcordant - very concordant), both rated on a 5-point Likert scale. The latter one, gave more insight into how well ChatGPT 3.0 can imitate a politician in terms of word use and phrasing. Additionally, respondents were urged to comment on their rating with the intention of finding out about the strongest properties of each audio clip which predominantly influenced their perception.

Independent variable **Dependent variables**



Procedure

Before starting the experiment after filling in demographic data, participants were asked to complete a short survey aimed at gathering data on their opinion about AI concerning their experience, knowledge and expectations. This was aimed to find out about the participant's general attitude towards and familiarity with AI.

After being given the instructions for the experiment, participants first took a Cognitive Reflection Test (CRT) which is considered a "robust test for measuring an individual's tendency towards reflecting on questions before answering" (Groh et al., 2022; Rand et al., 2016). While the questions given in the classic CRT might already be known to some people, alternative questions were chosen to avert that participants do not need to reflect on the question (Thomson & Oppenheimer, 2016). It gave an impression of how much a participant relies on their intuition or tends to rethink when making a decision, for instance when choosing an answer option. Additionally, this gave a better impression of a participant's thought process and the kind of reasoning that they might have applied when analyzing an audio clip.

Then participants were presented with the stimuli and determined the degree of genuineness through two measures: On one hand they rated how confident they were that a given audio clip was fabricated. On the other hand, they rated how concordant an audio clip sounded regarding their general perception of a speaker's political identity (Groh et al., 2022). Consequently, we found out how well participants are able to discern truth from falsehood – genuine versus spoof – across 12 political speeches by three well-known U.S. American politicians.

Statistical testing

In order to test the collected data statistically, first of all Cronbach's α was computed for all variables which are coded with several items. This is especially important for analysing the dependent variables Confidence and Concordance as this measurement is not entirely based on well-found literature and must therefore be validated. Subsequently, a paired samples t-test was carried out aiming to compare the average rating of the dependent variables for each audio clip. However, due to the measurement and the coding not being ideal or even partly failing to measure what it aims to measure, descriptive statistics will be key to make the most of the results of this experiment.

Results

First we will cover the section of participant's stance on AI. It was divided in familiarity with the topic of AI, use of AI tools, perception of development and integration of AI, and perceived impact of AI.

In terms of Familiarity, the reliability of knowledge about AI comprising 3 items was acceptable (α = .82). Consequently, the mean of all 3 items was used to calculate the compound variable 'Familiarity with the topic of AI'. Participants generally indicated moderate familiarity with the topic of AI (*M*=3.0, *SD*=0.97).

In terms of Use of AI tools, 70% (N=35) of respondents stated they have used AI tools before while 30% (N=15) answered they have not used any before. More specifically, 10% (N=5) of participants indicated they use AI tools every day, 16% (N=8) indicated a use of a few times a week, 18% (N=9) answered they use AI tools once a week, 10% (N=5) answered they use AI tools once a month, 22% (N=11) stated they rarely use AI tools and 24% (N=12) stated they have not used AI tools within the last 12 months. This user behavior is concluded as frequent use (44%, N=22) for the first three answer options and seldom use (56%, N=28) for the last three answer options.

Moving on to Development and Integration of AI, the reliability of feelings about the development and integration of AI comprising 5 items was acceptable (α = .79). Consequently, the mean of all 5 items was used to calculate the compound variable 'Perception of development and integration of AI'. Participants on average showed a slight tendency of negative feelings about the further development and integration (*M*=2.37, *SD*=0.64).

Finally, regarding the concept Perceived impact of AI, the statement 'The further development of AI has little impact on my life' was reverse coded so the polarity was switched for the analyses. Since this concept comprises 2 items, they remain separate in the analyses and the means should not be compiled to calculate a compound variable. Given the statement 'The further development of AI has little impact on my life' participants were also rather neutral even though a slight tendency to disagree with this statement was found (*M*=3.32, *SD*=0.87). Given the statement 'The advance of AI generally imposes more opportunities than threats' participants were indecisive not wanting to exclude either scenario (*M*=3.26, *SD*=0.90).

With respect to the CRT it can be said that it was rather unnecessary. Initially it was intended as a means for checking on participants' attention and also to find out what depth of reflection was most likely involved when assessing the audio clips but since the audio quality of the AI synthesis is quite high there is no other option than guessing or

deciding based on a feeling is some cases. More detail on that will be provided in the limitations section of the discussion.

Looking at the analysis of the dependent variables, the reliability of the variable 'Confidence that an audio clip was fabricated' comprising 12 items was acceptable: α = .71. Consequently, the mean of all 12 items was used to calculate the compound variable Confidence. However, the reliability of 'Concordance of an audio clip with the general perception of a speaker's speech style' comprising 12 items was not acceptable: α = .68. Therefore, the mean of these items cannot be used to calculate the compound variable Concordance. Due to the an inappropriate sample size and missing homogeneity of variance paired samples t-tests would not be valid in this context but it was conducted nevertheless to identify possible trends.

A paired samples t-test for Confidence showed a significant difference between genuine and spoof clips (t (49) = 6.13, p < .001). Confidence for spoof audio clips having been fabricated (M = 2.95, SD = 0.79) was shown to be higher than for genuine audio audio clips having been fabricated (M = 2.23, SD = 0.76)'. Another paired samples t-test for Concordance showed a significant difference between genuine and spoof clips (t (49) = 2.58, p = .013). Concordance for genuine audio clips (M = 3.66, SD = 0.64) was perceived to be higher than for spoof audio clips (M = 3.35, SD = 0.71). The means can be seen in comparison in Table 1.

	genuine (N=6)	spoof (N=6)
Confidence	M=2.23, SD=0.76	M=2.95, SD=0.79
Concordance	M=3.66, SD=0.64	M=3.35, SD=0.71

Table 1: Means and standard deviations of the dependent variables for each condition. The order of the audio clips can be found in block 6 of the questionnaire in the appendix to see which clips were genuine and which were spoof (AI) (1, 4, 6, 8, 9, 11-> spoof clips; 2, 3, 5, 7,10, 12-> genuine clips).

It can be noted that as for Confidence, participants' answers tended towards slight confidence that a genuine audio clip was fabricated while for spoof audio clips participants were somewhat more confident an audio clip was fabricated. Concerning Concordance, respondents' answers tended towards genuine audio clips sounding more concordant than spoof audio clips meanwhile for spoof audio clips respondents perceived the clip to be neither concordant nor disconcordant.

Concerning the open questions respondents were asked to shortly mention a specific reason for their rating? For spoof audio clips (1, 4, 6, 8, 9, 11) the top 75% comments talked about: uncertainty about rating (gut feeling), unnatural echo, audio clip sounding natural/human (breathing), audio clips sounding authentic/original, monotone/ robotic accentuation, weird flow of speech/intonation and pauses sounding unnatural. Additionally, in isolated cases the content of speech was not perceived as congruent with the speaker (e.g., Clip 8: Biden 3).

For genuine (2, 3, 5, 7,10, 12) audio clips the top 75% of comments were about: uncertainty about rating (gut feeling), audio clip sounding natural/human like (breathing), no suspicion of a fake (not fabricated/authentic), natural intonation/pauses, familiarity with speech of the speaker (certainty it's real), authentic sounding pronunciation, hearing emotions in speech, background noises making it sound real, strange background noises, natural reverb. In one case a participant remarked that the slur in Biden's his real speech is unmistakable.

Conclusion

In essence, the findings of the online experiment overall indicate that subjects had an inclination to rate genuine audio clips as genuine and to rate spoof audio clips as spoof. Nonetheless, the figures differ only marginally and are partly grouped around neutral values which suggests that respondents were unable to make a clear decision whether an audio clip was genuine or spoof in a lot of cases. Despite these marginal differences it is visible that genuine and spoof audio clips were not perceived to sound very different. Moreover, all averages show a relatively low standard deviation which demonstrates that respondents predominantly made the right assumption. A trend can be identified but inferences can only be made carefully in terms of the problems descending from the research design.

In response to the research question it can be claimed that humans can still debunk synthesized voices to a certain degree, but only under uncertainty. It can be concluded that the most regularly mentioned reasons that an audio clip is perceived as spoof are an unnatural echo, monotone/robotic accentuation, weird flow of speech/ intonation and pauses sounding unnatural. On the other hand, the most regularly mentioned reasons that a audio clip is perceived as genuine are.

The involvement of several confounding factors which are discussed closer in the limitations call for more research of this kind in order to create a more valid and diverse base to draw conclusions from.

Discussion

Referring to Groh et al. (2022) it can be claimed that hearing is not believing as seeing is believing. Nevertheless, the experiment is not an example of a day-to-day encounter with deepfake content which means that it is less likely that spoof content raises suspicion in someone who gets targeted by it. Some participants even expressed their astonishment in feedback they gave requesting the solution to the experiment about only being able to hear subtle differences between some of the audio clips and it giving them an eery feeling about how AI TTS synthesis could disrupt the future.

Even though, based on Halpern et al. (2020) it could still be argued that it is not easy to get sufficient data to build convincing fake-voice biometrics with the purpose of creating voice clones of ordinary people, privacy becomes a yet so important topic to regulate how assailable we are. The authors were here only determining likelihood ratios of a detector tool whether a recording is genuine or spoof but this is not comparable to the arbitrary human behavior.

Lorenzo-Trueba et al. (2018) laid the base for the research design at hand but here different approaches to generate spoofed speech where used. They used shorter utterances and also synthesized from transcripts of speech that has already been uttered while in our study we made up utterances for the audio clips containing spoof speech which shifts the meaning of the evidence that is concluded from the sole convincingness of the speech output to also its potential in its use to create novel utterances.

As of its current level of development, PrimeVoice AI (Eleven Labs homepage, 03.06.23) is in a number of cases quite capable of cloning and synthesizing a voice that can deceive an individual into thinking that it is the actual speaker whose voice was cloned. It is obvious that, if ideally configured, voice clones synthesized by PrimeVoice AI can sound indistinguishable. Next it is crucial to find out in what kind of scenarios someone is most vulnerable to falling for its persuasiveness without suspecting anything.

Having carried out this research we saw how little effort it takes to put together an alarmingly realistic imitation of someones voice. The experiment demonstrated that people majorly make the right decision when being confronted with audio clips that possibly have been fabricated. Irregardless, the data gives an impression of the uncertainty a participant must have felt being confronted with the audio clips at hand. What is important to note after all is that participants were told that this experiment included voice clones which might have heightened their senses to pay close attention to the detection of inconsistencies. It would be worthwhile to see how a similar investigation would turn out where people are not made aware of the attempt to trick them into

believing that fake is real because, as one or the other may already have guessed, the devil lays in the unexpectedness.

Undoubtedly, some things went wrong creating and conducting this experiment. The method of this research needs to be revised as a means to increase validity and score generalizable results. However, as Anderson (2015) pointed out, “no result is worthless […]” which is why one should focus on the conclusions that can be drawn regarding the research design as it is the first of its kind. The priority of this study was to contribute to the prevention of fraudulent activities in an attempt to strive for ethical use of technology and AI with the goal of benefiting humans in the long run and avert any negative developments and potential for conflict.

Limitations

Moving on to arguably the most important section talking about this study, the limitations of this experiment pose the essential deductions that should be carried away from reading this article. As far as the theoretical framework did not overlook any identical predecessors to this study and as far as there are no identical investigations which are ongoing, this research is the first of its kind. Due to the amateurish implementation and execution of the research method, the study lacks in validity, precision and accuracy which decreases the overall generalizability of the findings.

First of all, only one type of speech synthesis software was employed to generate the stimulus material. Even though it is one of the most advanced commercial TTS programs on the internet, other softwares should also be tested regarding their performance. Besides that, it is very likely that there are non-commercial speech synthesis softwares which are fantastically superior to PrimeVoice AI and could make it much harder for humans as well as machines to distinguish between genuine and spoof. There might as well already be a superior commercial software by the time this article will be published.

Secondly and unfortunately, only 50 participants were gathered in total during the data collection process which cannot be regarded as fully representative for the population that could be affected most by the malicious use of such an AI tool. Adding to that, a small proportion of native English speakers, especially Americans, decreases the precision of implications that can be drawn from this research as it would only be logical to have US Americans evaluate speech of US presidents. Nonetheless, it is relevant to see how susceptible subjects are to spoof speech who only have English as a second language.

Thirdly, analyzing the instrumentation and design of this investigation, there are several adjustments that can be made to rule out inaccuracy and decrease the potential for confusion distorting the real world connection of the results. Hence, there is room for improvement respecting the wording of the measurement of the variables as a means to answering the hypothesis. For instance, instead of asking “How confident are you that the given audio clip is fabricated” one could ask “Is the given audio clip fabricated” which could be answered by “is definitely fabricated, “might be fabricated”, “hard to distinguish”, “might be real” and “is definitely real”, or simply “yes” or “no”. The current wording of this variable assumes that someone has a minimum level of confidence that an audio clip is fabricated which creates a bias and might influence a respondent to be more suspicious of it having been fabricated than they would usually be. Accordingly, the measuring procedure was not completely free of bias and it would have been a smarter decision to keep the method simple in that regard. Adding a pre-test would have also been advisable to eliminate uncertainty and make the measurement of the variables less vague.

Furthermore, the order of audio clips should be randomized to balance out the procedure better and prevent that every respondent will be exposed to the same audio clip in the same order which could end up serving as a reference to which the following audio clips might be compared to, either consciously or subconsciously. This frame of bias through which the remaining stimulus material would be seen increases the noise (external influences that distort the results) in the data collection. Here, a between-subjects design would also be well applied.

Getting back to the CRT, the main purpose was to keep the participant’s attention engaged. In Groh et al. (2022) it was more meant to make respondents pay close attention to their assessment but in that specific case paying close attention still made a difference in distinguishing what is real and what is fake while in the experiment we conducted paying close attention will not help definitively to clearly tell apart real from fake. Given that answering the questionnaire already took some time to complete, in exchange for the CRT a simple question, for instance “Answer with strongly disagree when you are still paying attention”, could have been sufficient. In that way a quick attention check could have been done without adding to participants’ fatigue.

Ultimately, researchers aiming to make similar investigation in the field of evaluating speech synthesis software are encouraged to revise the method of this study with focus on the instrumentation and develop an improved version to examine the outcome under

varying conditions and with more expertise. With the aim of gathering more research data on AI TTS programs, the language of the stimulus material can be adjusted, as well the speakers whose recordings and voices are employed to create the stimulus are interchangeable, also the focus can be put on different populations, etc. There are numerous factors that can be modified in a way in which the general concept of evaluating the authenticity of TTS output is used to investigate the performance of AI or the ability of humans to tell real from fake, genuine from spoof. It should be pointed out that it would be another valuable step concerning research in this field to have participants evaluate TTS output without giving away that some part or even all of the stimulus is fake for the purpose of investigating what it takes to make people suspicious and under which circumstances people would not even contemplate anymore that the content they are being exposed to could be fake. No matter how aware one is of the advance of such technology or how prepared and protected one thinks they are, once you get caught of guard it can cause considerable chaos.

References:

Anderson, G. (2015). No result is worthless: the value of negative results in science. *On Medicine*. https://blogs.biomedcentral.com/on-medicine/2012/10/10/no-result-is-worthless-the-value-of-negative-results-in-science/

Brashier, N. M., & Schacter, D. L. (2020). Aging in an era of fake news. *Current Directions in Psychological Science*, *29*(3), 316–323. https://doi.org/10.1177/0963721420915872

Brunjes, K. (2023, January 19). *Age Range by Generation - Beresford Research*. Beresford Research. https://www.beresfordresearch.com/age-range-by-generation/

BuzzFeedVideo. (2018, April 17). *You Won't Believe What Obama Says In This Video!*. YouTube. https://www.youtube.com/watch?v=cQ54GDm1eL0

ElevenLabs || Prime Voice AI. (10.03.23 - 30.05.23). https://beta.elevenlabs.io/

Groh, M., Sankaranarayanan, A., & Picard, R. W. (2022). Human Detection of Political Deepfakes across Transcripts, Audio, and Video. *ArXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2202.12883

Halpern, B. M., Kelly, F. (2022) Can DeepFake voices steal high-profile identities?. *Work-in-Progress, Student paper.* http://oxfordwaveresearch.com/wp-content/uploads/2022/07/Abstract-Can-DeepFake-voices-steal-high-profile-identities.pdf

Hancock, J. T., & Bailenson, J. N. (2021). The social impact of Deepfakes. *Cyberpsychology, Behavior, and Social Networking*, *24*(3), 149–152. https://doi.org/10.1089/cyber.2021.29208.jth

Hern, A., & Milmo, D. (2023, February 24). Everything you wanted to know about AI – but were afraid to ask. *The Guardian*. https://www.theguardian.com/technology/2023/feb/24/ai-artificial-intelligence-chatbots-to-deepfakes

Howard, R. (2020). *Deepfake scams will be hard to detect. Here are some tips*. (07.06.23). https://blog.ting.com/internet/deepfake-scams

Johnson, D. G., & Diakopoulos, N. (2021b). What to do about deepfakes. *Communications of the ACM*, *64*(3), 33–35. https://doi.org/10.1145/3447255

Lee, K., Hitt, G., Terada, E., & Lee, J. H. (2022). Ethics of Singing Voice Synthesis: Perceptions of Users and Developers. *Zenodo (CERN European Organization for Nuclear Research)*. https://doi.org/10.5281/zenodo.7316768

Lorenzo-Trueba, J., Fang, F., Wang, X., Echizen, I., Yamagishi, J., & Kinnunen, T. (2018). Can we steal your vocal identity from the Internet?: Initial investigation of cloning Obama's voice using GAN, WaveNet and low-quality found data. *The Speaker and Language Recognition Workshop (Odyssey 2018)*. https://doi.org/10.21437/odyssey.2018-34

MalachyWho. (2023, February 25). *Presidents Play Rocket League Ranked* [Video]. YouTube. https://www.youtube.com/watch?v=Ovi1pz89U6o

MattVidPro AI. (2023, February 3). *This AI Speaks Just like a Human & CLONES Voices Flawlessly | ElevenLabs* [Video]. YouTube. https://www.youtube.com/watch?v=kqzI91YIfmw

Mirsky, Y., & Lee, W. (2021). The Creation and Detection of Deepfakes. *ACM Computing Surveys*, *54*(1), 1–41. https://doi.org/10.1145/3425780

Pennycook, G., Cheyne, J. A., Koehler, D. J. & Fugelsang, J. A. (2016). Is the cognitive reflection test a measure of both reflection and intuition? *Behav. Res. Methods* 48, 341–348. https://doi.org/10.3758/s13428-015-0576-1

Schneider, N. M. J. (2022, March 18). Wie Deepfakes im Ukraine-Krieg genutzt werden. *Wie Deepfakes Im Ukraine-Krieg Genutzt Werden - ZDFheute*. https://www.zdf.de/nachrichten/politik/selenskyj-deepfake-video-ukraine-krieg-russland-100.html (Date of access: 03.03.2023)

Thomas, D.: Deepfakes: A threat to democracy or just a bit of fun? (2020), https://www.bbc.com/news/business-51204954 (Date of access: 07.03.2023)

Thomson, K. S., & Oppenheimer, D. M. (2016). Investigating an alternate form of the cognitive reflection test. *Judgment and Decision Making*, *11*(1), 99–113. https://doi.org/10.1017/s1930297500007622

U.S. Presidents | Miller Center. (n.d.). Miller Center. https://millercenter.org/president

Vaccari, C., & Chadwick, A. (2020). Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News. *Social Media + Society*, *6*(1). https://doi.org/10.1177/2056305120903408

Van Heuven, V. J., & Van Bezooijen, R. (1995). Quality evaluation of synthesized speech. *Elsevier Science EBooks*, 707–738. https://openaccess.leidenuniv.nl/bitstream/handle/1887/1065/5_167_054.pdf?sequence=1

Wakefield, J.: Deepfake presidents used in Russia-Ukraine war (2022), https://www.bbc.com/news/technology-60780142 (Date of access: 03.03.23)

Wang, X., Yamagishi, J., Todisco, M., Delgado, H., Nautsch, A., Evans, N., ... & Ling, Z. H. (2020). ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech. *Computer Speech & Language*, *64*, 101114. https://doi.org/10.1016/j.csl.2020.101114

Appendix

Material for the synthesis of the spoofed audio clips

Presidential speeches from which extracts were used as genuine recordings and as training data for the voice clones:

Obama:
https://millercenter.org/the-presidency/presidential-speeches/september-10-2013-address-nation-syria

https://millercenter.org/the-presidency/presidential-speeches/december-4-2013-speech-economic-mobility

https://millercenter.org/the-presidency/presidential-speeches/november-20-2014-address-nation-immigration

Trump:
https://millercenter.org/the-presidency/presidential-speeches/february-15-2019-speech-declaring-national-emergency

https://millercenter.org/the-presidency/presidential-speeches/march-11-2020-statement-coronavirus

https://millercenter.org/the-presidency/presidential-speeches/january-13-2021-statement-about-violence-capitol

Biden:
https://millercenter.org/the-presidency/presidential-speeches/july-8-2021-speech-drawdown-us-forces-afghanistan

https://millercenter.org/the-presidency/presidential-speeches/september-9-2021-remarks-fighting-covid-19-pandemic

https://millercenter.org/the-presidency/presidential-speeches/february-24-2022-remarks-russian-invasion-ukraine

PrimeVoice AI settings applied to generate the spoof audio clips:

Stability: 70%
-> “Increasing variability can make speech more expressive with output varying between re-generations. It can also lead to instabilities.”

Clarity + Similarity Enhancement: 50%
-> “Low values are recommended if background artifacts are present in generated speech.”

(ElevenLabs || Prime Voice AI, 17.05.23)

Transcripts of spoof audio clips:

1_AI_Obama:
For too long, we have ignored the warning signs, turned a blind eye to the facts, and pushed the consequences of our actions onto future generations. We have exploited our planet's resources, polluted our air and water, and caused irreparable damage to our ecosystems.
The transition to a green economy will not happen overnight. It will require patience, perseverance, and cooperation from all of us. But I have faith in our ability to rise to the occasion, to innovate, and to create a better world for ourselves and for future generations.
We have the power to be the change that we seek. Let us embrace this challenge with courage and conviction, knowing that we are fighting for a cause that is greater than ourselves.

2_AI_Obama:
My fellow Americans,
Today I come before you to talk about one of the defining issues of our time: Inequality. Across our nation there are millions of hardworking families who are struggling to make ends meet. There are children who go to bed hungry, parents who can't afford healthcare, and seniors who are unable to retire with dignity.
But let me be clear - this is not just an issue of compassion or morality. This is an economic issue, a national security issue, and a moral imperative. When we have an economy that works for everyone, not just those at the top, we unleash the full potential of our nation.

2_AI_Trump:
So, let me express my heartfelt condolences to the families who have lost loved ones during this trying time. Our thoughts and prayers are with you, and we will do everything in our power to support you through this difficult period.
This virus came upon us unexpectedly, but we will not let it define us. We will rise above it and emerge stronger than ever before. And I want to assure you that my administration is taking every necessary step to protect the American people and safeguard our economy.

3_AI_Trump:
I have been in contact with local authorities and law enforcement, and I want to commend their swift and heroic response to this tragedy. Their bravery saved countless lives, prevented further harm and helped a tremendous deal to get a hold of the shooting that occurred last night.
To the people of Las Vegas, know that we are with you. We will stand by your side as you begin the process of healing and we will ensure that justice is served for those responsible for this heinous act.

2_AI_Biden:
We cannot afford to ignore this reality any longer. We must take bold and decisive action to address climate change and protect our planet for future generations.
That is why I am proud to announce that my administration is taking significant steps to combat climate change. We have rejoined the Paris Climate Agreement, which is a critical global effort to reduce greenhouse gas emissions and limit the worst effects of climate change.
We are also investing in clean energy technologies and infrastructure, such as wind and

solar power, to reduce our reliance on fossil fuels and create good-paying jobs for Americans.

3_AI_Biden:
Over the past year, our nation has faced unprecedented challenges. The COVID-19 pandemic has wreaked havoc on our economy, threatened the health and well-being of our citizens, and caused immense suffering for millions of Americans.
In response to this crisis, my administration worked tirelessly to craft a comprehensive relief package that would provide immediate assistance to those in need and pave the way for our recovery.

Questionnaire

Block 1: CONSENT FORM
-> I agree
-> I do not want to participate

Block 2: Demographic data
- Age: __
- Gender: male, female, non-binary, prefer not to say
- Native language: English, German, Dutch, French, Spanish, Other, namely:
- Highest education: primary school, high school, apprenticeship, bachelor's degree, master's degree, doctoral degree, Other, namely:
- Estimated English proficiency: A1 (beginner/elementary), A2 (pre-intermediate), B1 (intermediate), B2 (upper intermediate), C1 (advanced), C2 (proficient/native speaker)

Block 3: Survey aimed at gathering data on participant's opinion about AI
- I am up to date on the topic of artificial intelligence: strongly agree – strongly disagree
- I know a lot about AI: strongly agree – strongly disagree
- I know how AI works: strongly agree – strongly disagree
- I have used AI tools before (e.g. ChatGPT, Dall-E 2, Lumen5, etc.): yes, no
- I use AI tools… : every day, a few times a week, once a week, once a month, rarely, not within the last 12 months
- The development and integration of AI into our daily lives is… :
  Good – Bad
  Pleasant – Unpleasant
  Beneficial – Harmful
  Interesting – Boring
  Wise – Unwise
- The advance of AI generally imposes more opportunities than threats:
  strongly agree – strongly disagree
- The further development of AI has little impact on my life:
  strongly agree – strongly disagree

Block 4: Cognitive Reflection Test (CRT)
Before we will start the experiment you will complete a short test engaging your attention and tendency to rethink a "gut" response and engage in further reflection to find a correct answer.

1. If you're running a race and you pass the person in second place, what place are you in? (intuitive answer: first; correct answer: second)
2. A farmer had 15 sheep and all but 8 died. How many are left? (intuitive answer: 7; correct answer: 8)
3. Emily's father has three daughters. The first two are named April and May. What is the third daughter's name? (intuitive answer: June; correct answer: Emily)

Block 5: Introduction
Now you will be presented with 12 audio clips of Obama, Trump and Biden:
- Pay attention to how natural the speech sounds (how confident you are that a given audio clip is fabricated), and
- how congruent with the speech style of the given speaker (how concordant or discordant) you you think the clip sounds.

Try to only listen to each audio clip once and move on to evaluating the next clip.

Block 6: Experiment

Order of audio clips:
1. 1_AI_Obama
2. 3_Trump
3. 2_Trump
4. 3_AI_Biden
5. 3_Obama
6. 2_AI_Trump
7. 3_Biden
8. 2_AI_Biden
9. 2_AI_Obama
10. 1_Obama
11. 3_AI_Trump
12. 2_Biden

- Rate how confident you are that the given audio clip is fabricated (5-point Likert scale).
- Rate how concordant the given audio clip sounds (5- point Likert scale).
- Shortly mention a specific reason for your rating? (e.g. characteristics of the audio, gut feeling, etc.).