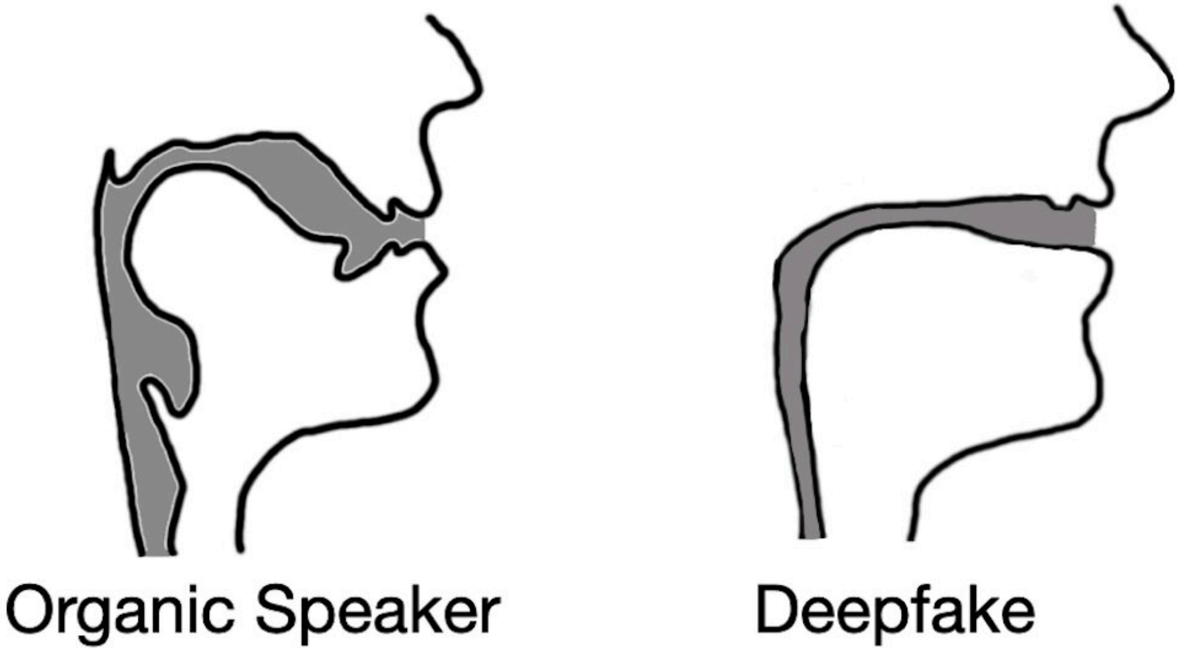
Deepfake audio has a tell: Researchers use fluid dynamics to spot artificial imposter voices

September 21 2022, by Logan Blue and Patrick Traynor



Deepfaked audio often results in vocal tract reconstructions that resemble drinking straws rather than biological vocal tracts. Credit: Logan Blue et al., CC BY-ND

Imagine the following scenario. A phone rings. An office worker answers it and hears his boss, in a panic, tell him that she forgot to transfer money to the new contractor before she left for the day and needs him to do it. She gives him the wire transfer information, and with the money transferred, the crisis has been averted.

The worker sits back in his chair, takes a deep breath, and watches as his boss walks in the door. The voice on the other end of the call was not his boss. In fact, it wasn't even a human. The voice he heard was that of an audio deepfake, a machine-generated audio sample designed to sound exactly like his boss.

Attacks like this using recorded audio have already occurred, and conversational audio deepfakes might not be far off.

Deepfakes, both audio and video, have been possible only with the development of sophisticated machine learning technologies in recent years. Deepfakes have brought with them a new level of uncertainty around digital media. To detect deepfakes, many researchers have turned to analyzing visual artifacts—minute glitches and inconsistencies—found in video deepfakes.

Audio deepfakes potentially pose an even greater threat, because people often communicate verbally without video—for example, via phone calls, radio and voice recordings. These voice-only communications greatly expand the possibilities for attackers to use deepfakes.

To detect audio deepfakes, we and our research colleagues at the University of Florida have developed a technique that measures the acoustic and fluid dynamic differences between voice samples created organically by human speakers and those generated synthetically by computers.

Organic vs. synthetic voices

Humans vocalize by forcing air over the various structures of the vocal tract, including vocal folds, tongue and lips. By rearranging these structures, you alter the acoustical properties of your vocal tract, allowing you to create over 200 distinct sounds, or phonemes. However, human anatomy fundamentally limits the acoustic behavior of these different phonemes, resulting in a relatively small range of correct sounds for each.

In contrast, audio deepfakes are created by first allowing a computer to listen to audio recordings of a targeted victim speaker. Depending on the exact techniques used, the computer might need to listen to as little as 10 to 20 seconds of audio. This audio is used to extract key information about the unique aspects of the victim's voice.

The attacker selects a phrase for the deepfake to speak and then, using a modified text-to-speech algorithm, generates an audio sample that sounds like the victim saying the selected phrase. This process of creating a single deepfaked audio sample can be accomplished in a matter of seconds, potentially allowing attackers enough flexibility to use the deepfake voice in a conversation.

Detecting audio deepfakes

The first step in differentiating speech produced by humans from speech generated by deepfakes is understanding how to acoustically model the vocal tract. Luckily scientists have techniques to estimate what someone—or some being such as a dinosaur— would sound like based on anatomical measurements of its vocal tract.

We did the reverse. By inverting many of these same techniques, we were able to extract an approximation of a speaker's vocal tract during a segment of speech. This allowed us to effectively peer into the anatomy of the speaker who created the audio sample.

From here, we hypothesized that deepfake audio samples would fail to be constrained by the same anatomical limitations humans have. In other words, the analysis of deepfaked audio samples simulated vocal tract shapes that do not exist in people.

Our testing results not only confirmed our hypothesis but revealed something interesting. When extracting vocal tract estimations from deepfake audio, we found that the estimations were often comically incorrect. For instance, it was common for deepfake audio to result in vocal tracts with the same relative diameter and consistency as a drinking straw, in contrast to human vocal tracts, which are much wider and more variable in shape.

This realization demonstrates that deepfake audio, even when convincing to human listeners, is far from indistinguishable from human-generated speech. By estimating the anatomy responsible for creating the observed speech, it's possible to identify the whether the audio was generated by a person or a computer.

Why this matters

Today's world is defined by the digital exchange of media and information. Everything from news to entertainment to conversations with loved ones typically happens via digital exchanges. Even in their infancy, deepfake video and audio undermine the confidence people have in these exchanges, effectively limiting their usefulness.

If the digital world is to remain a critical resource for information in people's lives, effective and secure techniques for determining the source of an audio sample are crucial.

Provided by The Conversation

Citation: Deepfake audio has a tell: Researchers use fluid dynamics to spot artificial imposter voices (2022, September 21) retrieved 3 May 2024 from https://techxplore.com/news/2022-09-deepfake-audio-fluid-dynamics-artificial.html