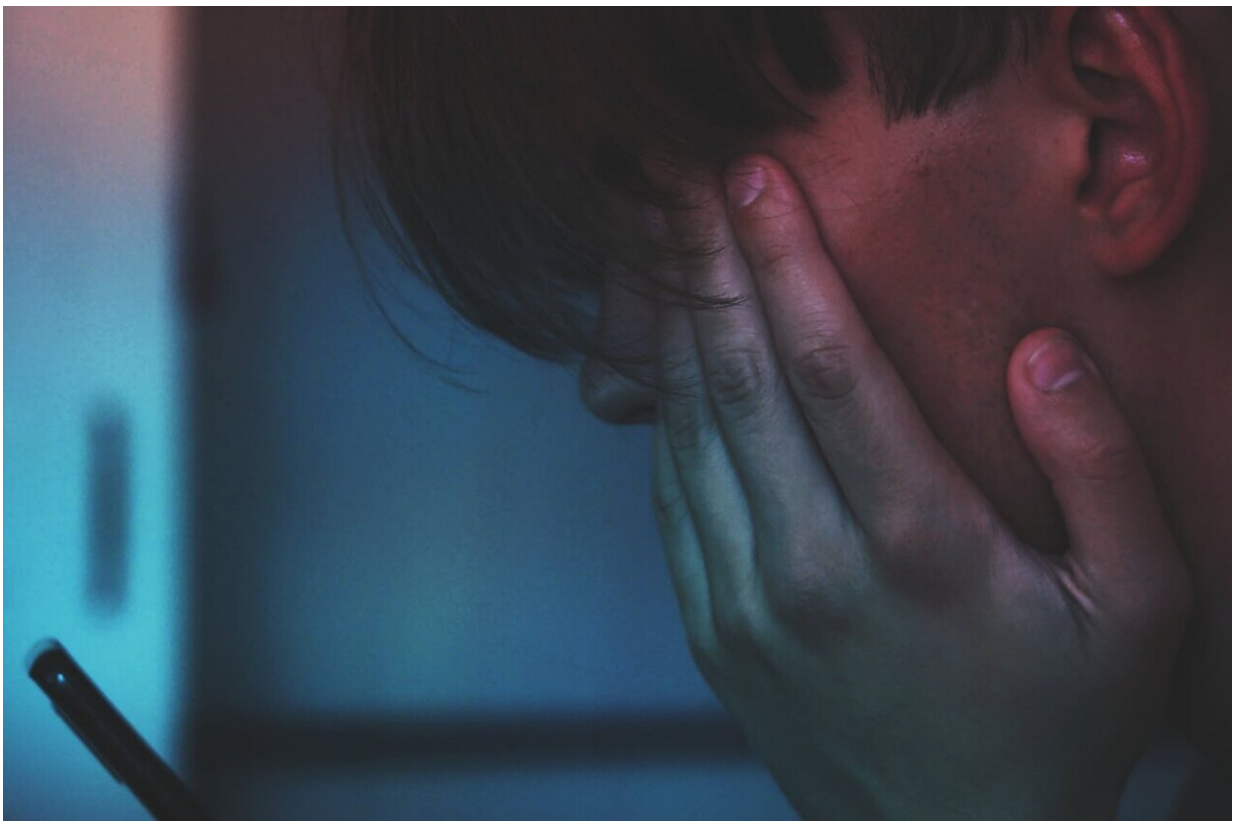
Study shows speech deepfakes frequently fool people, even after training on how to detect them

August 2 2023



The researchers suggest that training people to detect speech deepfakes is unrealistic, and efforts should focus on improving automated detectors. Credit: Adrian Swancar, Unsplash, CC0 (creativecommons.org/publicdomain/zero/1.0/)

In a study involving more than 500 people, participants correctly identified speech deepfakes only 73% of the time, and efforts to train participants to detect deepfakes had minimal effects. Kimberly Mai and colleagues at University College London, UK, presented these findings in the open-access journal *PLOS ONE* on August 2, 2023.

Speech deepfakes are synthetic voices produced by machine-learning models. Deepfakes may resemble a specific real person's voice, or they may be unique. Tools for making speech deepfakes have recently improved, raising concerns about security threats. For instance, they have already been used to trick bankers into authorizing fraudulent money transfers.

Research on detecting speech deepfakes has primarily focused on automated, machine-learning detection systems, but few studies have addressed humans' detection abilities.

Therefore, Mai and colleagues asked 529 people to complete an online activity that involved identifying speech deepfakes among multiple audio clips of both real human voices and deepfakes. The study was run in both English and Mandarin, and some participants were provided with examples of speech deepfakes to help train their detection skills.

Participants correctly identified deepfakes 73% of the time. Training participants to recognize deepfakes helped only slightly. Because participants were aware that some of the clips would be deepfakes—and because the researchers did not use the most advanced speech synthesis technology—people in real-world scenarios would likely perform worse than the study participants.

English and Mandarin speakers showed similar detection rates, though when asked to describe the speech features they used for detection, English speakers more often referenced breathing, while Mandarin

speakers more often referenced cadence, pacing between words, and fluency.

The researchers also found that participants' individual-level detection capabilities were worse than that of top-performing automated detectors. However, when averaged at the crowd-level, participants performed about as well as automated detectors and better handled unknown conditions for which automated detectors may not have been directly trained.

Speech deepfakes are likely to only become more difficult to detect. Given their findings, the researchers conclude that training people to detect speech deepfakes is unrealistic, and efforts should focus on improving automated detectors. However, they suggest that crowdsourcing evaluations on potential deepfake speech is a reasonable mitigation for now.

More information: Mai KT, Warning: Humans cannot reliably detect speech deepfakes, *PLoS ONE* (2023). DOI: 10.1371/journal.pone.0285333 , journals.plos.org/plosone/arti … journal.pone.0285333

Provided by Public Library of Science