OriginStory

Authenticating the human origin of voice at the time of recording

As AI becomes more prevalent in our daily lives, the specific challenge of distinguishing human-generated voices from those cloned by AI technologies becomes increasingly critical. To address these issues, OriginStory is developing a new technique that authenticates the human origin of voice recordings at the point of creation and then embeds this authentication as a watermark or signature in the stream, establishing a chain of trust from the moment the voice is captured to when it reaches the listener. This provides a proactive approach to protecting against the misuse of AI-enabled voice cloning.

AI-based detection isn't a long-term solution

Existing solutions to the voice cloning problem rely on AI-based detection, where pre-trained models identify AI-generated content. However, this approach faces challenges, including the need for continual model updates to keep pace with advancing deepfake technologies. The open-sourcing and independent training of deepfake models further complicates the problem, making it increasingly difficult to train models that distinguish AI-generated content from real signals, especially in high-stakes scenarios requiring high accuracy.

Authenticating the human origin of voice: a core solution to voice cloning

OriginStory posits a fundamentally new approach to media creation, one that requires authentication that there is a live human producing speech before it can be recorded. Typically microphones only record speech acoustics. We use off-the-shelf sensors already integrated in many devices to simultaneously measure speech acoustics and the co-occurring biosignals in the throat and mouth as a person is speaking (vibration of vocal folds and movement of articulators). By validating that these two signals have the same origin (e.g. the human speaker), we can validate that the speaker is human as no other system produces speech in the same way. Furthermore, the approach is privacy-preserving as the biosignals are not personally identifiable; and it is low in computational overhead and can be used for real-time communication (mobile phones, wireline, teleconferencing, VoIP) as well as recorded media. This solution can be thought of as *CAPTCHA for voice recording*: the voice cannot be verified as human unless the parallel, co-occurring biosignals are captured with the acoustics. Once a voice is recorded, the audio is watermarked or cryptographically-signed as "authentically human." At the receiver, this signature is decoded via an API and an "icon of human authenticity" is shown to the consumer, providing fool-proof evidence that the speech belongs to a human. Our validation studies demonstrate that human-generated content can be authenticated with an error rate less than 0.004%.

OriginStory's solution is feasible, resilient, and does not burden consumers

This solution leverages sensors already integrated in many consumer devices (and can be readily integrated into those that don't currently have them), making wide-scale adoption feasible. It places the primary responsibility for mitigating voice cloning harms on device manufacturers and service providers. By integrating the technology at the source of media creation (i.e., in devices), it intervenes at the origin, preventing the proliferation of synthetic media before it reaches consumers. Consumers are not required to take additional steps or possess technical expertise. The authentication process is seamlessly integrated into the device, operating automatically and passively in the background without altering the existing user experience. Similarly, at the receiver, the user is shown an easy-to-interpret "icon of human authenticity", fostering trust between the user and the media.

The solution is resilient to evolving voice cloning technology as it does not rely on the characteristics of the cloned voice at all, but on the presence of human biosignals during speech production. These signals are near impossible to spoof as they require modeling the intricate physiological phenomenon that results in speech and the material properties of human flesh, both of which impact the resulting biosignals.