F²Key: Dynamically Converting Your Face into a Private Key Based on COTS Headphones for Reliable Voice Interaction

Di Duan[1], Zehua Sun[1], Tao Ni[1], Shuaicheng Li[1], Xiaohua Jia[1], Weitao Xu[1], Tianxing Li[2]
[1]City University of Hong Kong, [2]Michigan State University

ABSTRACT

In this paper, we proposed F²Key, the first earable physical security system based on commercial off-the-shelf headphones. F²Key enables impactful applications, such as enhancing voiceprint-based authentication systems, reliable voice assistants, audio deepfake defense, and the legal validity of artifacts. The key idea of F²Key is to establish a stable acoustic sensing field across the user's face and embed the user's facial structures and articulatory habits into a user-specific generative model that serves as a private key. The private key can decrypt the Channel Impulse Response (CIR) profiles provided by the acoustic sensing field into an inferred spectrogram that can match the real one calculated from the corresponding speech, provided that the user's CIR-spectrogram mapping relationship is consistent with the one embedded in the generative model. Extensive experiments demonstrate that F²Key resists 99.9%, 96.4%, and 95.3% of speech replay attacks, mimicry attacks, and hybrid attacks, respectively. We discussed and evaluated F²Key from different perspectives, such as the health consideration and identical twins study, to show the practicality and reliability.

CCS CONCEPTS

• **Human-centered computing** → **Ubiquitous and mobile computing systems and tools**.

KEYWORDS

Acoustic sensing, Deepfake detection, Earable Sensing, Physical security system

ACM Reference Format:
Di Duan[1], Zehua Sun[1], Tao Ni[1], Shuaicheng Li[1], Xiaohua Jia[1], Weitao Xu[1], Tianxing Li[2]. 2024. F²Key: Dynamically Converting Your Face into a Private Key Based on COTS Headphones for Reliable Voice Interaction. In *The 22nd Annual International Conference on Mobile Systems, Applications and Services (MOBISYS '24), June 3–7, 2024, Minato-ku, Tokyo, Japan.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3643832.3661860

1 INTRODUCTION

With the rapid development of AI, deepfake scams [2, 24, 61] have emerged and have caused billions of dollars in losses [22]. Most deepfake scams are based on deep fakes of video and audio. For example, by forging deepfake voices to impersonate victims, scammers call others for fraud [24, 61]; using photos and speech from the Internet to create deeply synthesized celebrity videos for the purpose of commercial deception [2]. The counterfeit artifacts (*i.e.*, video, audio) reach a level of realism that is almost indistinguishable from the victims, either in terms of visual images or auditory sounds. It presents unprecedented challenges to online social networks and trust among individuals.

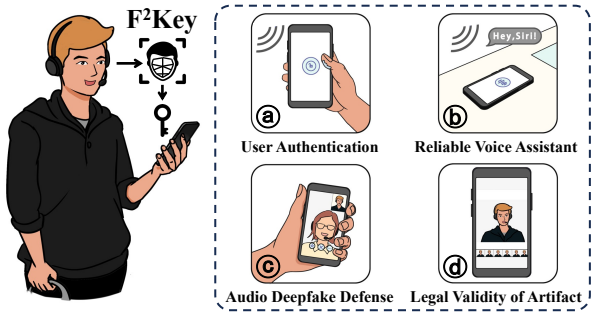


Figure 1: F²Key, the first earable physical security system based on COTS headphones and multi-modality that enables several security-related applications; it embeds the real-world information tied to the speaker in audio.

To detect these ubiquitous counterfeit artifacts (*i.e.*, audio, video) or artifacts that have been maliciously tampered with, using the voiceprint embedded in the audio as a basis for verification is an effective method to detect audio-visual forgeries [88]. However, since the obtrusive voiceprint in human speech is pervasive and easily accessible, it can be recorded and exploited for synthesis, mimicry, and replay attacks. It is challenging to defend against these attacks since the attacker can acquire the victim's real voiceprint and bypass any authentication systems solely based on voiceprint, provided the recorded audio has sufficient quality. Fortunately, the audio recorded from the victim is not arbitrary; generating specific speech requires the use of synthesis attacks and mimicry attacks. Researchers have investigated how to defend against these attacks based on spectrum [58, 78, 79], utterance level average [15], relationships with breathing [18], and injecting subtle adversarial perturbations [32, 74]. However, they all rely on the nuance between fake and real speech and cannot defend against replay attacks that use the victim's speech. Existing solutions also exploit liveness detection to verify whether the voiceprints in the audio are produced by a live person or a machine, according to cues such as breaths [64] and multiple channels [44, 82, 87]. However, multi-channel approaches require additional sensors, such as a microphone array [44], and the liveness detection cannot defend against hybrid attack. For example, an attacker provides liveness

MOBISYS '24, June 3–7, 2024, Minato-ku, Tokyo, Japan

ACM ISBN 979-8-4007-0581-6/24/06…$15.00
https://doi.org/10.1145/3643832.3661860

information by speaking silently while simultaneously replaying the recorded victim's speech.

Recent work proposed a speech verification system that can protect live speeches from malicious alterations by matching the QR code (registered by a speaker's speech) placed on site and the speaker's speech through meta-information [62]. However, it only focuses on malicious tampering with speech in videos, neglecting the role of faces in the video and overlooking how to prevent forgery and tampering in pure audio without QR codes. Motivated by this research gap, we propose leveraging the facial information to verify the user and designing an anti-counterfeiting solution applicable to pure audio. To extract facial biometric information, existing work leverages visual information (*e.g.*, facial images from the video) and computer vision techniques [40, 84]. However, the visually observable faceprint can easily be captured and accessed on the Internet, and facial images generated with cutting-edge tools [33, 69] have reached a degree of realism that can easily fool a human. Although some studies [30, 83] investigate the relationship between mouth movements and speech to jointly detect deepfakes, they are still based on 2D images. Recently, more tricky methods of forgery have emerged; the methods [17, 23, 71] developed to generate realistic images of the mouth area according to speech content can also be applied to manufacture deepfakes, which leads to 2D facial images becoming unreliable.

To address the above issues, we utilize non-visual sensing modalities (*e.g.*, ultrasound) to perceive facial biometric information and link it with the corresponding speech. Recent studies have demonstrated that low-frequency ultrasound can detect human facial movements by establishing an acoustic sensing field across the user's face [41, 65]. However, they focused only on sensing applications, overlooking their research value in security. To fill this research gap, we propose F^{2}Key (Fig. 1), the first earable physical security system that converts a user's face into a private key via commercial off-the-shelf (COTS) headphones. F^{2}Key offers a range of applications, including enhancing voiceprint-based authentication systems against replay and mimicry attacks, securing voice assistants to prevent unauthorized access, defending against audio deepfake fraud in digital communications, and providing legal validity to media files by embedding the real-world information tied to the speaker in them.

However, it is non-trivial to instantiate F^{2}Key in practice due to three main challenges: **(1)** The two speakers of COTS headphones are occluded by the user's head when worn. Therefore, using a speaker as an ultrasound transmitter, even if ultrasound can leak out, the signal-to-noise ratio (SNR) would be very low, unsatisfying the need for fine-grained perception. **(2)** It is challenging to model the ambiguous relationships between facial structures, articulatory habits, and the corresponding speech using an earable acoustic sensing field. **(3)** Since static features (*e.g.*, facial images) can be easily stolen and reused, it is challenging to dynamically leverage the aforementioned relationships in practice.

We propose three countermeasures to address these challenges. First, we propose a new hardware setup that uses COTS headphones equipped with a boom microphone and our proposed "auxiliary spacer" (elaborated in § 5) to increase the quality and SNR of the ultrasound received to detect fine-grained facial articulatory gestures. Second, we design a challenge-response mechanism to model the relationships that can tie the acoustic features with the corresponding speech. Finally, we design a generative model to embed this mapping relationship from the acoustic features to the spectrogram of user speech as a private key. The private key can dynamically decrypt the variational channel impulse response (CIR) profiles of the acoustic sensing field into an inferred spectrogram that can match the real one, provided that the wearer's CIR-spectrogram mapping relationship is consistent with that embedded in the generative model. The contributions of this paper can be summarized as follows:

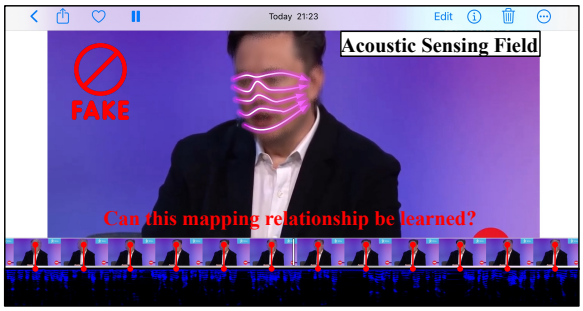


Figure 2: The image is from a deepfake video, which indirectly led to a cryptocurrency scam [2].

- We design and implement the first earable physical security system based on COTS headphones equipped with a boom microphone. By addressing the low SNR and low-quality issues of ultrasound, the hardware setup establishes a stable acoustic sensing field across the user's face and enables fine-grained articulatory gesture detection.
- We are the first to investigate linking the user's facial structures and articulatory habits with the corresponding speech in a non-visual way. We illustrate the relationship by proposing a challenge-response mechanism.
- We propose a holistic solution encompassing hardware design, verification, and deep learning models. Comprehensive evaluations demonstrate that F^{2}Key is reliable and practical and can defend against 99.9% of replay attacks, 96.4% of mimicry attacks, and 95.3% of hybrid attacks.

2 MOTIVATION & THREAT MODEL

2.1 Motivation

The severity of voiceprint-based security threats, such as replay attacks and audio deepfakes, has escalated dramatically. These attacks replicate or manipulate video or audio, posing significant challenges to traditional social networks and trust between individuals on the Internet. Motivated by the absence of a wearable physical security system that can effectively defend against these attacks, we attempt to fill this research gap by proposing a multi-modality solution.

We also provide a motivating example shown in Fig. 2. The artifacts created by deepfake technologies have reached a level of realism that is almost indistinguishable for a human, whether it be in terms of visual images or auditory sounds. To effectively prevent artifact forgery, embedding physical information from the real world into artifacts is a wise choice. We found that there is a

corresponding relationship between the speaker's speech and facial movements. If we can perceive the face in a non-visual way and combine it with the speaker's speech, it can greatly increase the difficulty of video counterfeiting. Since our target is low-frequency human speech (< 8,000 Hz [70]), using ultrasound to perceive the face is the best choice. This is because the frequency band of ultrasound is much higher than that of speech and can be received by the same microphone (perfect synchronization) along with the speech. In addition, it does not introduce additional sensors, thus avoiding an increase in implementation costs. When considering speech recording and facial perception, it naturally leads to the idea of using portable and head-mounted headphones in our system.

In this paper, we aim to pioneer the development of the first earable physical security system based on COTS headphones equipped with a boom microphone. The system can defend against attacks by jointly using both non-visual facial information and speech, and it can robustly resist various types of attacks.

2.2 Threat Model

We consider the following attacks to pose a threat to audio artifacts and voiceprint-based authentication systems:

Synthetic Attack. An attacker uses technologies such as text-to-speech [35] or voice cloning [3, 16] to synthesize audio forgeries to impersonate the victim.

Replay Attack. An attacker records the voice of the victim and plays the recording when required to pass through the authentication system. Since the playback theoretically contains the leaked ultrasounds and the voiceprint that is identical to that of the victim, replay attacks present a considerable challenge to voiceprint-based authentication systems.

Mimicry Attack. An attacker attempts to deceive the authentication system based on voiceprint by speaking and mimicking the voice characteristics (such as pitch, rhythm, accent, etc.) of the victim. This strategy typically requires a certain level of skill and familiarity with the victim.

Hybrid Attack. It is a combination of the above attack types. For example, an attacker might execute a replay attack and a mimicry attack simultaneously. They could replay the victim's speech while pretending to perform the corresponding articulatory gestures (*i.e.*, silent speech), thereby deceiving liveness detection or multi-modal authentication systems.

3 FEASIBILITY STUDY

Inspired by the discovery in § 2.1, we investigate the feasibility of using COTS headphones equipped with a boom microphone to create a stable acoustic sensing field across a user's face and discuss the defense rationale of F²Key.

3.1 Direct Propagation Path

In this section, we first abstract the representative propagation paths using the speaker and the boom microphone to detect facial movements. Then, we analyze the acoustic impedance of the ultrasound wave in different propagation mediums to indicate the strongest propagation path.

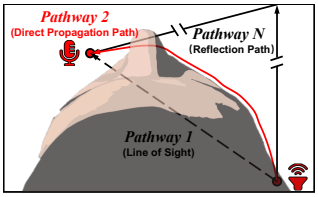


Figure 3: Signal propagation paths.

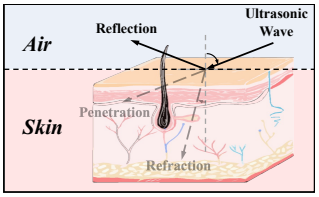


Figure 4: Ultrasonic waves propagate from air to skin.

Table 1: Acoustic impedance of different mediums [56].

Medium	Air (20 °C)	Fat	Muscle	Bone
Acoustic Impedance (kg/m^2s)	0.0004×10^6	1.34×10^6	1.71×10^6	7.8×10^6

In the process of propagating ultrasound from a speaker to the opposite boom microphone, the multipath effect can be abstracted into the three most representative paths (*i.e.*, Pathway 1, Pathway 2, and Pathway N) shown in Fig. 3. Pathway 1 represents the straight path (*i.e.*, line of sight) in which the signal travels through the user's head and reaches the microphone; Pathway 2 represents the propagation path fitting the user's face, formed by complex diffraction and reflection processes; Pathway N represents a path in the classic multipath propagation, where the signal is reflected by distant surfaces and subsequently received by the microphone.

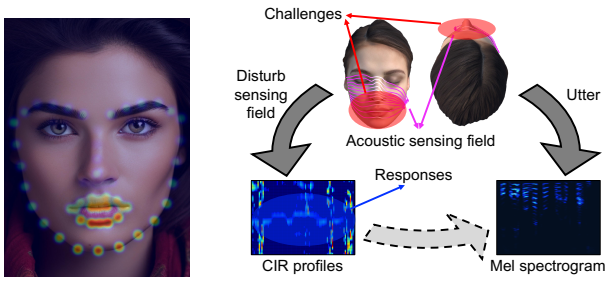
As Fig. 4 shows, when ultrasound propagates from air to skin, significant reflection and attenuation occur, especially when there is a significant difference in acoustic impedance between the mediums. Due to a four-orders-of-magnitude difference in acoustic impedance between human tissue and air (Tab. 1), most of the ultrasonic energy is reflected and absorbed upon contact with the skin surface, resulting in extremely low penetration. Therefore, the energy of Pathway 1, which alternates between different mediums, such as air, fat, bone, and muscle, can be considered negligible. If ultrasound can leak from an ear pad of headphones and be captured by the microphone (Pathway 2 and N), the ultrasound energy is significantly stronger than that of Pathway 1.

Furthermore, due to the far longer reflection distance, the energy of the ultrasonic wave along Pathway N is significantly lower than that of Pathway 2, which undergoes complex reflections and diffractions at the skin surface. Therefore, Pathway 2 is the direct propagation path with the strongest energy among the numerous components and is sensitive to perturbations of facial movements; it establishes the theoretical foundation for the proposed system.

3.2 Acoustic Challenge-Response

To dynamically convert the user's face into a private key that can be used in conjunction with the user's speech, we first conduct an in-depth analysis of the active areas and informative characteristics of the human face during speaking.

First, we detect face landmarks on a single-person TV address video from BBC News to obtain each landmark's activity level.



(a) Active areas. (b) Challenge-response relationships.

Figure 5: Acoustic challenge-response rationales. The faces in the images are not real for anonymity.

Specifically, we use OpenCV [6] and dlib [39] to detect 68 facial landmarks in each frame and analyze the cumulative changes between adjacent frames to measure the activity level for each landmark. Next, normalize these activity values to a range of 0–100 and project them onto an AI-generated face to visualize the result. In Fig. 5(a), the colors represent the activity level in the corresponding facial regions. Areas that are relatively active during user vocalization are concentrated in the lower face, with the mouth and chin dominant. If we establish an acoustic sensing field that can cover these areas, the sensing field will be perturbed by facial articulatory gestures, resulting in variations of CIR profiles along the time dimension.

Since the acoustic sensing field is sensitive to the perturbations of speech-related active areas, we can consider the user's articulatory gestures as challenges to the acoustic sensing field, with the corresponding responses being variations in the CIR profiles along the time dimension. The differences in individual facial structures, articulatory habits, and the consistency of each individual's challenge-response mechanism lay the foundation for implementing an anti-spoofing verification approach by F^{2}Key.

As Fig. 5(b) shows, when the user performs articulatory gestures to utter some speech, the gestures can be considered as some challenges (red areas) for the stable acoustic sensing field. Owing to the perturbation caused by the facial articulatory gestures, there will be some responses (variations in the blue area) in the CIR profiles. Since the speech and variations in the CIR profiles are caused by the same articulatory gestures, there are strong intrinsic relationships between them, and we can learn a generative model to translate CIR profiles into corresponding spectrograms similar to the principle of acoustic-based silent command recognition. It effectively becomes a private key by embedding the unique CIR-spectrogram mapping relationship into a generative model. This key decrypts CIR profiles into a spectrogram that should match the real one, provided that the user's CIR-spectrogram mapping relationship is consistent with the one embedded in the generative model.

Key insight: Dynamic verification of user legitimacy is possible by embedding the unique CIR-spectrogram mapping relationship of each individual into a user-specific generative model.

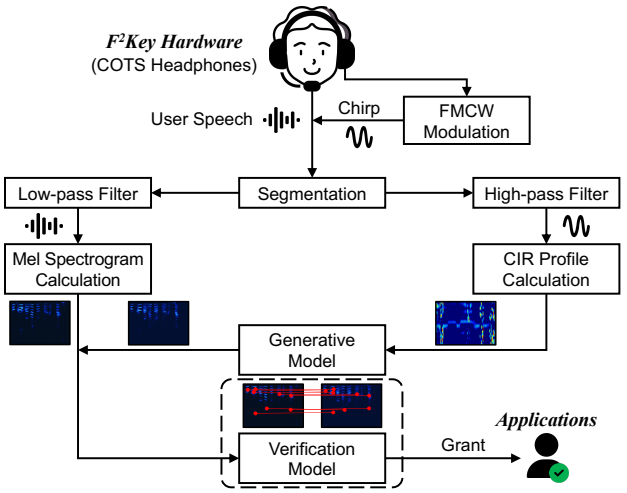


Figure 6: F^{2}Key system overview.

By combining facial structures and articulatory gestures, this challenge-response mechanism greatly enhances the reliability of voiceprint security, which is otherwise highly susceptible to theft and vulnerable to replay attacks.

4 SYSTEM OVERVIEW

Based on the above analysis and study, we propose F^{2}Key, the first earable physical security system based on COTS headphones which can provide speech-included artifact verification and deepfake detection. It effectively resists replay, mimicry, and hybrid attacks by verifying user identity via the CIR-spectrogram mapping relationship rather than using static features as template.

Fig. 6 illustrates the overview of F^{2}Key. When the legitimate user utters a speech segment, the speaker on the opposite side of the boom microphone emits frequency modulated continuous wave (FMCW). With the help of auxiliary spacers, the "escaped" chirp signal establishes a stable acoustic sensing field across the user's face, which precisely tracks the user's facial articulatory gestures. The user's speech and FMCW signals are received by the microphone at the same time. F^{2}Key first filters in user speech and high-frequency FMCW signals using a low-pass filter and a high-pass filter, respectively. Thereafter, the low-frequency and high-frequency components will be calculated into multi-scale Mel spectrograms and informative CIR profiles, respectively. The CIR profiles are used as conditions for a registered generative model to generate inferred Mel spectrograms. Finally, the user-specific verification model will calculate the similarity between the inferred Mel spectrograms and the real ones, returning a verification result according to the verification model's decision threshold.

Note that, although F^{2}Key targets speech, the generative model is registered by a user's facial structures and articulatory habits; F^{2}Key links them with the corresponding speech. Therefore, in addition to serving as a second authentication factor for the voiceprint-based two-factor authentication (2FA) system, F^{2}Key can also be used to verify the authenticity of speech-included artifacts.

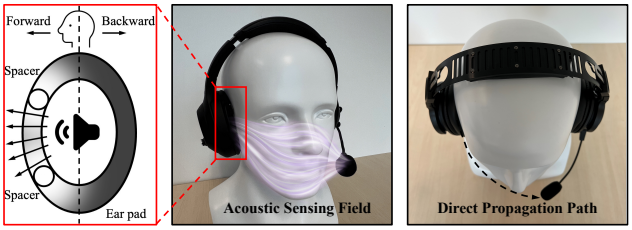


Figure 7: F²Key uses auxiliary spacers to intentionally create a gap that allows the ultrasound to "escape" and pass by the user's face, thus establishing an acoustic sensing field that is perturbed by articulatory gestures.

5 HARDWARE DESIGN

To increase the SNR of the FMCW, we propose attaching two auxiliary spacers to enhance the leaked ultrasonic waves without any modification to the COTS headphones.

5.1 Auxiliary Spacers Design

As depicted in Fig. 7, we use skin-friendly material spacers, asymmetrically attaching them on one side of the ear pad. The asymmetric design intentionally creates a gap on one side for the ultrasonic waves to "escape" towards the direction of the microphone, while maintaining close contact with the skin behind the user's ear on the opposite side. As a result, a stable acoustic sensing field is established across the user's face as shown in the middle figure of Fig. 7. We also evaluate the impact of auxiliary spacers in terms of privacy leakage and listening experience in § 9.7.

5.2 FMCW Modulation

The FMCW, known for its ideal autocorrelation property, enables the separation of signal propagation paths by estimating their CIR [41, 72]. In this paper, we use the speaker on the headphone and the opposite-side boom/modular microphone to transmit and receive the 15 kHz–20 kHz FMCW signal to capture the subtle displacement of the human face caused by articulatory gestures. There are several reasons why we choose FMCW in this frequency range:

- The upper limit of the frequency response range for most COTS headphone speakers is 20 kHz and the frequency range is within the capabilities of most COTS headphones, making it a practical choice.
- The frequency range was selected because it is higher than the perceptual upper limit (closer to 15 kHz [55]) of most adults and thus inaudible for most users, which has been verified in other studies [41].
- The frequency range provides sufficient bandwidth with a lower level of autocorrelation side lobes [68], resulting in adequate perceptual granularity and sensitivity to subtle skin deformations [41].

For some auditory-sensitive users, such as teenagers, we can appropriately reduce the bandwidth and use 16 kHz as the starting frequency of the chirp. In F²Key, we configure the FMCW signal to have a period of 1200 samples at a sampling rate of 48 kHz.

Algorithm 1 CIR profiles informative area extraction

Input: S: received audio, R: reference chirp
Output: C_i: CIR profiles' informative area

$S_h \leftarrow$ Butterworth(S) ▷ Obtain high frequency component
$Cor \leftarrow$ CrossCorrelation(S_h, R)
$C \leftarrow$ StackSlices(Cor, LENGTH(R)) ▷ Obtain full view
$m \leftarrow$ MeanCurve(C) ▷ Calculate mean curve
$m_s \leftarrow$ SavitzkyGolay(m) ▷ Smooth mean curve
$peaks \leftarrow$ FindPeaks(m_s) ▷ Find all peaks
$index \leftarrow$ Index($peaks$, 1) ▷ Get the highest peak's index
$startIndex \leftarrow index - 50$
$endIndex \leftarrow index + 50$ ▷ Set start and end indices
if $startIndex \geq 0$ and $endIndex \leq$ LENGTH(R) **then**
    $C_i \leftarrow C[startIndex : endIndex]$
else if $startIndex < 0$ **then**
    $startIndex \leftarrow startIndex +$ LENGTH(R)
    $C_i \leftarrow$ Concatenate($C[startIndex :], C[: endIndex]$)
else if $endIndex >$ LENGTH(R) **then**
    $endIndex \leftarrow endIndex -$ LENGTH(R)
    $C_i \leftarrow$ Concatenate($C[startIndex :], C[: endIndex]$)
end if ▷ Segment and extract informative area
$C_i \leftarrow C_i -$ Mean(C_i, 1) ▷ Subtract mean along time dimension
$C_i \leftarrow$ Clip($C_i, threshold$) ▷ Clip values below threshold

Then, F²Key can update the user's facial structures 40 times per second. Additionally, the minimal distinguishable difference in length between adjacent ultrasonic propagation paths is 34000/48000 = 0.708 cm. The sensing granularity is adequately fine to capture facial articulatory gestures. This is corroborated by the consistency and reproducibility of the CIR profiles obtained in § 6.1.

We also investigate the health implications of F²Key by measuring the sound pressure level (SPL) of its ultrasound. The SPL at the external auditory meatus of a human head model and the microphone to receive ultrasound are approximately 65 dB and 33 dB, respectively. These levels are within the recommended limits of the World Health Organization (WHO) of 70 dB for 24 hours to prevent hearing impairment [25], offering a safety margin of 5 dB (3.16 times), thus ensuring the safety of F²Key for users.

6 SIGNAL PROCESSING

In this section, we elaborate on calculating the CIR profiles and the multi-scale Mel spectrograms that serve as generative conditions and targets, respectively.

6.1 CIR Profiles

We apply a high-pass Butterworth filter with a cutoff frequency of 15 kHz to remove the vocal component and obtain ultrasonic signals. These signals are used to analyze variations in the acoustic sensing field. The CIR profiles calculation and informative area extraction can be summarized as Algorithm 1.

First, we conduct a cross-correlation analysis to determine the distance variations of multi-path signal propagation. Then, we obtain a complex sequence that can approximate the CIR of the echo propagation channels [11]. To isolate acoustic channels affected by articulatory gestures and remove irrelevant ones (*e.g.*, those influenced by surroundings), we segment the CIR values into slices, each with 1200 consecutive values, covering all channels in a chirp

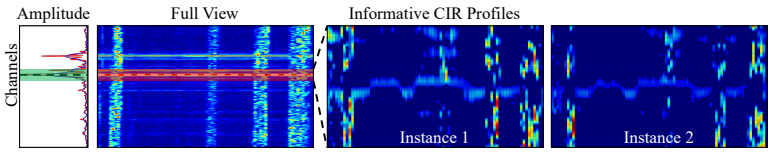


Figure 8: Imformative CIR profiles calculation.

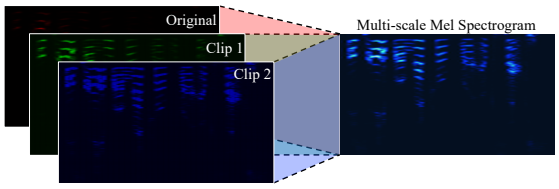


Figure 9: Multi-scale Mel spectrogram.

sequence. Then, we stack these slices along the time dimension as full-view CIR profiles. Subsequently, we calculate each channel's mean over time and smooth the curve using a Savitzky–Golay filter. We then locate the highest peak in the smoothed curve and select 50 channels before and after it as the informative area. If the peak is near edge channels, we scroll upward or downward to concatenate channels, forming this area. To emphasize changes in informative channels, we subtract the mean values over time, clipping the values below a threshold to zero. This threshold is set at 10% of the maximum value.

Fig. 8 shows informative CIR profiles of an ultrasound segment, reflecting articulatory gesture-induced perturbations in the acoustic field. It also presents two instances where the same user, wearing F^{2}Key, utters the same sentence twice. Noticeable similarities between these instances confirm that the CIR profiles accurately track and consistently reproduce the user's articulatory gestures.

6.2 Multi-scale Mel Spectrogram

After obtaining the high-frequency component from the received audio, we apply a low-pass Butterworth filter with a cut-off frequency of 8 kHz to filter in vocal components. Since signals above 8 kHz have minimal impact on speech intelligibility and human perception [51], we resample the filtered vocal component to 16 kHz to reduce computational burden, according to the Shannon sampling theorem [52, 63].

We propose a multi-scale Mel spectrogram extraction method that stacks Mel spectrograms of various numerical scales to enhance the information carried in speech signals, effectively utilizing vocalprint. Specifically, we first generate an original Mel spectrogram using select hyperparameters: a 1600-point Fast Fourier Transform (FFT) window, a hop length of 400 samples, and a setting of 256 Mel frequency bands. Next, we adopt two levels of clipping to the original Mel spectrogram. The first level clips the values between 0 and 10, while the second level clips the values between 0 and 1. The results of the two clips are shown in Fig. 9. This step is performed to mitigate the effect of extreme values in the spectrograms and to enhance the contrast between different frequency components. Finally, the three scales of the Mel spectrograms are stacked as three channels in RGB format, creating a three-channel RGB image. This approach aims to facilitate subsequent training of a generative model, treating the transformation from CIR profiles to Mel spectrograms as an image-based task.

7 MODEL DESIGN

We now present the design of the F^{2}Key models. F^{2}Key requires a pair of a generative model and a verification model for each user, which are responsible for generating the inferred spectrograms and determining the match between the generated and the real spectrograms, respectively.

7.1 Generative Model

We use pix2pix [34] based on the conditional generative adversarial network (cGAN) to generate inferred spectrograms S_{inf} using CIR profiles C as conditions. The generative model extracts the mapping relationship from CIR profiles to corresponding speech and embeds it into a trained model for further verification. The implementation of pix2pix follows the open source repository [89]. Pix2pix is a powerful generative model that has shown remarkable capabilities in transformations between paired training data. Furthermore, unlike diffusion models, which require a step-by-step denoising process that results in exceedingly long generation times [13], models trained with cGAN take less time, making them suitable for deployment on mobile devices.

The pix2pix model consists of two primary components: a generator G_θ and a discriminator D_ϕ. The generator is tasked with generating inferred spectrograms S_{inf} that are indistinguishable from real spectrograms S_r given the CIR profiles. On the other hand, the discriminator tries to distinguish between S_r and S_{inf}. The generator G_θ takes CIR profiles C as input and generates an inferred spectrogram $S_{inf} = G_\theta(C)$. The discriminator D_ϕ receives a real image pair (S_r, C) or a fake image pair (S_{inf}, C) and attempts to classify them as real or fake. The training process involves alternating between updating the discriminator and the generator. The discriminator is trained to maximize its ability to correctly classify real and generated spectrograms, while the generator is trained to fool the discriminator.

The objective function comprises two components: adversarial loss and L1 loss. The adversarial loss ensures that the generated spectrograms are close to those of the real one, while the L1 loss ensures fidelity to the input CIR profiles. The overall loss function can be expressed as:

$$\mathcal{L}(\theta, \phi) = \mathcal{L}_{adv}\left(G_\theta, D_\phi\right) + \lambda \mathcal{L}_{L1}\left(G_\theta\right),$$

where $\mathcal{L}adv$ is the adversarial loss, $\mathcal{L}_{L1}$ is the L1 loss, and λ is a hyperparameter that balances the two components.

The adversarial loss and the L1 loss can be formulated as:

$$\begin{aligned}\mathcal{L}_{\text{adv}}\left(G_\theta, D_\phi\right) =& \mathbb{E}_{S_r, C\sim p(S_r,C)}\left[\log D_\phi\left(S_r, C\right)\right] \\ &+ \mathbb{E}_{C\sim p(C), S_{inf}\sim G_\theta(C)}\left[\log\left(1 - D_\phi\left(S_{inf}, C\right)\right)\right],\end{aligned}$$

$$\mathcal{L}_{L1}\left(G_\theta\right) = \mathbb{E}_{S_r, C\sim p(S_r,C), S_{inf}\sim G_\theta(C)}\left[\left|S_r - S_{inf}\right|\right].$$

We set the batch size at 1 and use the Adam optimizer for both the generator and the discriminator with a learning rate of 0.0002 and the momentum term of Adam β_1 of 0.5. CIR profiles and spectrograms are resampled to 256×256, and the generator uses UNet [57]

as its backbone. We set $\lambda = 100$ to ensure that the L1 loss has sufficient weight in the overall loss function. The model is trained for 200 epochs.

7.2 Verification Model

Once the generative model M_G is prepared, the discriminator will be discarded, and we use the generator to generate each S_{inf} using the corresponding C as conditions. The generation process can be represented as $S_{inf} = M_G(C)$. Then, the generated spectrograms S_{inf} and the real spectrograms S_r will be used to train a verification model based on Siamese neural network [7]. Since the verification model aims to learn unobtrusive patterns that may appear anywhere in spectrograms, we use the advanced Vision Transformer (ViT) network as the backbone of the Siamese neural network, following the implementation of the open source repository [49].

We introduce contrastive learning between positive and negative samples to improve the verification performance and adopt triplet loss [60] as our training loss. Specifically, the ViT network takes each Anchor-Positive or Anchor-Negative pair as input, where the Anchor, Positive, and Negative are selected as follows:

- Anchor: The real spectrogram S_r of a legitimate user that corresponds to CIR profiles C.
- Positive: The inferred spectrogram S_{inf} generated from C and the legitimate user's M_G.
- Negative: The inferred spectrogram S'_{inf} generated from unpaired C' or M'_G, where C' denotes the CIR profiles of other people when uttering the same sentence, and M'_G represents the generative model of another person. Or the inferred spectrogram $\tilde{S}_{inf}$ generated from $\tilde{C}$ and the legitimate user's M_G, where $\tilde{C}$ indicates CIR profiles of the utterance of the legitimate user's other sentences.

For a given Anchor A_i, the Positive P_i is explicit. We randomly select a Negative N_i that satisfies the aforementioned situations. Subsequently, we can form two pairs: A_i-P_i and A_i-N_i. The Siamese network makes inferences twice on A_i-P_i and A_i-N_i pairs to obtain the deep features A_{iout}, P_{iout}, N_{iout} after ViT embedding. The triplet loss can be represented as:

$$\mathcal{L}_t = \sum_{i=1}^{N} [d(f(A_{iout}), f(P_{iout})) - d(f(A_{iout}), f(N_{iout})) + \alpha]_+,$$

where d represents the Euclidean distance, f denotes the transition of ViT, and N is the batch size. The $[]_+$ operation ensures non-negativity of the loss values by taking the maximum of the quantity inside and zero. The hyperparameter α controls the margin between positive and negative pairs. Then, the Euclidean distance between P_{iout} or N_{iout} and A_{iout} is regressed to a range of 0–1 through a fully connected layer and a sigmoid activation function to represent the similarity between them. Note that we set the decision threshold for determining the final results as a trainable parameter, which participates in the optimization process. This trainable decision threshold is initially set to 0.5. Finally, we optimize a Hinge loss between the predictions for positive and negative pairs to further enhance the model's ability to distinguish between them. The Hinge loss aims to increase the gap between the Siamese neural network output similarity and the decision threshold, ensuring that the model's predicted similarities for positive pairs are significantly above the threshold and for negative pairs well below it. The Hinge loss is computed as follows:

$$\mathcal{L}_{\text{hinge}} = \sum_{i=1}^{N} \big([\text{margin} - (\mathbf{Pred}(A_i, P_i) - \tau)]_+ + [\text{margin} + (\mathbf{Pred}(A_i, N_i) - \tau)]_+ \big),$$

where τ denotes the decision threshold, **Pred** presents the output similarity between inputs calculated by the Siamese network. Therefore, the objective function is $\mathcal{L} = \mathcal{L}_t + \mathcal{L}_{\text{hinge}}$.

The training settings for the verification model are similar to those of the generative model. Specifically, the batch size is set to 16, the learning rate is set to 0.0001 (0.001 for τ only), and the maximum epoch is set to 200. The margins used in Triplet loss and Hinge loss are 3.0 and 0.2 , respectively.

8 VERIFICATION PIPELINE

Fig. 10 shows the interactive verification pipeline of F²Key, which is comprised of two phases: a registration phase (dashed line) and a verification phase (solid line).

Registration Phase. F²Key first performs signal processing on the received audio. The low-pass and high-pass filters will separate user speech and ultrasonic waves from each other; the user speech and ultrasound are calculated in spectrograms and CIR profiles, respectively. Then the spectrograms and CIR profiles will be used to train a generative model that embeds the unique mapping relationship from CIR profiles to spectrograms. Thereafter, the user-specific generative model was frozen, and a verification model was trained using a Siamese neural network, triplet loss, and Hinge loss.

Verification Phase. When a user is speaking while wearing F²Key, the generative model generates an inferred spectrogram according to the embedded mapping relationship, and the verification model outputs the similarity between the generated and the real spectrogram to further indicate that the wearer's mapping relationship is consistent with the registered one or not. Finally, the system outputs the verification result—grant or deny.

When attacked, the CIR profiles of an attacker will differ from those of the victim, even when uttering the same sentence. Consequently, the generative model will convert the attacker's CIR profiles into an inferred spectrogram that cannot match the attacker's real spectrogram (in the case of zero-effort and mimicry attacks) or the victim's real spectrogram (in the case of replay and hybrid attacks). The victim's articulatory gestures cause variational CIR profiles C_L and result in a spectrogram S_L. The victim-registered generative model embeds the mapping relationship $\mathbf{G_L}$ that maps C_L to S_L such that $S_L \approx \mathbf{G_L}(C_L)$ when $\mathbf{Sim}(\mathbf{G_L}(C_L), S_L) > \tau$, where **Sim** and τ denote the calculation of similarity and the decision threshold, respectively. If an attacker attempts to access F²Key, they must provide variational CIR profiles and perform mimicry or hybrid attacks (§ 2.2). However, the attacker's CIR profiles C_A differ from those of the victim. Therefore, a fake inferred spectrogram $S_{inf_A} = \mathbf{G_L}(C_A)$ will neither match the victim's S_L (ideal playback in hybrid attack) due to the different CIR profiles provided nor match the attacker's S_A because $\mathbf{G_L}$ is registered to the victim's facial structures and articulatory habits.

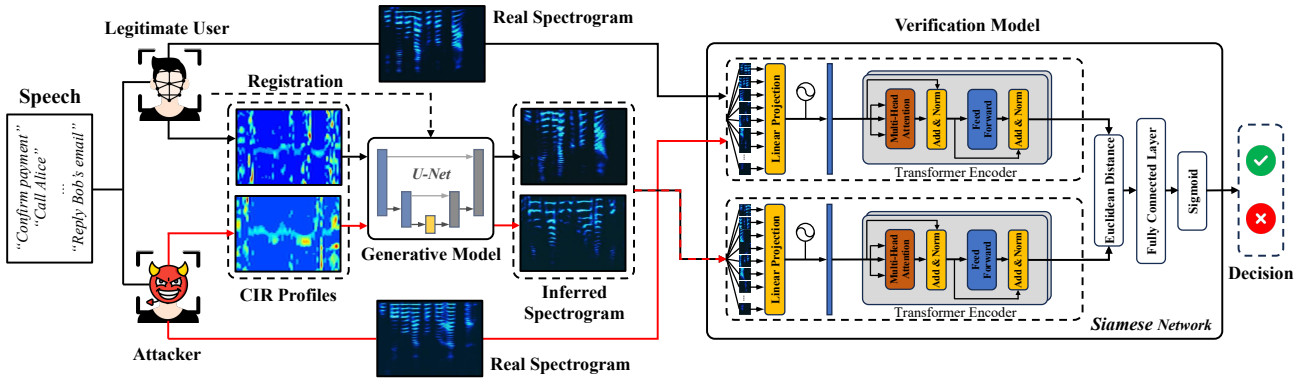


Figure 10: Verification pipeline of F²Key. The legitimate user and attackers will provide distinct CIR profiles based on their facial structures and articulatory habits. Verification will be granted only if the embedded CIR-spectrogram mapping relationship is consistent with that of the wearer.

9 EVALUATION

9.1 Experimental Setup

9.1.1 Methodology. Each participant was informed about the purpose and procedure of the experiment. An experimenter assisted each participant in wearing the F²Key hardware. Then, the experimenter controlled the hardware to emit FMCW signals and collected the participant's speech (along with ultrasound) using the boom microphone.

9.1.2 Dataset and Implementation. We recruited 26 participants (14 females, 12 males), aged 19 to 35, averaging 26.1 years (SD = 3.3), all experienced with headphones. The experiments were carried out in a room with the size of 4 × 4 meters. Each participant was asked to speak 15 security-related sentences (*i.e.*, "Confirm payment"), each sentence repeated 30 times[1], while wearing F²Key. Then, we built a dataset that contains 26 × 15 × 30 = 11700 utterances. For each participant, the data collection process took an average of 17.8 min to complete using an Antlion Mod Microphone [4] mounted on a Sony WH-1000XM4 [66]. We also implemented F²Key on two other COTS headphones (*i.e.*, Logitech G733 [48] and Audio-Technica ATH-G1WL [5]) to verify the generalization of F²Key on different headphones. We also leverage the dataset from previous study [19], which contains 13,680 five-second speech segments collected from speakers of various accents across 11 countries, to develop a pre-trained model that incorporates acoustic-speech priors.

Thereafter, we trained a user-specific generative model M_G and a verification model M_V for each participant. The input to the system is a segment of user speech along with the corresponding collected CIR profiles, and the output is the verification result. Our deep learning models were trained on a workstation equipped with an AMD Ryzen 3955WX, 4 × 64 GB of RAM, and three NVIDIA RTX 3090 GPUs with 24 GB of memory each. The models were trained with only one single 3090 GPU and implemented using the PyTorch framework version 1.13. Unless specified, the datasets used for training, validation, and testing in this section are randomly split in a ratio of 80%, 10%, 10%, respectively.

[1]Ethical approval has been obtained (No. H002969).

9.1.3 Evaluation Metrics. As a physical security system that aims to enable the verification of audio artifacts and enhance voiceprint-based authentication systems as the second factor of 2FA, we consider the following four metrics:

- True Acceptance Rate (TAR): The rate at which the legitimate user is granted access correctly.
- False Acceptance Rate (FAR): The rate at which attackers are granted access as a legitimate user.
- False Rejection Rate (FRR): The rate at which the legitimate user is denied access as an attacker.
- Equal Error Rate (EER): The point at which the FAR and FRR are equal. It represents the threshold at which the system's sensitivity and specificity are balanced.

9.2 Overall Performance

To evaluate the overall performance of F²Key among all participants. We assembled the trained M_G and M_V into an end-to-end model M for each participant. We used each participant's testing dataset D_{t_i} to test all user-specific models M_j, where i and j indicate the participant's ID number. We conducted 26 × (26 − 1) = **650** zero-effort attack experiments to demonstrate the specificity of each participant's CIR-spectrogram mapping relationship. Then, we obtained a heatmap shaped in 26 × 26, shown in Fig. 11. The values on the diagonal of the heatmap represent the TAR for each legitimate user with a minimum of 96% TAR. The other values represent the FAR of user-specific models when subjected to zero-effort attacks by attackers. This group of zero-effort attack experiments presents a series of FARs with a mean of 4.78%. Figs. 12 show that the curves of all participants follow similar trends. On average, F²Key achieves an EER of 2.7%. We also estimated the EER of each participant, averaging 2. 89% at a mean decision threshold of 0.53.

Despite these results demonstrating the verification capabilities of F²Key, using it solely for authentication poses risks (*e.g.*, FAR of 4.78%). F²Key is designed to enhance voice-print-based authentication and detect deepfakes. It compensates for the shortcomings of traditional voiceprint-based authentication, significantly reducing the risk of attacks and the threat posed by audio deepfakes.

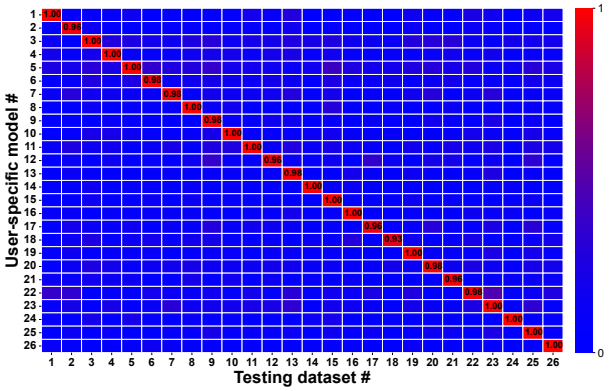
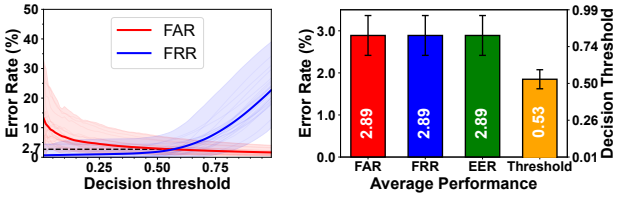


Figure 11: Overall performance heatmap.



(a) FAR and FRR. (b) Participants mean performance.

Figure 12: Average performance among 26 participants.

9.3 System Robustness

9.3.1 Impact of Unseen Sentence. To assess the practicality of F²Key, designed to address the challenge of speech verification regardless of speech content, we invited ten participants (from the 26 participants) to conduct a secondary unseen data collection. Each participant was asked to speak each of the 20 randomly selected sentences from the TIMIT corpus [28] three times. The TAR of their user-specific model was evaluated for each legitimate user using these 60 utterances. Furthermore, we conducted a zero-effort attack using other participants' utterances to assess the FAR. The three utterances of each sentence are subjected to a majority voting process to determine the final result. As Fig. 13(a) shows, the TAR is all above 89% with a mean value of 95.7%. Zero-effort attacks have a low passing rate (< 2%), even if sentences are not included in the training set. This experiment demonstrates that F²Key can be used in the open world.

9.3.2 Impact of Headphone Model. As mentioned in § 9.1.2, we implement F²Key on three COTS headphones. We selected two participants to collect two sets of data (450 sentences for each) using G733 and ATH-G1WL. Their data is used to attack each other to assess FAR. Fig. 13(b) illustrates the performance achieved by pairing a modular Antlion+XM4 is the best, owing to the latter's high-quality speakers with a broad frequency response of 4 Hz to 40,000 Hz, compared to the G733 and G1WL's range capping at 20,000 Hz. Therefore, the latter two devices exhibit lower performance (2.7%–4.8%) than the former due to greater signal distortions caused by speaker limitations.

9.3.3 Impact of Ambient Noise. We evaluate the performance of F²Key under three different types of noise: air conditioning noise (∼40 dB), soft instrumental music (∼50 dB), and competing speech (∼60 dB) at about 0.5 m from the hardware. Three participants are involved in this experiment; we test 20 sentences on their user-specific models without retraining or fine-tuning. From the result (Fig. 13(c)), we find that the operation of an air conditioner causes extra background noise in the real spectrograms, leading to a slight performance degradation (about 6%) in TAR. However, when faced with noise, such as music and competing speech, the mixing of noise directly changes the spectrograms, challenging F²Key from its principle. As a result, performance is severely impacted. It is a universal limitation in voiceprint-based systems; we will discuss the promising solution in § 11. Furthermore, when countering these troublesome noises, attackers will also fail.

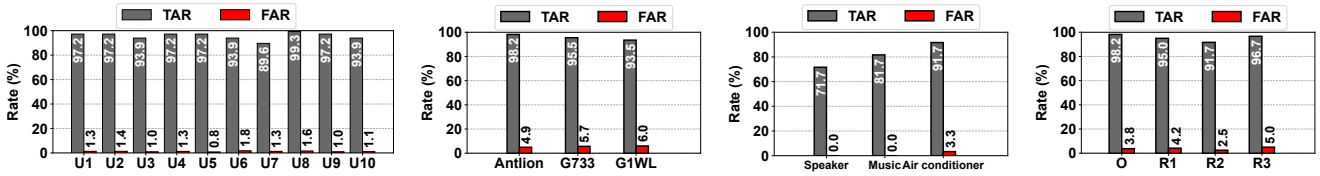
9.3.4 Impact of Re-wearing. To assess the robustness of F²Key under re-wear with an inevitable displacement each time, we invite three participants to take off and re-wear F²Key, repeating three times. We collect 20 sentences each time and test them using their user-specific models without retraining or fine-tuning. Fig. 13(d) shows that the re-wearing experiments present a TAR of 94.5% on average, and the mean FAR is 3.9%. In general, re-wearing does not significantly affect usability and security.

9.3.5 Impact of Facial Variations. In practice, a user's face may exhibit slight variations due to reasonable factors, such as wearing glasses or applying makeup. Since makeup primarily affects visual appearances without altering actual facial structures, it does not affect our system. To investigate the impact of common add-ons, we invite three participants to conduct a control experiment. Specifically, we collect 450 utterances from the three participants both with and without glasses. First, the data of a training group were used to fine-tune the basic pre-trained model to obtain a trained model, which was then tested directly with the control group's test data. Then, we switch the training and control groups and repeat the experiment. By averaging the absolute values of the differences in FRR of both tests, we evaluate the impact of wearing glasses as a facial variation. The results show that this condition increases the FRR by 0.36%, indicating a slight effect on the system performance. The reasons behind it can be found in Fig. 5(a), the areas around the user's eyes are less active when speaking. Moreover, since wearing glasses is a static feature that does not dynamically affect multipath during speech, its influence on extracting CIR-spectrogram mapping relationships is negligible.

9.4 Attack Experiments

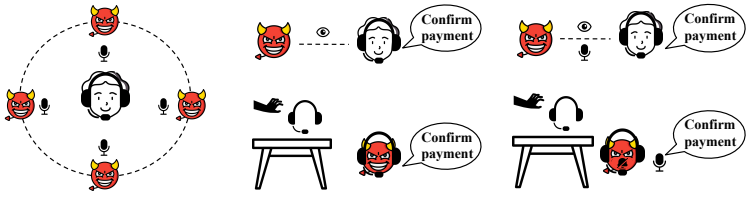
In this section, we consider the most promising attacks to challenge F²Key. Since there are bound to be nuances between the synthesized and real speech of a victim, and given the varied and often unfair synthesis methods, we do not consider synthetic attacks. Instead, we directly challenge the most intractable replay and hybrid attacks that use real speech. The results are integrated in Fig. 15.

9.4.1 Replay Attack. Fig. 14(a) illustrates the replay attack scenarios where an attacker records audio from the victim's headphones using a microphone positioned 0.5 meters away in four directions



(a) Unseen sentence impact. (b) Headphone model impact. (c) Ambient noise impact. (d) Re-wearing impact.

Figure 13: F^{2}Key robustness.



(a) Replay attacks. (b) Mimicry attacks. (c) Hybrid attacks.

Figure 14: Attack experiments scenarios.

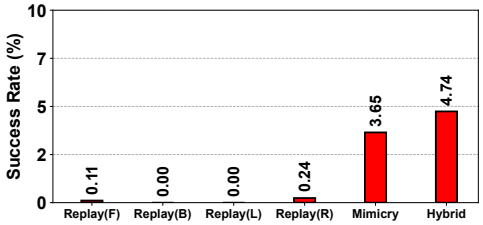


Figure 15: Attack experiments results.

(front, back, left, and right). The attacker then blocks the headphone's earpad with soundproof material and plays the recorded audio to attack F^{2}Key. However, F^{2}Key successfully resists these attacks, with a success rate below 0.3%. The reason is that the leaked FMCW signals did not sweep across the user's face and did not yield informative CIR profiles, especially when recording from the front, back, and left. Additionally, audio from the boom microphone side (right) has poor resolution for small-scale perturbations and extremely low SNR in the FMCW signals, rendering the recorded audio unable to effectively breach F^{2}Key's defense.

9.4.2 Mimicry Attacks. Fig. 14(b) illustrates the mimicry attack scenarios where attackers observe the victim, including facial gestures and speech, and then try to access F^{2}Key by imitating these observations when unattended. The result shows that by intentionally mimicking the victim's articulatory gestures and timbre, the success rate of the mimicry attack presents an increase of about 0.9% compared to the zero-effort attack elaborated in § 9.2. Although attackers can observe the victim freely, the similarity achieved by imitation is limited without professional training, allowing F^{2}Key to resist 96.4% of mimicry attacks.

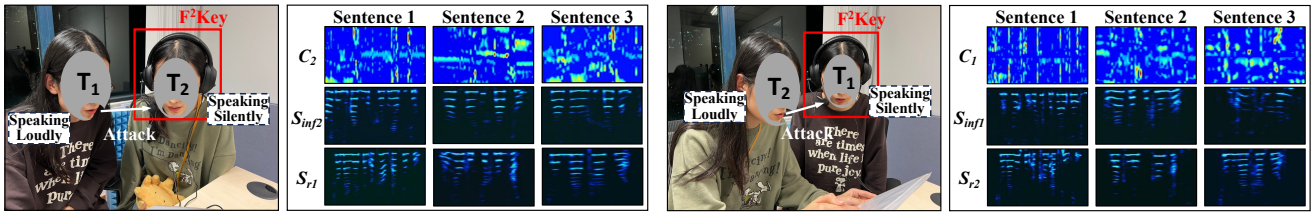
9.4.3 Hybrid Attacks. Fig. 14(c) illustrates the hybrid attack scenarios similar to those of mimicry attacks. The difference is that the attacker replays the recording of the victim's speech and simultaneously performs silent speech. The results show that even if the victim's voiceprint is stolen, the attacker's unregistered facial biometric information fails to provide effective CIR profiles and further match the CIR-spectrogram mapping relationship embedded in the generative model. As such, F^{2}Key can resist 95.3% hybrid attacks and significantly improve the reliability of voiceprint-based authentication and the anti-counterfeiting of audio artifacts.

9.5 Identical Twins Study

Among our 26 participants, there is a special case that the 22nd and 23rd participants are a pair of identical twins. They have extremely similar facial features. To verify the similarity of their faces, we tested the Face ID function [1] on each of their iPhones. The results showed that both could unlock each other's iPhones within three attempts. In such a special case, the authentication methods based solely on facial images or structure become ineffective; furthermore, liveness detection also fails to compensate for this.

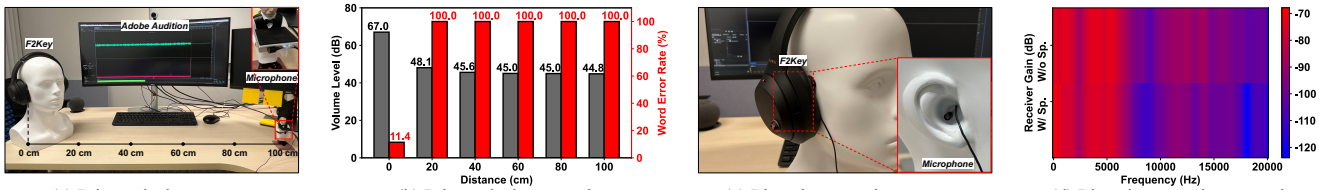
We consider an extremely special case—*the spoofing attack between identical twins*. Specifically, we invited the twins to repeat the content of the experiment in § 9.4.3, as shown in Figs. 16(a) and 16(c). One of them was wearing F^{2}Key and speaking in silence, while the other was looking at the same text, loudly reading the same sentences. Note that this is different from § 9.4.3, the reasons are two-fold: **(1)** Different from playingback the previously recorded audio on electronic devices. This experiment uses the live voice of a victim. In this situation, even the most advanced voice-liveness detection methods will fail. **(2)** The two individuals have extremely similar faces; therefore, the "key" embedded in the generative model may be more similar.

The results show in Figs. 16(b) and 16(d), we find that there are noticeable differences between the inferred spectrogram and the real one produced by the attacker; the attack experiment achieves an average success rate of 24.5%. The reason is that although the twins have very similar facial features, there are noticeable differences in their timbre of speech and habits in articulatory gestures. Therefore, the CIR-spectrogram mapping relationships have significant differences. These experiments were carried out merely to investigate the effectiveness of the proposed joint defense mechanism in defending against attacks between identical twins with highly similar facial features. However, in real-world scenarios, it is impossible for a



(a) Identical twin T_1 attacks T_2 (b) $T_1 \rightarrow T_2$ results (c) Identical twin T_2 attacks T_1 (d) $T_2 \rightarrow T_1$ results

Figure 16: Attack experiments between identical twins. (a) and (c) show the experimental scenarios in which identical twins attack each other; (b) and (d) present the results.



(a) Privacy leakage set up (b) Privacy leakage result (c) Listening experience set up (d) Listening experience result

Figure 17: The impact of auxiliary spacers. (a) and (c) show the experimental scenarios that assess the privacy leakage risk and listening experience degradation; (b) and (d) present the corresponding results.

victim to collaborate with an attacker to compromise their own devices.

9.6 Computational Delay

We evaluated F²Key's computational delay on four devices (*i.e.*, a laptop with an RTX 2080Ti GPU, a server with an RTX 3090 GPU, Samsung S10, and OnePlus 8T), focusing on three stages: signal processing, generative model, and verification model. Signal processing takes about 244 ms, while the generative model needs 198 ms on the 3090 GPU and 318 ms on the 2080Ti. The verification model needs a minimal delay (12.3 ms on 3090, 14.6 ms on 2080Ti). F²Key processes up to 5 s speech clips under 0.6 s if GPUs are available. For the resource-constrained Samsung S10 and OnePlus 8T, the overall computational delays are 1.97 s and 1.67 s, respectively. F²Key facilitates real-time voiceprint authentication and deepfake defense in VoIP applications, and can be used on both local mobile devices and cloud servers.

9.7 Auxiliary Spacers Impact Study

In this section, we assess the risk of privacy leakage and the impact on listening experience caused by auxiliary spacers, the playback volume is set to 50% of the maximum volume.

9.7.1 Privacy Leakage Study. Fig. 17(a) illustrates a test where a microphone, placed at distances from 20 cm to 100 cm from F²Key mounted on a head model, captures sound leakage. The test measured sound leakage from F²Key, from Fig. 17(b), we observed that the volume of the music decreased and plateaued beyond 40 cm. Speech recognition[2] was evaluated using the word error rate (WER) metric, the results showing that the ASR system failed to recognize speech due to the low SNR (the WER defined as 100%), which confirms the limited privacy risks associated with F²Key.

[2]The speech recognition is performed by Google Cloud Speech API [29].

9.7.2 Listening Experience Study. To evaluate sound reception, we used a 10 s chirp spanning 20 Hz to 20,000 Hz. The chirp, played by F²Key, was recorded by a microphone attached to the external auditory canal of a head model, as shown in Fig. 17(c). The receiver gain analysis measured in dB is presented in Fig. 17(d), which shows consistent receiver gain patterns across frequencies, with a minor drop at higher frequencies when using spacers, particularly above 8,000 Hz. However, since most human audio lies below 8,000 Hz [11], the auxiliary spacers have a negligible effect on the intelligibility of the audio but can affect the listening experience of some high-fidelity music, which is a minor limitation of F²Key.

10 RELATED WORK

Audiovisual Deepfake Detection. Fake media creation, especially audiovisual deepfakes, is an active research area, thus leading to significant efforts in the anti-counterfeiting of artifacts. Existing deepfake detection methods for video rely mainly on inconsistency detection (*e.g.*, artifact warping [43], blending boundaries [42], and fingerprinting [86]) and metadata embedding (*e.g.*, provenance credential [47], quick response code [46, 62], and temporal content hashes [9]). However, deepfake detection in the audio domain remains relatively underdeveloped [85]. The existing methods rely on spectral features in terms of magnitude spectrum [79], phase spectrum [78], and modulation spectrum [58] to detect spoofing attacks. Recent research has focused on more discriminative deep features that can be learned, such as relationships among breathing, talking, and silence sounds [18], utterance level average [15]. In contrast to these methods, F²Key establishes a stable acoustic sensing field to link user facial structures and articulatory habits with the corresponding speech in a multi-modality way, allowing anti-spoofing user authentication and audio deepfake defense.

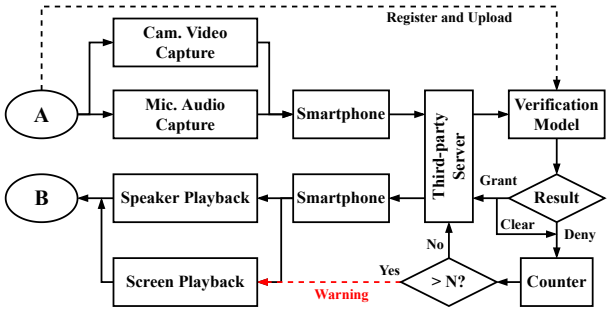


Figure 18: Integration flowchart. The flowchart shows the principle of integrating F^{2}Key into the third-party server to termly detect audio deepfake.

Earable Sensing. Recently, researchers have explored earable devices, given their portability and potential for various applications, such as on-face interaction [81], face reconstruction [41, 77], speech enhancement [14, 31, 59], and silent command recognition [38, 67]. In numerous studies, they can be divided into two categories, custom prototypes and COTS devices. Some researchers fabricated custom prototypes for novel applications, such as blood pressure measurement [8], microsleep event detection [53], eavesdropping [45]. Furthermore, numerous studies have incorporated in-ear microphones into COTS earbuds to explore wide-ranging applications in areas such as health monitoring [8, 10, 37], behavior recognition [36, 50, 54, 65], user authentication [21, 26, 27, 75, 76, 80], etc. However, these studies rely on custom hardware or require earphone remodeling. Therefore, considering the constraints of dedicated and remodeled earable devices, researchers have shifted their focus to implementing intelligent applications on COTS earable devices. EarphoneTrack [12] proposed an innovative acoustic motion tracking approach using earphones; HeadFi [20] measured the imbalance between two earphones using a Wheatstone Bridge, allowing applications such as user identification, heart rate monitoring, and gesture recognition; FaceOri [73] achieved head pose detection by coordinating with a smartphone. However, to our knowledge, no COTS-headphone-based physical security system has been proposed. We propose F^{2}Key, which is designed for the anti-counterfeiting of artifacts, ensuring reliable voice interaction, and enhancing voiceprint-based authentication systems.

11 DISCUSSION & FUTURE WORK

Integration with Mobility Systems. To illustrate the integration of F^{2}Key with existing mobility systems and voice-based applications, Fig. 18 shows the flowchart that combines F^{2}Key with smartphones in Voice over Internet Protocol (VoIP) applications.

When user A and user B make an Internet call, the voice or video of user A is transmitted to a third-party server for relay via the Internet, and then transmitted to user B's smartphone. In this process, the verification of F^{2}Key can be integrated into the third-party server. Specifically, both users should register and upload their user-specific model that embeds the unique CIR-spectrogram mapping relationships in advance. During the call, the server captures the user's voice at fixed intervals (*e.g.*, five seconds without overlap) and verifies it; if N consecutive segments are all denied, the system highly suspects that the user's voice may involve deepfake and sends a warning to the client of the other party. When the status indicator that represents credibility on the call interface changes, one should be alert to the risk of fraud.

Ambient Noise Issue. As mentioned in § 9.3.3, ambient noise, such as competing speech, will pollute the real spectrogram and severely affect the performance. Although it is a universal limitation for all voiceprint-based systems, the hardware configuration of F^{2}Key has great potential to address it. In our previous work EarSE [19], we used the same hardware as F^{2}Key to explore the possibility of multi-modal speech enhancement. Since only the user's articulatory gestures perturb the acoustic field, affecting the CIR profiles, while ambient noise does not, ambient noise can be isolated and user speech can be filtered in for enhancement. The results indicate that the hardware setting can increase SiSDR by 14.61 dB when facing competing speakers and background noise. With the support of multi-modal speech enhancement, the impact of noise will be mitigated to some extent. Given that EarSE and F^{2}Key are independent systems that perform different tasks, utilize the same hardware. The effective integration of them will be the main focus of our future research.

Comfort and Headphone Types. Besides the above discussion, we also investigated the comfort of the hardware configuration in previous work. The survey shows that about 70% users regard the hardware to be comfortable to wear [19]. Although F^{2}Key can be established on COTS headphones without a boom microphone (*e.g.*, Sony WH-1000XM4), its application to earbuds still presents limitations. Adding an outward-facing speaker module to the earbuds and combining it with the ambient noise microphone of the opposite earbud to establish the acoustic sensing field presents a promising solution.

12 CONCLUSION

In this paper, we proposed the first earable physical security system, F^{2}Key. It leverages the form factors of COTS headphones equipped with a boom microphone and auxiliary spacers to create a stable acoustic sensing field across the user's face, which enables fine-grained articulatory gesture detection. We modeled the relationship between facial structures, articulatory habits, and the corresponding speech using a challenge-response mechanism and embed the mapping relationship into a generative model for further verification. The proposed verification model will grant access or confirm the authenticity, provided the wearer's CIR-spectrogram mapping relationship is consistent with the embedded one. The results demonstrate that F^{2}Key has both reliability and practicality; it can resist 99.9%, 96.4%, and 95.3% of speech replay, mimicry, and hybrid attacks, respectively.

ACKNOWLEDGMENTS

The work was supported by the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. CityU 21201420 and CityU 11201422). The work was also partially supported by CityU APRC grant 9610471, CityU MFPRC grant 9680333, CityU SIRG grant 7020057, CityU SRG-Fd grant 7005666 and 7005984.

REFERENCES

[1] 2023. Face ID Security. Apple Support Page. https://support.apple.com/en-us/102381 Accessed on: 2023-10-15.

[2] Lawrence Abrams. 2023. Elon Musk Deep Fakes Promote New BitVex Cryptocurrency Scam. https://www.bleepingcomputer.com/news/security/elon-musk-deep-fakes-promote-new-bitvex-cryptocurrency-scam Accessed: 2023-11-16.

[3] Sercan Arik, Jitong Chen, Kainan Peng, Wei Ping, and Yanqi Zhou. 2018. Neural voice cloning with a few samples. *Advances in neural information processing systems (NeurIPS)* 31 (2018).

[4] Antlion Audio. 2023. *Antilion Mod Mic.* Retrieved April 6, 2023 from https://antlionaudio.com/collections/microphones/products/modmic-usb

[5] Audio-Technica. 2023. *ATH-G1WL.* Retrieved April 6, 2023 from https://www.audio-technica.com/en-us/ath-g1wl

[6] G. Bradski. 2000. The OpenCV Library. *Dr. Dobb's Journal of Software Tools* (2000).

[7] Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. 1993. Signature verification using a" siamese" time delay neural network. *Advances in neural information processing systems (NeurIPS)* 6 (1993).

[8] Nam Bui, Nhat Pham, Jessica Jacqueline Barnitz, Zhanan Zou, Phuc Nguyen, Hoang Truong, Taeho Kim, Nicholas Farrow, Anh Nguyen, Jianliang Xiao, et al. 2019. ebp: A wearable system for frequent and comfortable blood pressure monitoring from user's ear. In *Proceedings of the 25th annual international conference on mobile computing and networking (MobiCom)*. 1–17.

[9] Tu Bui, Daniel Cooper, John Collomosse, Mark Bell, Alex Green, John Sheridan, Jez Higgins, Arindra Das, Jared Keller, Olivier Thereaux, et al. 2019. Archangel: Tamper-proofing video archives using temporal content hashes on the blockchain. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 0–0.

[10] Kayla-Jade Butkow, Ting Dang, Andrea Ferlini, Dong Ma, and Cecilia Mascolo. 2023. hEARt: Motion-resilient Heart Rate Monitoring with In-ear Microphones. In *Proceedings of the IEEE International Conference on Pervasive Computing and Communications (PerCom)*. IEEE, 200–209.

[11] Chao Cai, Rong Zheng, and Jun Luo. 2022. Ubiquitous acoustic sensing on commodity iot devices: A survey. *IEEE Communications Surveys & Tutorials* 24, 1 (2022), 432–454.

[12] Gaoshuai Cao, Kuang Yuan, Jie Xiong, Panlong Yang, Yubo Yan, Hao Zhou, and Xiang-Yang Li. 2020. Earphonetrack: involving earphones into the ecosystem of acoustic motion tracking. In *Proceedings of the 18th Conference on Embedded Networked Sensor Systems (SenSys)*. 95–108.

[13] Hanqun Cao, Cheng Tan, Zhangyang Gao, Yilun Xu, Guangyong Chen, Pheng-Ann Heng, and Stan Z Li. 2022. A survey on generative diffusion model. *arXiv preprint arXiv:2209.02646* (2022).

[14] Ishan Chatterjee, Maruchi Kim, Vivek Jayaram, Shyamnath Gollakota, Ira Kemelmacher, Shwetak Patel, and Steven M Seitz. 2022. ClearBuds: wireless binaural earbuds for learning-based speech enhancement. In *Proceedings of the 20th Annual International Conference on Mobile Systems, Applications and Services (MobiSys)*. 384–396.

[15] Nanxin Chen, Yanmin Qian, Heinrich Dinkel, Bo Chen, and Kai Yu. 2015. Robust deep feature for spoofing detection—The SJTU system for ASVspoof 2015 challenge. In *Sixteenth Annual Conference of the International Speech Communication Association*.

[16] Qi Chen, Mingkui Tan, Yuankai Qi, Jiaqiu Zhou, Yuanqing Li, and Qi Wu. 2022. V2C: visual voice cloning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 21242–21251.

[17] Kun Cheng, Xiaodong Cun, Yong Zhang, Menghan Xia, Fei Yin, Mingrui Zhu, Xuan Wang, Jue Wang, and Nannan Wang. 2022. Videoretalking: Audio-based lip synchronization for talking head video editing in the wild. In *Proceedings of the SIGGRAPH Asia 2022 Conference Papers (SIGGRAPH-Asia)*. 1–9.

[18] Thien-Phuc Doan, Long Nguyen-Vu, Souhwan Jung, and Kihun Hong. 2023. BTS-E: Audio Deepfake Detection Using Breathing-Talking-Silence Encoder. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.

[19] Di Duan, Yongliang Chen, Weitao Xu, and Tianxing Li. 2024. EarSE: Bringing Robust Speech Enhancement to COTS Headphones. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)* 7, 4 (2024), 1–33.

[20] Xiaoran Fan, Longfei Shangguan, Siddharth Rupavatharam, Yanyong Zhang, Jie Xiong, Yunfei Ma, and Richard Howard. 2021. HeadFi: bringing intelligence to all headphones. In *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking (MobiCom)*. 147–159.

[21] Andrea Ferlini, Dong Ma, Robert Harle, and Cecilia Mascolo. 2021. EarGate: gait-based user identification with in-ear microphones. In *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking (MobiCom)*. 337–349.

[22] Pablo Ferrezuelo. 2023. Why Deepfake Fraud Losses Should Scare Financial Institutions. (2023). https://www.finextra.com/blogposting/23223/why-deepfake-fraud-losses-should-scare-financial-institutions Accessed: 2023-11-17.

[23] Panagiotis P Filntisis, George Retsinas, Foivos Paraperas-Papantoniou, Athanasios Katsamanis, Anastasios Roussos, and Petros Maragos. 2023. SPECTRE: Visual Speech-Informed Perceptual 3D Facial Expression Reconstruction From Videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 5744–5754.

[24] Emily Flitter and Stacy Cowley. 2023. Voice Deepfakes Are Coming for Your Bank Balance. (2023). https://www.nytimes.com/2023/08/30/business/voice-deepfakes-bank-scams.html Accessed: 2023-11-17.

[25] Centers for Disease Control and Prevention. 2023. Public Health and Scientific Information. https://www.cdc.gov/nceh/hearing_loss/public_health_scientific_info.html Accessed: 2023-07-31.

[26] Yang Gao, Yincheng Jin, Jagmohan Chauhan, Seokmin Choi, Jiyang Li, and Zhanpeng Jin. 2021. Voice in ear: Spoofing-resistant and passphrase-independent body sound authentication. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)* 5, 1 (2021), 1–25.

[27] Yang Gao, Wei Wang, Vir V Phoha, Wei Sun, and Zhanpeng Jin. 2019. EarEcho: Using ear canal echo for wearable authentication. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)* 3, 3 (2019), 1–24.

[28] John S Garofolo et al. 1988. DARPA TIMIT acoustic-phonetic speech database. *National Institute of Standards and Technology (NIST)* 15 (1988), 29–50.

[29] Google. 2023. Google Cloud Speech-to-Text. https://cloud.google.com/speech-to-text/. Accessed: 2023-11-13.

[30] Alexandros Haliassos, Rodrigo Mira, Stavros Petridis, and Maja Pantic. 2022. Leveraging real talking faces via self-supervision for robust forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 14950–14962.

[31] Lixing He, Haozheng Hou, Shuyao Shi, Xian Shuai, and Zhenyu Yan. 2023. Towards Bone-Conducted Vibration Speech Enhancement on Head-Mounted Wearables. In *Proceedings of the 21st Annual International Conference on Mobile Systems, Applications and Services*. 14–27.

[32] Chien-yu Huang, Yist Y Lin, Hung-yi Lee, and Lin-shan Lee. 2021. Defending your voice: Adversarial attack on voice conversion. In *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 552–559.

[33] iperov. 2023. DeepFaceLab is the leading software for creating deepfakes. https://github.com/iperov/DeepFaceLab.

[34] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. 2017. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*. 1125–1134.

[35] Ye Jia, Yu Zhang, Ron Weiss, Quan Wang, Jonathan Shen, Fei Ren, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno, Yonghui Wu, et al. 2018. Transfer learning from speaker verification to multispeaker text-to-speech synthesis. *Advances in neural information processing systems (NeurIPS)* 31 (2018).

[36] Nan Jiang, Terence Sim, and Jun Han. 2022. EarWalk: towards walking posture identification using earables. In *Proceedings of the 23rd Annual International Workshop on Mobile Computing Systems and Applications (HotMobile)*. 35–40.

[37] Yincheng Jin, Yang Gao, Xiaotao Guo, Jun Wen, Zhengxiong Li, and Zhanpeng Jin. 2022. EarHealth: an earphone-based acoustic otoscope for detection of multiple ear diseases in daily life. In *Proceedings of the 20th Annual International Conference on Mobile Systems, Applications and Services (MobiSys)*. 397–408.

[38] Yincheng Jin, Yang Gao, Xuhai Xu, Seokmin Choi, Jiyang Li, Feng Liu, Zhengxiong Li, and Zhanpeng Jin. 2022. EarCommand: " Hearing" Your Silent Speech Commands In Ear. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)* 6, 2 (2022), 1–28.

[39] Vahid Kazemi and Josephine Sullivan. 2014. One millisecond face alignment with an ensemble of regression trees. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*. 1867–1874.

[40] Jan Niklas Kolf, Tim Rieber, Jurek Elliesen, Fadi Boutros, Arjan Kuijper, and Naser Damer. 2023. Identity-driven Three-Player Generative Adversarial Network for Synthetic-based Face Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 806–816.

[41] Ke Li, Ruidong Zhang, Bo Liang, François Guimbretière, and Cheng Zhang. 2022. Eario: A low-power acoustic sensing earable for continuously tracking detailed facial movements. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)* 6, 2 (2022), 1–24.

[42] Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo. 2020. Face x-ray for more general face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*. 5001–5010.

[43] Yuezun Li and Siwei Lyu. 2018. Exposing deepfake videos by detecting face warping artifacts. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*.

[44] Zhuohang Li, Cong Shi, Tianfang Zhang, Yi Xie, Jian Liu, Bo Yuan, and Yingying Chen. 2021. Robust detection of machine-induced audio attacks in intelligent audio systems with microphone array. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security (CCS)*. 1884–1899.

[45] Qianru Liao, Yongzhi Huang, Yandao Huang, Yuheng Zhong, Huitong Jin, and Kaishun Wu. 2022. MagEar: eavesdropping via audio recovery using magnetic side channel. In *Proceedings of the 20th Annual International Conference on Mobile*

Systems, Applications and Services (MobiSys). 371–383.
[46] Xiaomei Liu and Xin Tang. 2020. Image authentication using QR code watermarking approach based on image segmentation. In *2020 IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*. IEEE, 1572–1577.
[47] Yuxin Liu, Yoshimichi Nakatsuka, Ardalan Amiri Sani, Sharad Agarwal, and Gene Tsudik. 2022. Vronicle: verifiable provenance for videos from mobile devices. In *Proceedings of the 20th Annual International Conference on Mobile Systems, Applications and Services (MobiSys)*. 196–208.
[48] Logitech. 2023. *G733*. Retrieved April 6, 2023 from https://www.logitechg.com/en-us/products/gaming-audio/g733-rgb-wireless-headset.981-000863.html
[49] lucidrains. 2020. vit-pytorch: An implementation of Vision Transformers in PyTorch. https://github.com/lucidrains/vit-pytorch.
[50] Dong Ma, Andrea Ferlini, and Cecilia Mascolo. 2021. OESense: employing occlusion effect for in-ear human sensing. In *Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services (MobiSys)*. 175–187.
[51] Brian B Monson, Eric J Hunter, Andrew J Lotto, and Brad H Story. 2014. The perceptual significance of high-frequency energy in the human voice. *Frontiers in psychology* 5 (2014), 587.
[52] Harry Nyquist. 1928. Certain topics in telegraph transmission theory. *Transactions of the American Institute of Electrical Engineers* 47, 2 (1928), 617–644.
[53] Nhat Pham, Tuan Dinh, Taeho Kim, Zohreh Raghebi, Nam Bui, Hoang Truong, Tuan Nguyen, Farnoush Banaei-Kashani, Ann Halbower, Thang N Dinh, et al. 2021. Detection of Microsleep Events with a Behind-the-ear Wearable System. *IEEE Transactions on Mobile Computing (TMC)* (2021).
[54] Jay Prakash, Zhijian Yang, Yu-Lin Wei, Haitham Hassanieh, and Romit Roy Choudhury. 2020. EarSense: earphones as a teeth activity sensor. In *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking (MobiCom)*. 1–13.
[55] D. Purves, G.J. Augustine, D. Fitzpatrick, et al. 2001. *The Audible Spectrum*. Sinauer Associates.
[56] Radiopaedia. 2023. *Acoustic impedance*. Retrieved November 17, 2023 from https://radiopaedia.org/articles/acoustic-impedance
[57] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Springer, 234–241.
[58] Md Sahidullah, Tomi Kinnunen, and Cemal Hanilçi. 2015. A comparison of features for synthetic speech detection. (2015).
[59] Philipp Schilk, Niccolò Polvani, Andrea Ronco, Milos Cernak, and Michele Magno. 2023. In-Ear-Voice: Towards Milli-Watt Audio Enhancement With Bone-Conduction Microphones for In-Ear Sensing Platforms. In *Proceedings of the 8th ACM/IEEE Conference on Internet of Things Design and Implementation (IOTDI)*. 1–12.
[60] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*. 815–823.
[61] Avast security news team. 2023. Voice fraud scams company out of $243,000. (2023). https://blog.avast.com/deepfake-voice-fraud-causes-243k-scam Accessed: 2023-11-17.
[62] Irtaza Shahid and Nirupam Roy. 2023. " Is this my president speaking?" Tamper-proofing Speech in Live Recordings. In *Proceedings of the 21st Annual International Conference on Mobile Systems, Applications and Services (MobiSys)*. 219–232.
[63] Claude E Shannon. 1949. Communication in the presence of noise. *Proceedings of the IRE* 37, 1 (1949), 10–21.
[64] Sayaka Shiota, Fernando Villavicencio, Junichi Yamagishi, Nobutaka Ono, Isao Echizen, and Tomoko Matsui. 2015. Voice liveness detection algorithms based on pop noise caused by human breath for automatic speaker verification. In *Sixteenth annual conference of the international speech communication association*.
[65] Xingzhe Song, Kai Huang, and Wei Gao. 2022. FaceListener: Recognizing Human Facial Expressions via Acoustic Sensing on Commodity Headphones. In *Proceedings of the 21st ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*. IEEE, 145–157.
[66] Sony. 2023. *WH-1000XM4*. Retrieved April 6, 2023 from https://electronics.sony.com/audio/headphones/headband/p/wh1000xm4-b
[67] Tanmay Srivastava, Prerna Khanna, Shijia Pan, Phuc Nguyen, and Shubham Jain. 2022. MuteIt: Jaw Motion Based Unvoiced Command Recognition Using Earable. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)* 6, 3 (2022), 1–26.
[68] Ke Sun, Ting Zhao, Wei Wang, and Lei Xie. 2018. Vskin: Sensing touch gestures on surfaces of mobile devices using acoustic signals. In *Proceedings of the 24th Annual International Conference on Mobile Computing and Networking (MobiCom)*. 591–605.
[69] Midjourney Team. 2023. Midjourney Image Generation. https://www.midjourney.com/ Accessed: 2023-11-18.
[70] Ingo R Titze and Daniel W Martin. 1998. Principles of voice production.
[71] Jiadong Wang, Xinyuan Qian, Malu Zhang, Robby T Tan, and Haizhou Li. 2023. Seeing What You Said: Talking Face Generation Guided by a Lip Reading Expert. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 14653–14662.
[72] Tianben Wang, Daqing Zhang, Yuanqing Zheng, Tao Gu, Xingshe Zhou, and Bernadette Dorizzi. 2018. C-FMCW based contactless respiration detection using acoustic signal. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)* 1, 4 (2018), 1–20.
[73] Yuntao Wang, Jiexin Ding, Ishan Chatterjee, Farshid Salemi Parizi, Yuzhou Zhuang, Yukang Yan, Shwetak Patel, and Yuanchun Shi. 2022. FaceOri: Tracking Head Position and Orientation Using Ultrasonic Ranging on Earphones. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*. 1–12.
[74] Yuanda Wang, Hanqing Guo, Guangjing Wang, Bocheng Chen, and Qiben Yan. 2023. VSMask: Defending Against Voice Synthesis Attack via Real-Time Predictive Perturbation. *arXiv preprint arXiv:2305.05736* (2023).
[75] Zi Wang, Yili Ren, Yingying Chen, and Jie Yang. 2022. Toothsonic: Earable authentication via acoustic toothprint. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)* 6, 2 (2022), 1–24.
[76] Zi Wang, Sheng Tan, Linghan Zhang, Yili Ren, Zhi Wang, and Jie Yang. 2021. EarDynamic: An Ear Canal Deformation Based Continuous User Authentication Using In-Ear Wearables. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)* 5, 1 (2021), 1–27.
[77] Yi Wu, Vimal Kakaraparthi, Zhuohang Li, Tien Pham, Jian Liu, and Phuc Nguyen. 2021. BioFace-3D: continuous 3d facial reconstruction through lightweight single-ear biosensors. In *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking (MobiCom)*. 350–363.
[78] Zhizheng Wu, Xiong Xiao, Eng Siong Chng, and Haizhou Li. 2013. Synthetic speech detection using temporal modulation feature. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 7234–7238.
[79] Xiong Xiao, Xiaohai Tian, Steven Du, Haihua Xu, Engsiong Chng, and Haizhou Li. 2015. Spoofing speech detection using high dimensional magnitude and phase features: the NTU approach for ASVspoof 2015 challenge. In *Interspeech*. 2052–2056.
[80] Yadong Xie, Fan Li, Yue Wu, Huijie Chen, Zhiyuan Zhao, and Yu Wang. 2022. TeethPass: Dental Occlusion-based User Authentication via In-ear Acoustic Sensing. In *Proceedings of the 2022-IEEE Conference on Computer Communications (INFOCOM)*. IEEE, 1789–1798.
[81] Xuhai Xu, Haitian Shi, Xin Yi, WenJia Liu, Yukang Yan, Yuanchun Shi, Alex Mariakakis, Jennifer Mankoff, and Anind K Dey. 2020. Earbuddy: Enabling on-face interaction via wireless earbuds. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*. 1–14.
[82] Ryoya Yaguchi, Sayaka Shiota, Nobutaka Ono, and Hitoshi Kiya. 2019. Replay attack detection using generalized cross-correlation of stereo signal. In *2019 27th European Signal Processing Conference (EUSIPCO)*. IEEE, 1–5.
[83] Wenyuan Yang, Xiaoyu Zhou, Zhikai Chen, Bofei Guo, Zhongjie Ba, Zhihua Xia, Xiaochun Cao, and Kui Ren. 2023. AVoiD-DF: Audio-Visual Joint Learning for Detecting Deepfake. *IEEE Transactions on Information Forensics and Security (TIFS)* 18 (2023), 2015–2029.
[84] Xiaolong Yang, Xiaohong Jia, Dihong Gong, Dong-Ming Yan, Zhifeng Li, and Wei Liu. 2023. LARNeXt: End-to-End Lie Algebra Residual Network for Face Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* (2023).
[85] Jiangyan Yi, Chenglong Wang, Jianhua Tao, Xiaohui Zhang, Chu Yuan Zhang, and Yan Zhao. 2023. Audio Deepfake Detection: A Survey. *arXiv preprint arXiv:2308.14970* (2023).
[86] Ning Yu, Larry S Davis, and Mario Fritz. 2019. Attributing fake images to gans: Learning and analyzing gan fingerprints. In *Proceedings of the IEEE/CVF international conference on computer vision*. 7556–7566.
[87] Linghan Zhang, Sheng Tan, Jie Yang, and Yingying Chen. 2016. Voicelive: A phoneme localization based liveness detection for voice authentication on smartphones. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. 1080–1091.
[88] Yipin Zhou and Ser-Nam Lim. 2021. Joint audio-visual deepfake detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (CVPR)*. 14800–14809.
[89] Jun-Yan Zhu and Taesung Park. 2023. pytorch-CycleGAN-and-pix2pix: Image-to-Image Translation in PyTorch. https://github.com/junyanz/pytorch-CycleGAN-and-pix2pix.