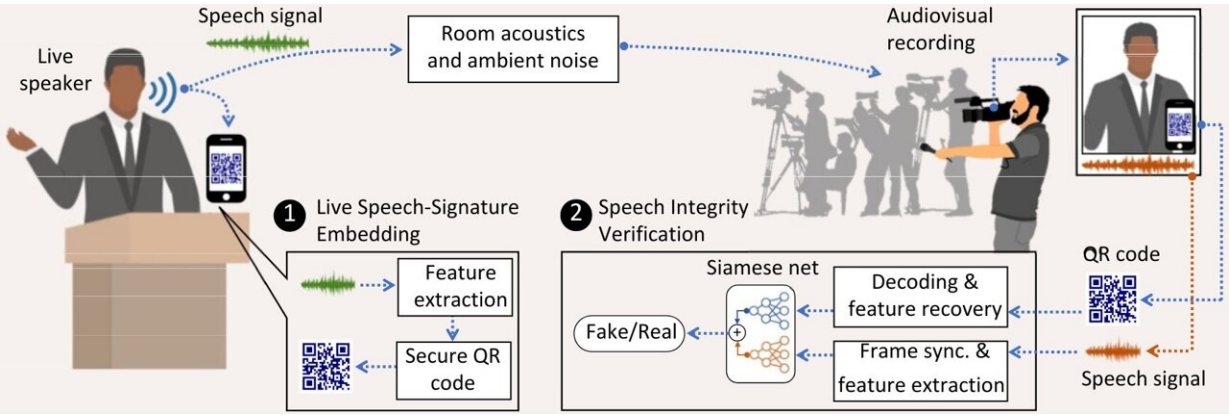
Fighting deepfakes, shallowfakes and media manipulation

January 30 2024



A diagram overview of how TalkLock works. Users can embed a secure QR code in a live recording of an event in real-time and later verify other multimedia content against this code. Credit: Nirupam Roy

Photo, audio and video technologies have advanced significantly in recent years, making it easier to create convincing fake multimedia content, like politicians singing popular songs or saying silly things to get a laugh or a click. With a few easily accessible applications and some practice, the average person can alter the face and voice of just about anyone.

But University of Maryland Assistant Professor of Computer Science Nirupam Roy says media manipulation isn't just fun and games—a bit of

video and audio editing can quickly lead to life-changing consequences in today's world. Using increasingly sophisticated technologies like artificial intelligence and machine learning, bad actors can exploit the lines between fiction and reality more easily than ever.

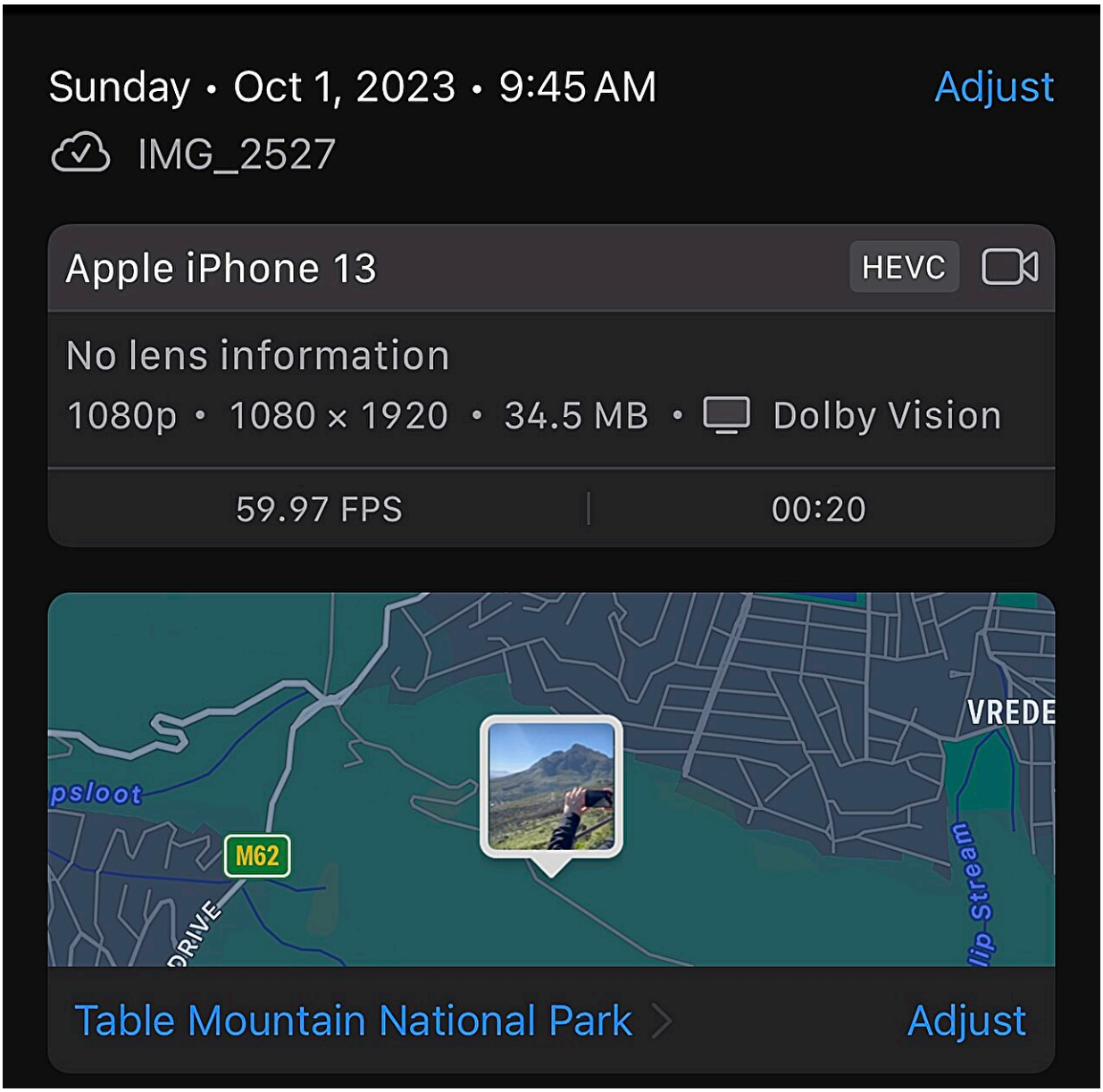
To combat this growing threat, Roy is developing TalkLock, a cryptographic QR code-based system that can verify whether content has been edited from its original form.

"In 2022, a doctored video of Ukrainian President Volodymyr Zelenskyy was circulated on the internet by hackers. In that fake video, he appeared to tell his soldiers to lay down their arms and give up fighting for Ukraine," said Roy, who holds a joint appointment in the University of Maryland Institute for Advanced Computer Studies.

"The clip was debunked, but there was already an impact on morale, on democracy, on people. You can imagine the consequences if it had stayed up for longer, or if viewers couldn't verify its authenticity."

Roy explained that the fake Zelenskyy video is just one of many maliciously edited video and audio clips circulating on the web, thanks to recent surges in multimedia content called deepfakes and shallowfakes.

"While deepfakes use artificial intelligence to seamlessly alter faces, mimic voices or even fabricate actions in videos, shallowfakes rely less on complex editing techniques and more on connecting partial truths to small lies," Roy said. "Shallowfakes are equally—if not more—dangerous because they end up snowballing and becoming easier for people to accept smaller fabrications as the truth. They make us raise questions about how accurate our usual sources of information can be."



Metadata provides basic information about a piece of multimedia. Credit: Georgia Jiang

Finding a way to tamper-proof recordings of live events

After observing the fallout from the viral falsified Zelenskyy video and

others like it, Roy realized that fighting deepfakes and shallowfakes was essential to preventing the rapid spread of dangerous disinformation.

"We already have a few ways of counteracting deepfakes and other audio-video alterations," Roy said. "Beyond just looking for obvious discrepancies in the videos, websites like Facebook can automatically verify the metadata of uploaded content to see if it's altered or not."

Metadata contains information about a piece of media, such as when it was recorded and on what device. Image editors like Photoshop also leave editing history in a photo's metadata. The metadata embedded in a file can be used to cross-check the origins of the media, but this commonly used authentication technique isn't foolproof.

Some types of metadata can be added manually after a video or audio clip is recorded while other types can be stripped entirely. These shortfalls make using default metadata alone as an authenticator unreliable, especially for recordings of live events.

"A big problem we're running into is what happens at a live event, like a public speech or press conference," he added. "Any audience member can technically record a video of a speech and upload it somewhere at their convenience. And once it's up, it's free to be downloaded and re-uploaded again and again, recirculating through multiple people who might have malicious intentions. We just don't have source control over media that's recorded at a live event by a member of the audience."

To solve this problem, Roy and his team created TalkLock, a system that can generate a QR code capable of protecting the authenticity of a public figure's likeness.

"The main idea is to use a device like a smartphone or tablet to continuously generate cryptographic sequences created from little bits of

the live speech, forming a unique QR code. That QR code captures carefully extracted features of the speech," Roy explained.

"Because the QR code will be displayed on the device's screen with the speaker, any authentic recordings of the speaker will also contain the QR code. The presence of the QR code marks the verifiability of the live recording, even if it's posted in different formats, uploaded on different social media platforms or shown on TV."

In addition to its ability to place a unique marker on a video or audio clip, TalkLock can also systematically analyze features from a recording and check them against the code sequence generated from the original live version. Any discrepancies found by TalkLock would indicate that the content was altered.

"As long as the generated QR code is recorded along with the speaker, political leaders, public figures and celebrities would be able to protect their likenesses from being exploited," Roy said. "It's the first step to preserving the integrity of our information, protecting people from crimes like targeted defamation."

Not just for celebrities and politicians

It's only the beginning for TalkLock and its capabilities, according to Roy.

"Although it may seem like actors and politicians are the only ones who should be worried, they're no longer the only targets of malicious media manipulation," Roy said. "Ordinary people are at risk now, too. Their likenesses can also be used to create false narratives, fraud and extortion attempts, blackmail and more."

Roy noted that publicly posted photos and videos on social media

platforms like Instagram and Facebook make it easier than ever for abuse and privacy violations to occur.

To address this need for protection at the individual level, his team is developing a mobile app version of TalkLock, which will be more tailored to the average person's needs and can be used by anyone who owns a smartphone. He expects the app to be completed in summer 2024.

"People can just hold their phone nearby with the app on as they speak and just doing that will create a layer of protection from malicious editing," he explained. "Users will be able to control their own audio-video footprint online with just their phones."

Roy hopes that similar protections will be available to the public as default settings on all mobile devices soon. Computer science Ph.D. students Irtaza Shahid and Nakul Garg and undergraduates Robert Estan and Aditya Chattopadhyay are working with Roy to develop an open-source implementation of the TalkLock software stack and the mobile app. The team recently published a paper explaining the key concept of the project in the proceedings of MobiSys '23, the International Conference on Mobile Systems, Applications and Services.

"Our ultimate goal is to make sure that everyone can have equal access to real, genuine information," Roy added. "Only then can we make a step closer to a truly equitable and democratic society."

More information: Irtaza Shahid et al, "Is this my president speaking?" Tamper-proofing Speech in Live Recordings, *Proceedings of the 21st Annual International Conference on Mobile Systems, Applications and Services* (2023). DOI: 10.1145/3581791.3596862