# Q&A: What you need to know about audio deepfakes

March 19 2024, by Rachel Gordon

Credit: Unsplash/CC0 Public Domain

Audio deepfakes have had a recent bout of bad press after an AI-generated robocall purporting to be the voice of Joe Biden hit up New Hampshire residents, urging them not to cast ballots. Meanwhile, spear-phishers—phishing campaigns that target a specific person or group, especially using information known to be of interest to the target—go fishing for money, and actors aim to preserve their audio likeness.

What receives less press, however, are some of the uses of audio deepfakes that could actually benefit society. In this Q&A prepared for MIT News, postdoc Nauman Dawalatabad addresses concerns as well as potential upsides of the emerging tech. A fuller version of this interview can be seen at the video at the end of the article.

## What ethical considerations justify the concealment of the source speaker's identity in audio deepfakes, especially when this technology is used for creating innovative content?

Dawalatabad: The inquiry into why research is important in obscuring the identity of the source speaker, despite a large primary use of generative models for audio creation in entertainment, for example, does raise ethical considerations.

Speech does not contain the information only about "who you are?" (identity) or "what you are speaking?" (content); it encapsulates a myriad of sensitive information including age, gender, accent, current health, and even cues about the upcoming future health conditions.

For instance, our recent research paper, "Detecting Dementia from Long Neuropsychological Interviews," demonstrates the feasibility of

detecting dementia from speech with considerably high accuracy.

Moreover, there are multiple models that can detect gender, accent, age, and other information from speech with very high accuracy. There is a need for advancements in technology that safeguard against the inadvertent disclosure of such private data.

The endeavor to anonymize the source speaker's identity is not merely a technical challenge but a moral obligation to preserve individual privacy in the digital age.

**How can we effectively maneuver through the challenges posed by audio deepfakes in spear-phishing attacks, taking into account the associated risks, the development of countermeasures, and the advancement of detection techniques?**

The deployment of audio deepfakes in spear-phishing attacks introduces multiple risks, including the propagation of misinformation and fake news, identity theft, privacy infringements, and the malicious alteration of content.

The recent circulation of deceptive robocalls in Massachusetts exemplifies the detrimental impact of such technology. We also recently spoke with the spoke with the *Boston Globe* about this technology, and how easy and inexpensive it is to generate such deepfake audios.

Anyone without a significant technical background can easily generate such audio, with multiple available tools online. Such fake news from deepfake generators can disturb financial markets and even electoral outcomes. The theft of one's voice to access voice-operated bank accounts and the unauthorized utilization of one's vocal identity for financial gain are reminders of the urgent need for robust countermeasures.

Further risks may include privacy violation, where an attacker can utilize the victim's audio without their permission or consent. Further, attackers can also alter the content of the original audio, which can have a serious impact.

Two primary and prominent directions have emerged in designing systems to detect fake audio: artifact detection and liveness detection. When audio is generated by a generative model, the model introduces some artifact in the generated signal. Researchers design algorithms/models to detect these artifacts. However, there are some challenges with this approach due to increasing sophistication of audio deepfake generators.

In the future, we may also see models with very small or almost no artifacts. Liveness detection, on the other hand, leverages the inherent qualities of natural speech, such as breathing patterns, intonations, or rhythms, which are challenging for AI models to replicate accurately. Some companies like Pindrop are developing such solutions for detecting audio fakes.

Additionally, strategies like audio watermarking serve as proactive defenses, embedding encrypted identifiers within the original audio to trace its origin and deter tampering. Despite other potential vulnerabilities, such as the risk of replay attacks, ongoing research and development in this arena offer promising solutions to mitigate the threats posed by audio deepfakes.

## Q: Despite their potential for misuse, what are some positive aspects and benefits of audio deepfake technology? How do you imagine the future relationship between AI and our experiences of audio perception will evolve?

Contrary to the predominant focus on the nefarious applications of audio deepfakes, the technology harbors immense potential for positive impact across various sectors. Beyond the realm of creativity, where voice conversion technologies enable unprecedented flexibility in entertainment and media, audio deepfakes hold transformative promise in health care and education sectors.

My current ongoing work in the anonymization of patient and doctor voices in cognitive health-care interviews, for instance, facilitates the sharing of crucial medical data for research globally while ensuring privacy. Sharing this data among researchers fosters development in the areas of cognitive health care.

The application of this technology in voice restoration represents a hope for individuals with speech impairments, for example, for ALS or dysarthric speech, enhancing communication abilities and quality of life.

I am very positive about the future impact of audio generative AI models. The future interplay between AI and audio perception is poised for groundbreaking advancements, particularly through the lens of psychoacoustics—the study of how humans perceive sounds. Innovations in augmented and virtual reality, exemplified by devices like the Apple Vision Pro and others, are pushing the boundaries of audio experiences towards unparalleled realism.

Recently we have seen an exponential increase in the number of sophisticated models coming up almost every month. This rapid pace of research and development in this field promises not only to refine these technologies but also to expand their applications in ways that profoundly benefit society. Despite the inherent risks, the potential for audio generative AI models to revolutionize health care, entertainment, education, and beyond is a testament to the positive trajectory of this research field.

Provided by Massachusetts Institute of Technology