# MagLive: Robust Voice Liveness Detection on Smartphones Using Magnetic Pattern Changes

Xiping Sun, Jing Chen, *Senior Member, IEEE*, Cong Wu, Kun He, *Member, IEEE*, Haozhe Xu, Yebo Feng, Ruiying Du, and Xianhao Chen, *Member, IEEE*

***Abstract*—Voice authentication has been widely used on smartphones. However, it remains vulnerable to spoofing attacks, where the attacker replays recorded voice samples from authentic humans using loudspeakers to bypass the voice authentication system. In this paper, we present MagLive, a robust voice liveness detection scheme designed for smartphones to mitigate such spoofing attacks. MagLive leverages the differences in magnetic pattern changes generated by different speakers (i.e., humans or loudspeakers) when speaking for liveness detection, which are captured by the built-in magnetometer on smartphones. To extract effective and robust magnetic features, MagLive utilizes a TF-CNN-SAF model as the feature extractor, which includes a time-frequency convolutional neural network (TF-CNN) combined with a self-attention-based fusion (SAF) model. Supervised contrastive learning is then employed to achieve user-irrelevance, device-irrelevance, and content-irrelevance. MagLive imposes no additional burden on users and does not rely on active sensing or specialized hardware. We conducted comprehensive experiments with various settings to evaluate the security and robustness of MagLive. Our results demonstrate that MagLive effectively distinguishes between humans and attackers (i.e., loudspeakers), achieving an average balanced accuracy (BAC) of 99.01% and an equal error rate (EER) of 0.77%.**

***Index Terms*—Liveness detection, voice authentication, smartphone, magnetic sensing, user security and privacy.**

## I. Introduction

VOICE authentication technologies, as a well-known form of biometrics, have seen a marked increase in adoption for executing sensitive operations on modern smartphones [1], [2]. These operations encompass a range of activities, from secure logins to mobile banking transactions. Notable implementations include WeChat's Voiceprint feature, which facilitates user login through voice passwords [3], and Citi's deployment of voice biometrics for customer identification [4]. Despite the convenience and user-friendly nature of these systems, the inherent public exposure of human voices introduces significant vulnerabilities. Given the easily accessible nature of voice data, these systems are particularly prone to spoofing attacks wherein attackers could record and manipulate voice samples, then replay them with loudspeakers to unauthorizedly access secure services. Such vulnerabilities not only compromise personal privacy but also pose substantial risks of financial loss and unauthorized access to sensitive information.

Efforts to counteract spoofing attacks have led to advancements in liveness detection technology, designed to determine whether a voice is genuinely human or artificially reproduced and replayed by a loudspeaker. Research in this area spans various devices like smartphones, smart speakers, and wearables. Techniques include using microphone arrays [5], [6], [7], [8] and other specialized hardware (wireless, mmWave radars) [9], [10], [11], [12] for smart speakers, and methods tailored for wearable devices [13], [14], [15].

Voice liveness detection on smartphones can be categorized into three key areas, each exploiting distinct features of voice interaction. Firstly, sound field features analyze the acoustic energy created as the sound propagates over the air [16], but require the user to remain stationary gesture for accurate results. Secondly, human features target the biological mechanisms of voice production [17], [18], [19], [20], [21], [22], yet may introduce discomfort with the need for high-frequency sound active sensing or reliance on specialized hardware. Thirdly, loudspeaker features focus on detecting anomalies in speaker output [23], [24], [25]. However, some struggle against advanced attacks [26] that closely imitate human voices [23], [24], while others are unsuitable for diverse environments and demand specific user actions, compromising practicality and convenience [25].

### A. MagLive

In this work, we introduce MagLive, a robust approach for detecting the liveness of voices on smartphones by utilizing changes in magnetic patterns, overcoming the shortcomings of existing methods that lack user-friendliness and security. The key idea is to discern the distinctive variations in magnetic pattern generated by human speech as opposed to those produced by loudspeakers, as depicted in Fig. 1. This technique

Received 16 August 2024; revised 4 June 2025 and 2 February 2026; accepted 6 March 2026. Date of publication 11 March 2026; date of current version 20 March 2026. This work was supported in part by the National Natural Science Foundation of China under Grant 62302343 and Grant 62472323, in part by the Key Research and Development Program of Hubei Province under Grant 2024BAB018, and in part by Wuhan Scientific and Technical Achievements Project under Grant 2024030803010172. The associate editor coordinating the review of this article and approving it for publication was Dr. Fernando Alonso-Fernandez. *(Corresponding author: Ruiying Du.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Institutional Review Board of Wuhan University.

Xiping Sun, Jing Chen, Cong Wu, Kun He, Haozhe Xu, and Ruiying Du are with the Key Laboratory of Aerospace Information Security and Trusted Computing, Ministry of Education, School of Cyber Science and Engineering, Wuhan University, Wuhan 430072, China (e-mail: xiping@whu.edu.cn; chenjing@whu.edu.cn; cnacwu@whu.edu.cn; hekun@whu.edu.cn; haozhexu@whu.edu.cn; duraying@whu.edu.cn).

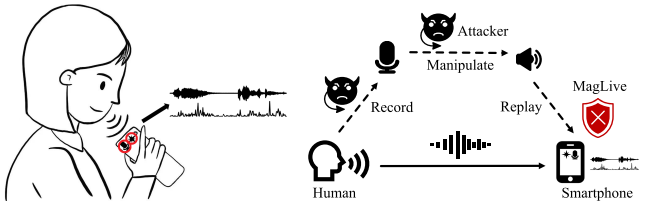
Yebo Feng is with the School of Computer Science and Engineering, Nanyang Technological University, Singapore 639798 (e-mail: yebo.feng@ntu.edu.sg).

Xianhao Chen is with the Department of Electrical and Electronic Engineering and the HKU Musketeers Foundation Institute of Data Science, The University of Hong Kong, Hong Kong, SAR, China (e-mail: xchen@eee.hku.hk).

Digital Object Identifier 10.1109/TIFS.2026.3673129

1556-6021 © 2026 IEEE. All rights reserved, including rights for text and data mining, and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: Nanyang Technological University Library. Downloaded on April 02,2026 at 16:02:23 UTC from IEEE Xplore. Restrictions apply.




(a) Scenario (b) Liveness detection

Fig. 1. Illustration of MagLive. (a) It uses the built-in magnetometer and microphone on the smartphone for voice liveness detection. (b) It detects the liveness of the voice to determine whether it is from an authentic human or artificially reproduced.

utilizes the inherent magnetometer and microphone already equipped in smartphones, enabling voice authenticity detection without the need for active signal sensing or specialized hardware. MagLive features minimal operational constraints and maintains its effectiveness across diverse environmental conditions, setting a new standard in voice liveness detection technology.

The development of MagLive presents three primary challenges. The first challenge is that the magnetic field changes induced by speakers are minute and weak, causing useful signals to be potentially submerged by noise in the magnetometer. We tackle this by first filtering out noise and neutralizing the effects of the Earth's magnetic field. Then, we leverage the voice data to aid in the segmentation of the magnetometer data. The second challenge is the intricate nature of magnetic field variations caused by speakers, which complicates the extraction of relevant magnetic patterns for voice detection. To overcome this, we deploy a TF-CNN-SAF model to achieve precise feature representation, combining a time-frequency convolutional neural network (TF-CNN) with a self-attention-based fusion (SAF) model. The third challenge stems from the susceptibility of magnetometer data to diverse external factors, including user interactions, device differences, voice variations, and environmental noise. We employ a supervised contrastive learning approach to maximize the differences between samples from humans and loudspeakers, while minimizing the differences between samples within the same class.

### B. Novelty

Significantly different from previous work, especially that of Chen et al. [25], the novelty of our paper lies in the following aspects: 1) Learning-based representation: Unlike prior approaches that rely on handcrafted signal-level features combined with heuristic thresholds, MagLive learns discriminative features automatically using a TF-CNN with self-attention fusion. 2) Suitability for diverse conditions: MagLive is designed to operate without environment-specific threshold calibration, which is a key limitation of prior methods. By leveraging supervised contrastive learning, MagLive reduces sensitivity to environmental variations, making it more suitable for deployment in real-world conditions. 3) Little usage constraints: MagLive does not require active sensing, predefined user motion trajectories, or specialized hardware. This passive design reduces user burden and facilitates practical deployment on commodity smartphones.

Our contributions can be summarized as follows:

- We propose MagLive, a robust voice liveness detection scheme that models magnetic pattern changes associated with speech to defend against spoofing attacks on smartphones.
- We devise a series of preprocessing methods to isolate the segments of magnetometer data induced by speech, and design 1) a TF-CNN-SAF-based feature extraction method to derive effective and robust feature representations, and 2) a supervised contrastive learning-based training method to achieve user-irrelevance, device-irrelevance, and content-irrelevance.
- We conduct comprehensive experiments on real devices and various settings to evaluate the performance of MagLive, achieving an average balanced accuracy (BAC) of 99.01% and an equal error rate (EER) of 0.77%.

### C. Code Availability

We aim to promote open, reproducible, and transparent research among the academic research community. Therefore, we publish MagLive's processing pipeline and the neural network codebase, which can be obtained at GitHub repository: https://github.com/sxp-up/MagLive.

## II. Related Work

Liveness detection enhances the security of voice authentication against spoofing attacks. In this section, we categorize existing voice liveness detection methods into two classes based on the authentication device type: smartphones and other devices (e.g., smart speakers and wearable devices). Table I summarizes the characteristics of some state-of-the-art voice liveness detection methods.

### A. Voice Liveness Detection on Smartphones

For voice liveness detection methods on smartphones, we further divide them into three classes based on the source of distinctiveness: sound field features [16], human features [17], [18], [19], [20], [21], [22], [27], [28], [29] and loudspeaker features [23], [24], [25].

*1) Sound Field Features:* These methods use the unique biometric information embedded in the sound field during the sound propagation stage. CaField [16] extracts the fieldprint from the sound field to detect speakers (either humans or loudspeakers). However, users are required to maintain a fixed manner to ensure the robustness of these fieldprints.

*2) Human Features:* These methods focus on unique features during human voice generation for liveness detection. Some methods use smartphones to actively transmit high-frequency acoustic signals to sense user's articulatory gestures [17], lip motions [18], [19] and chest motions [20] during human voice generation. VoiceLive [22] measures the time-difference-of-arrival (TDoA) changes in a sequence of phoneme sounds, which is sensitive to the placement of the smartphone. Some methods leverage the unique energy responses [27], breathing pop sounds [28], and the arrangement of the human vocal tract estimated by fluid dynamics [29] during speech generation. Wang et al. [21] detect the pressure of the oral airflow using specialized hardware.

Authorized licensed use limited to: Nanyang Technological University Library. Downloaded on April 02,2026 at 16:02:23 UTC from IEEE Xplore. Restrictions apply.

TABLE I
COMPARISON OF STATE-OF-THE-ART VOICE LIVENESS DETECTION METHODS

| Device | System | Distinctiveness | No active sensing | No specialized hardware | Little usage constraint | Resist spectrum modulated attacks | Diverse conditions[1] | Accuracy | EER |
|---|---|---|---|---|---|---|---|---|---|
| Smartphones | CaField [16] | Sound field | ✓ | ✓ | × | ✓ | ✓ | 99.16% | 0.85% |
| | VoiceGesture [17] | Mouth motion | × | ✓ | ✓ | ✓ | ✓ | ∼ 99% | ∼ 1% |
| | VoiceLive [22] | Phoneme location | ✓ | ✓ | × | ✓ | ✓ | ∼ 99% | ∼ 1% |
| | Wang *et al.* [21] | Oral airflow | ✓ | × | ✓ | ✓ | ✓ | 97.25% | 2.08% |
| | Void [24] | Hardware imperfections | ✓ | ✓ | ✓ | × | ✓ | ∼ 98% | ∼ 1% |
| | Chen *et al.* [25] | Magnetic field | × | ✓ | × | ✓ | × | 100% | 0% |
| | MagLive (our work) | Magnetic pattern changes | ✓ | ✓ | ✓ | ✓ | ✓ | 99.01% | 0.77% |
| Other devices | ArrayID [6] | Multi-channel audio | ✓ | × | ✓ | ✓ | ✓ | 99.84% | 0.17% |
| | WearID [13] | Device ownership | ✓ | × | ✓ | ✓ | ✓ | 97.2% | N/A |

[1] : Diverse conditions means that the methods is robust in various situations, such as different environments.

*3) Loudspeaker Features:* Researchers also detect voice liveness by concentrating on features that arise when a loudspeaker generates sound. Blue et al. [23] and Void [24] use distortions in the sound caused by the hardware imperfections of loudspeakers as features. These methods only rely on features extracted in the audio domain, making them vulnerable to acoustic attacks, such as spectrum modulated attacks [26]. As loudspeakers inevitably generate magnetic fields during sound emission, Chen et al. [25] exploit magnetometer readings for voice liveness detection. Their method relies on handcrafted signal-level features, namely magnetic strength and its temporal changing rate, combined with manually determined thresholds. To achieve high accuracy, the system requires active acoustic probing and explicit user actions to move the smartphone along a predefined trajectory. Moreover, as discussed in their work, the decision thresholds need to be calibrated for specific environments and are sensitive to magnetic interference. Such environment-dependent calibration limits the applicability of the method in real-world scenarios.

In contrast, our work is to explore the effective changes in magnetic patterns associated with speech. MagLive eliminates the need for active sensing and specialized hardware, imposing little usage constraint. It is robust against various attacks and effective across diverse environmental conditions.

### B. Voice Liveness Detection on Other Devices

Existing methods utilize microphone arrays [5], [6], [7], [8] and specialized hardware [9], [10], [11], [12] for voice liveness detection on smart speakers. Additionally, some methods require extra wearable devices to assist authentication [13], [14], [15].

Li et al. [5], ArrayID [6] and VoShield [7] utilize multi-channel audio from the microphone array to detect the difference between human speech and spoofing audio. Speaker-Sonar [8] leverages a circular microphone array to emit the inaudible sound and track the user's direction. All these methods require microphone arrays for detection. Some methods capture wireless signals to recognize mouth motions [9], [10] or unique breathing rates [11] to distinguish authentic voice commands from spoofed ones. Vocalprint [12] leverages skin-reflect mmWave signals to sense the minute vocal vibrations in the near-throat region of users. These methods require additional sensors such as wireless sensors and mmWave radars. WearID [13] leverages motion sensors on the user's wearable device to capture aerial speech in the vibration domain and verifies it with the speech captured in the audio domain. VAuth [14] is designed to fit in widely-adopted wearable devices, such as eyeglasses and earphones. 2MA [15] uses multiple devices operating in the same area to eliminate replay attacks. These methods require additional smart devices to assist with authentication, leading to extra costs.

## III. PRELIMINARIES

In this section, we first introduce the magnetic effect of speakers, and then show the motivating examples.

### A. Magnetic Effect of Speakers

The mechanisms behind sound production in humans and loudspeakers are distinct. Human speech is created through the coordination of vocal cords and various other organs, working in unison to produce sound [29]. On the other hand, an electric loudspeaker operates on electromagnetic induction [30], converting electrical currents into sound. This process results in a specific magnetic signature unique to loudspeakers.

Fig. 7 illustrates the basic components of an electric loudspeaker: a permanent magnet, a voice coil, and a diaphragm. The permanent magnet creates a steady magnetic field. When the loudspeaker is in use, electrical current passes through the voice coil, turning it into a temporary electromagnet that generates changing magnetic fields. This interaction causes the voice coil to move towards or away from the permanent magnet. The diaphragm, attached to the voice coil, vibrates in response to these movements, pushing against the air to create sound waves. These sound waves result from the shifts in the magnetic field around the coil, meaning that sound production in a loudspeaker is directly linked to variations in magnetic patterns.

### B. Motivating Examples

To investigate the variations in magnetic patterns associated with speech from different sources—namely, humans and loudspeakers—we carried out initial experiments. These involved recording both magnetic signals and voice data using the built-in magnetometer and microphone of a smartphone. For our experiments, we selected the iPhone 14 Pro as the authentication device and collected magnetometer data at its highest sampling rate of 100Hz.

Authorized licensed use limited to: Nanyang Technological University Library. Downloaded on April 02,2026 at 16:02:23 UTC from IEEE Xplore. Restrictions apply.

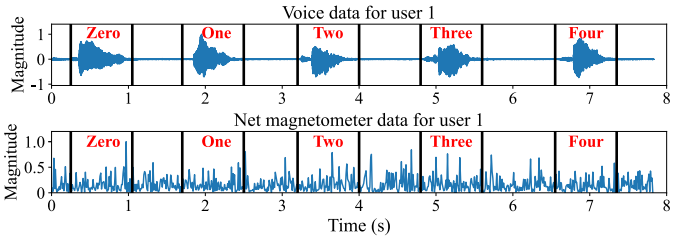



Fig. 2. An example of user 1 speaking digits from zero to four (human).

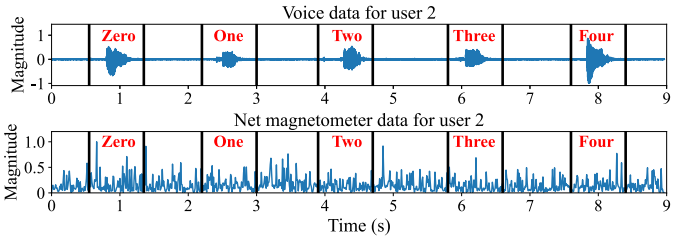



Fig. 3. An example of user 2 speaking digits from zero to four (human).

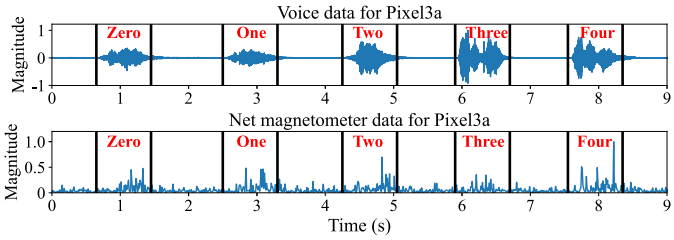



Fig. 4. An example of Pixel3a replaying the speech of user 1 (loudspeaker).

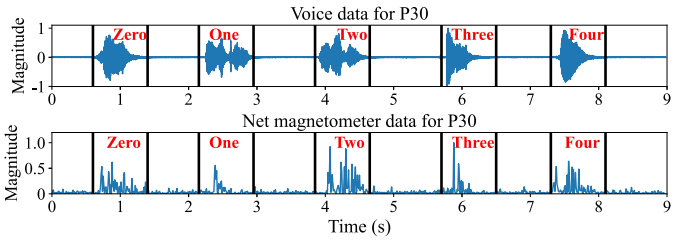



Fig. 5. An example of P30 replaying the speech of user 1 (loudspeaker).

To investigate the magnetic signatures associated with human speech, we enlisted two volunteers to speak a sequence of numbers from zero to four in English. The resulting voice and magnetometer data, normalized for comparison, are illustrated in Fig. 2 for the first user and Fig. 3 for the second. To assess the magnetic effects generated by loudspeakers, we played back the recording of the first user's speech through two different speaker models, Pixel3a and P30, without any direct contact with the testing smartphone. The outcomes of these tests are depicted in Fig. 4 and Fig. 5, showcasing the voice and magnetometer data during the simulated spoofing attacks. Distinct patterns can be observed from the magnetometer data triggered by loudspeakers compared to the natural human voice. These results are consistent with the analyses in Section III-A, which indicate that dynamic magnetic fields caused by loudspeakers when emitting sound create specific magnetic signature. Notably, the magnetic field changes varied not just between humans and loudspeakers but also among different individuals and spoofing devices, despite reproducing the same set of spoken digits. These variations underscore the unique magnetic footprints left by different sources and form the empirical basis for MagLive's design. Leveraging these distinct magnetic patterns, we refine our voice liveness detection methodology in MagLive.

## IV. Overview of MagLive

In this section, we first present the system overview of MagLive. Then, we introduce the threat model and design goals.

### A. System Overview

The basic idea of MagLive is that different speakers (i.e., humans or loudspeakers) generate unique magnetic pattern changes when speaking. Therefore, MagLive utilizes the commercial smartphone's built-in microphone and magnetometer for magnetic sensing-based voice liveness detection. Fig. 6 shows the workflow of MagLive, which comprises four modules: data capture, data preprocessing, feature extraction, and authentication.

In our MagLive framework, the data capture module concurrently gathers voice and magnetometer data from the authentication device, typically a smartphone, with subsequent analysis to determine the distance between the sound source and the device. During data preprocessing, we refine the magnetometer data by eliminating noise and reducing the Earth's magnetic field's impact, using voice data to segment the magnetometer readings effectively. The feature extraction module then isolates patterns of magnetic field fluctuations attributed to different sources, employing a TF-CNN-SAF model that combines a time-frequency convolutional neural network (TF-CNN) with a self-attention-based fusion (SAF) model for effective and robust feature derivation. This process is enhanced through supervised contrastive learning, tailored to ensure the system's independence from user identity, voice content, and device variation. Finally, the authentication module processes these refined features through a binary classifier to ascertain the voice sample's authenticity and differentiate between human and non-human sources efficiently.

### B. Threat Model

The goal of an attacker is to bypass the voice authentication system and perform sensitive operations. State-of-the-art voice authentication systems are robust against human-based voice impersonation attacks, where an attacker tries to mimic a target user's voice timbre and prosody without machines [25]. Therefore, such attacks do not pose a real threat [31]. In this paper, we focus on a more realistic spoofing attack scenario, where attackers record and manipulate voice samples, then replay them using spoofing devices (primarily loudspeakers). Thus, considering the attacker's ability and goal, spoofing attacks can be classified into three types [32]:

- *Replay.* The attacker uses loudspeakers to replay recorded voice samples collected from the target user [11].

Authorized licensed use limited to: Nanyang Technological University Library. Downloaded on April 02,2026 at 16:02:23 UTC from IEEE Xplore. Restrictions apply.

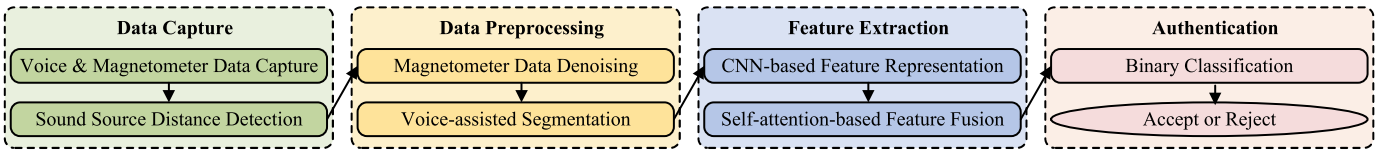



Fig. 6. Workflow of MagLive.

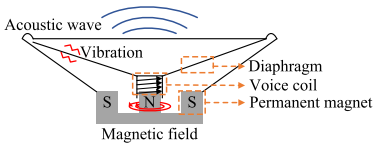



Fig. 7. The mechanical structure of an electric loudspeaker.

- *Speech synthesis.* The attacker generates intelligible artificial voice that sounds like the target user from text and plays it via loudspeakers [33].
- *Voice conversion.* The attacker manipulates the voice of a human speaker so that it resembles the target user and plays it with loudspeakers [34].

Besides spoofing attacks, we also consider advanced attacks such as modulated attacks [26] and adversarial attacks [35]. While the techniques of the above attacks differ, they rely on spoofing devices (primarily loudspeakers) to replay the manipulated voice samples, as shown in Fig. 1(b). Therefore, we focus on detecting the distinct characteristics of loudspeakers to cope with these attacks.

### C. Design Goals

A suitable liveness detection method for a voice authentication system should satisfy the following goals:

- *Security:* it should effectively distinguish between legitimate samples and various spoofing samples, ensuring that only authentic users are granted access.
- *Robustness:* it should perform robustly across diverse conditions, such as different spoofing devices and authentication devices, various users, different voice contents, and different environmental settings.
- *User-friendliness:* it should be easy to use for users, requiring minimal user intervention and not relying on active sensing or specialized hardware.

## V. Design of MagLive

MagLive is composed of four modules: data capture, data preprocessing, feature extraction, and authentication. In this section, we provide a detailed explanation of each module.

### A. Data Capture

MagLive simultaneously collects voice and magnetometer data from the authentication device (i.e., smartphone). The voice data is used for voice authentication, while the magnetometer data supports liveness detection. Given the intrinsic attenuation of magnetic signals [36], we ensure system performance by verifying that the sound source is within a short-range distance threshold.

*1) Sound Source Distance Detection:* To minimize user burden and avoid active signal sensing, we estimate the sound source distance by leveraging the distance and energy differences between the smartphone's two microphones [37]. Unlike prior work [25] requiring high-frequency acoustic signals and user motion, our method is passive and lightweight.

We apply the generalized cross correlation with phase transform (GCC-PHAT) [38] to estimate the time difference of arrival (TDOA) between two microphones. Assuming sound velocity $v \approx 340$ m/s [39], we estimate the distance difference. If we denote the distances from the sound source to the two microphones as $d_1$ and $d_2$ respectively, the distance difference can be expressed as $\triangle d = |d_1 - d_2|$. Let $E_1$ and $E_2$ denote the received energies at the two microphones, then from the inverse-square law [40]:

$$E_1 d_1^2 = E_2 d_2^2 \tag{1}$$

Assuming $d_1 > d_2$ (i.e., mic 1 is farther), we solve for $d_1$ and $d_2$ as:

$$d_1 = \frac{\sqrt{E_2}}{\sqrt{E_2} - \sqrt{E_1}} \triangle d \quad \text{and} \quad d_2 = \frac{\sqrt{E_1}}{\sqrt{E_2} - \sqrt{E_1}} \triangle d \tag{2}$$

To ensure reliability, we first check whether $\triangle d$ approximately equals the microphone spacing (i.e., the smartphone length), which implies that the sound source and microphones lie on a straight line. Only under this condition do we proceed to evaluate whether $d_2$ (the closer distance) is within the predefined threshold (6 cm), which is determined through experiments in Section VI-C. If $d_2$ exceeds the threshold, we discard the sample and prompt re-collection. This mechanism enables MagLive to verify whether the sound source is within range. We validate the effectiveness of this distance detection method through experiments in Section VI-B.

### B. Data Preprocessing

After capturing the voice and magnetometer data, we perform denoising and use the voice data to assist in segmenting the magnetometer data.

*1) Denoising:* Magnetic field changes induced by speakers are weak and can be easily submerged by noise in the magnetometer [41]. To address this, we first apply a high-pass Butterworth filter [42] to eliminate the noise. We set the filter's cut-off frequency to 5 Hz, based on empirical observations and our insights. Taking the voice command "OK Google Hey Siri" as an example, Fig. 8 demonstrates the denoising effect on the z-axis magnetometer data, while Fig. 9 shows the corresponding effect on the spectrogram. We can observe that the denoised magnetometer data exhibits more distinct magnetic patterns related to the voice data. Further evaluation is shown in Section VI-B.

Authorized licensed use limited to: Nanyang Technological University Library. Downloaded on April 02,2026 at 16:02:23 UTC from IEEE Xplore. Restrictions apply.

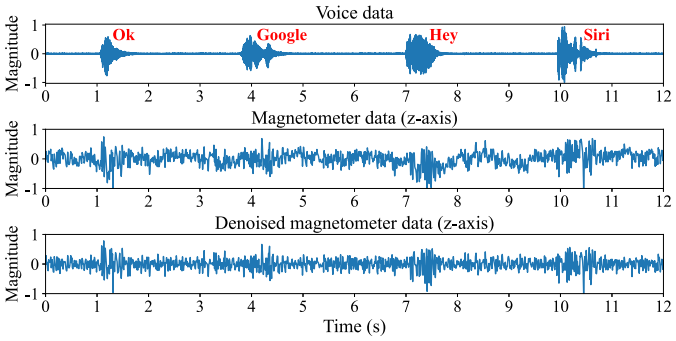
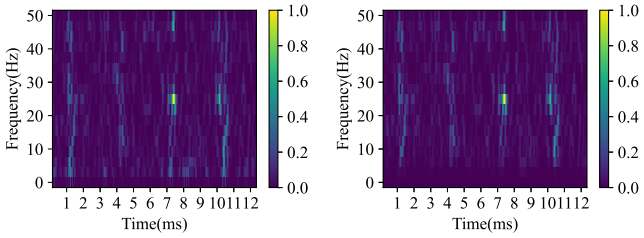



Fig. 8. An example of the denoising effect of the z-axis magnetometer data corresponding to the voice command "OK Google Hey Siri".




(a) Spectrogram of original data (b) Spectrogram of denoised data

Fig. 9. An example of the denoising effect applied to the spectrogram of the z-axis magnetometer data.

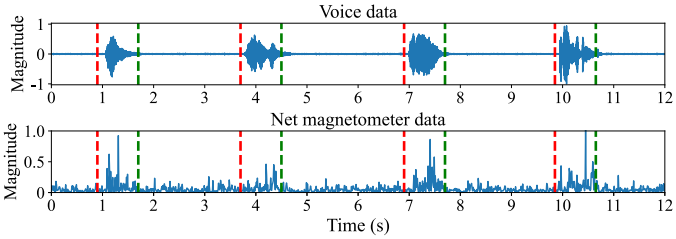



Fig. 10. An example of the segmentation result of the net magnetometer data for the voice command "OK Google Hey Siri", assisted by voice data.

Considering the influence of the Earth's magnetic field, the captured three-axis magnetometer data is geo-spatial dependent. To mitigate the impact of location, we aggregate the magnetometer data across the three axes. The net magnetometer data $\boldsymbol{m'}$ derived from the three-axis magnetometer data $\boldsymbol{m} = (\boldsymbol{m_x}, \boldsymbol{m_y}, \boldsymbol{m_z})$ is calculated as follows [43]:

$$m'(t) = \|\boldsymbol{m}(t)\| = \sqrt{m_x(t)^2 + m_y(t)^2 + m_z(t)^2} \tag{3}$$

*2) Voice-Assisted Segmentation:* To accurately pinpoint the magnetic changes induced by speech, we use the voice data to assist in segmenting the magnetometer data. First, we apply voice activity detection [16] to identify segments of speech in the recorded voice data, with each segment indicating the presence of a speech signal. Since the magnetometer and voice data are synchronized, we then use the voice segments to determine the corresponding segments in the magnetometer data. To ensure that each segment of magnetometer data covers an entire speech segment, we adjust the start and end points of each segment to span 100 sample points. This length is an empirical parameter based on our observations. This process

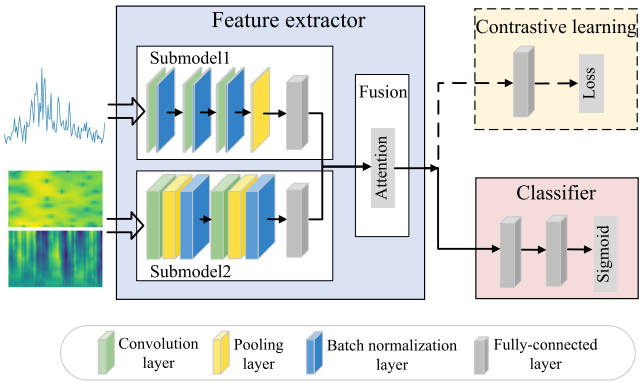



Fig. 11. The architecture of our feature extraction and classifier model.

ultimately yields magnetometer data segments corresponding to each word for feature extraction. Fig. 10 illustrates an example of the segmentation result of the net magnetometer data, demonstrating the effectiveness of our method. Further evaluation is shown in Section VI-B.

### C. Feature Extraction

To distinguish between human speech and loudspeaker playback, we design a tailored feature extractor named TF-CNN-SAF, which captures magnetic pattern variations induced by different types of speakers. Specifically, we first construct a time-frequency convolutional neural network (TF-CNN) to extract features from two complementary perspectives: time-domain envelope patterns and time–frequency spectrograms. To adaptively integrate these multi-view features, we introduce a self-attention-based fusion (SAF) model that learns dynamic weights based on channel importance. Finally, to promote robustness across different users, devices, and speech content, we employ a supervised contrastive learning strategy that enhances the discriminative power of the extracted features. Below, we provide a detailed description of each component in our model design.

*1) TF-CNN-Based Feature Representation:* To extract meaningful and effective features from magnetometer data, we construct a time-frequency convolutional neural network (TF-CNN) with a dual-branch design, which is commonly used in multi-representation modeling [44], [45], [46]. Specifically, as shown in Fig. 11, we first extract the envelope of the pre-processed magnetometer signal to capture its temporal variation, and feed it into a 1D-CNN branch (submodel1). In parallel, we apply the Short-Time Fourier Transform (STFT) [47] to generate spectrograms that retain both magnitude and phase information. These are then fed into a 2D-CNN branch (submodel2). This dual-branch structure enables the model to learn complementary magnetic patterns from two perspectives: the 1D-CNN captures the timing dynamics of magnetic fluctuations, while the 2D-CNN focuses on spectral and spatial characteristics.

Table II summarizes the architecture of our TF-CNN model. For submodel1, we first adopt three convolution blocks to learn the feature embedding. Each convolution block comprises a 1-D convolution (Conv1D) layer, followed by a batch normalization (BN) layer to accelerate the training process,

Authorized licensed use limited to: Nanyang Technological University Library. Downloaded on April 02,2026 at 16:02:23 UTC from IEEE Xplore. Restrictions apply.

TABLE II
THE STRUCTURE OF OUR TF-CNN MODEL

| Model name | Layer | Layer type | Output shape | # Param |
|---|---|---|---|---|
| Submodel1 | 1 | Conv1D + BN + ReLU | (98,16) | 128 |
| | 2 | Conv1D + BN + ReLU | (96,32) | 1,696 |
| | 3 | Conv1D + BN + ReLU | (94,16) | 1,616 |
| | 4 | Pooling | (47,16) | 0 |
| | 5 | Flatten + FC + ReLU | (64) | 48,192 |
| Submodel2 | 1 | Conv2D + ReLU | (15,67,16) | 304 |
| | 2 | Pooling + BN | (7,33,16) | 64 |
| | 3 | Conv2D + ReLU | (5,31,32) | 4,640 |
| | 4 | Pooling + BN | (2,15,32) | 128 |
| | 5 | Flatten + FC + ReLU | (64) | 61,504 |

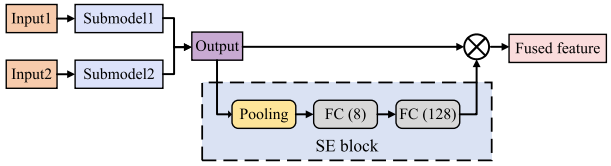



Fig. 12. The architecture of our self-attention-based fusion (SAF) model.

and a ReLU layer as the activation function. Subsequently, an average pooling layer is added. For submodel2, we use two 2-D convolution (Conv2D) layers with interleaved max pooling and BN layers. Specifically, the kernel sizes for the Conv2D and max pooling layers are set to $3 \times 3$ and $2 \times 2$, respectively. The outputs of both branches are flattened and compressed through fully connected layers into a unified 128-dimensional feature vector.

*2) Self-Attention-Based Feature Fusion:* Instead of simply concatenating the features extracted from the two submodels, we develop a self-attention-based fusion (SAF) model, which recalibrates the magnetic features with adaptive weights to selectively emphasize the significant ones. As shown in Fig. 12, we employ a Squeeze-and-Excitation (SE) block [48] with a global average pooling layer and two fully-connected (FC) layers to learn a weight vector in range of [0, 1]. Specifically, the global average pooling layer computes the average value for each channel, and then two fully-connected (FC) layers produce a weight vector. The obtained weight vector is used to scale the outputs of the two submodels, creating fused features that assign higher weights to the more informative ones. Finally, the feature extractor of MagLive produces a 128-dimensional vector as the feature representation.

*3) Contrastive Learning-Based Model Training:* In this paper, our focus lies in detecting the differences in magnetic pattern changes caused by humans and loudspeakers for voice liveness detection. To construct a user-irrelevant, device-irrelevant, and content-irrelevant liveness detection method, we use supervised contrastive learning [49], [50] to train the feature extractor. This method eliminates user specificity, device specificity, and content specificity by maximizing the differences between samples from humans and loudspeakers, while minimizing the differences between samples within the same class.

The structure of the contrastive learning network mainly consists of three components: the feature extractor, the projection head, and the supervised contrastive loss. The feature extractor is built upon the TF-CNN-SAF model. To

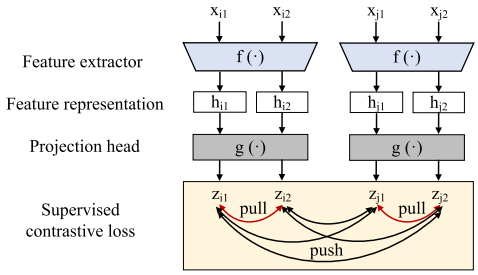



Fig. 13. Illustration of contrastive learning-based model training.

compute supervised contrastive loss during training, we append a projection head for the feature extractor, which includes a fully-connected layer and a ReLU layer. Specifically, for each batch $I$ containing samples and corresponding labels $\{x_i, y_i\}$, where $i \in I \equiv \{1, 2, \ldots, N\}$, the input $x_i$ is first passed through the feature extractor $f(\cdot)$ to obtain the feature representation $h_i$. Subsequently, the feature representation is further propagated through a projection head $g(\cdot)$, yielding the output $z_i$. The supervised contrastive loss is then computed on $z_i$ using Eq. 4, aiming to minimize the distances between feature representations of samples from the same class while maximizing the distances between those from different classes:

$$\mathcal{L} = \sum_{i \in I} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(z_i \cdot z_p / \tau)}{\sum_{a \in A(i)} \exp(z_i \cdot z_a / \tau)} \tag{4}$$

Here, $A(i) \equiv I \setminus \{i\}$, $P(i) \equiv \{p \in A(i) : y_p = y_i\}$, and $|P(i)|$ is its cardinality. The symbol $\cdot$ denotes the inner product, and $\tau$ is a scalar temperature parameter. In the case where $x_{i1}$ and $x_{i2}$ are from the same class, and $x_{j1}$ and $x_{j2}$ are from another class, as illustrated in Fig. 13, the aim is to pull together samples from the same class while pushing apart samples from different classes.

During training, hyperparameters are selected via a grid search on the validation set to identify an effective configuration. The model is optimized using the Adam optimizer under the supervised contrastive learning framework. The TF-CNN-SAF feature extractor is trained in a single stage. After training, the projection head and the supervised contrastive loss are discarded, and only the trained TF-CNN-SAF feature extractor is retained for inference.

To test the effectiveness of our feature representations, we randomly select 200 testing samples from a loudspeaker and a human. These samples are then fed into the trained feature extractor to extract their corresponding feature representations. Then we utilize t-Distributed Stochastic Neighbor Embedding (t-SNE) [51] to reduce the dimension of the feature representation from 128 to 2 and visualize these samples. As shown in Fig. 14, we can see that samples with the same label are closely clustered in the feature space, while samples from different classes are farther apart. The distinct feature distribution of the loudspeaker and human samples demonstrate the feasibility of our feature extraction approach.

### D. Authentication

Based on the feature representations generated by the feature extractor, MagLive uses binary cross-entropy as the loss

Authorized licensed use limited to: Nanyang Technological University Library. Downloaded on April 02,2026 at 16:02:23 UTC from IEEE Xplore. Restrictions apply.

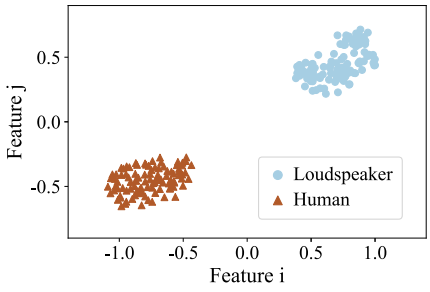



Fig. 14. T-SNE visualization of features for the loudspeaker and human.

function to train a classifier for voice liveness detection, distinguishing between human (1) and loudspeaker (0) samples. As shown in Fig. 11, the classifier consists of two fully-connected layers followed by a sigmoid layer. Since voice liveness detection is a binary classification task, the output of the sigmoid layer represents the probability that the voice sample belongs to a human. If the probability exceeds a predefined threshold, the sample is classified as originating from a real human. Otherwise, it is deemed to be from an attacker (i.e., a loudspeaker). In this work, we empirically set the threshold at 0.5.

## VI. Evaluation

In this section, we report the evaluation results of our proposed voice liveness detection system, MagLive. We first introduce the experiment setup, including the data collection and evaluation metrics. Then we present the overall performance of MagLive and evaluate the security of MagLive in defending against various attacks. We also conduct comprehensive experiments to evaluate the robustness and effectiveness of MagLive under different settings and factors.

### A. Experiment Setup

*1) Data Collection:* Our data collection procedure is designed to capture both bona fide and spoofing speech under a smartphone-based acquisition protocol, and consists of two phases: human data collection and spoofing data collection. We use an iPhone 14 Pro as the primary authentication device and utilize the Sensor Logger app [52] to gather sensor measurements from the smartphone's built-in sensors. The built-in magnetometer samples at a rate of 100 Hz, while audio is recorded at 44.1 kHz.

For human data collection, we recruited 20 participants in this study, aged from 21 to 28, including 9 males and 11 females. Participants were informed that the purpose of the experiments was to enhance the security of voice authentication on smartphones. The human dataset consists of two parts. Following the WeChat voiceprint authentication scheme [3], where users read a set of digits as passwords, each participant was first requested to speak ten digits from zero to nine for 20 repetitions. In addition, we selected 15 commonly used voice commands, as listed in Table III, and each participant repeated each command twice. In total, 50 voice commands and the corresponding magnetometer data were collected from each participant, resulting in 1000 voice commands (equivalent to 6000 words) and synchronized magnetometer measurements in the human dataset. The human data were collected in an office environment with natural background noises, including human conversations and heating, ventilation, and air conditioning (HVAC) noise, to reflect realistic smartphone usage conditions.

TABLE III
List of 15 Voice Commands Used in the Experiments

| No. | Voice commands |
|---|---|
| 1 | Alexa. |
| 2 | Cortana. |
| 3 | OK Google. |
| 4 | Hey Siri. |
| 5 | Hi Assistant. |
| 6 | Turn on Bluetooth. |
| 7 | Take a photo. |
| 8 | Open music player. |
| 9 | Mute the volume. |
| 10 | Show me my messages. |
| 11 | Where is my package? |
| 12 | Call the nearest computer shop. |
| 13 | Remind me to buy milk. |
| 14 | What is my schedule for tomorrow? |
| 15 | What is the time at home? |

TABLE IV
Experimental Devices and Their Basic Information

| No. | Type | Manuf. | Model | Size(L*W*H in cm) |
|---|---|---|---|---|
| 1 | Smartphone | Apple | iPhone 14 Pro | 14.8 × 7.2 × 0.8 |
| 2 | Smartphone | Apple | iPhone XR | 15.1 × 7.6 × 0.8 |
| 3 | Smartphone | Huawei | P30 | 14.9 × 7.1 × 0.8 |
| 4 | Smartphone | Google | Pixel 3a | 15.1 × 7.0 × 0.8 |
| 5 | Smartphone | Samsung | Galaxy S10 | 15.0 × 7.0 × 0.8 |
| 6 | Tablet | Apple | iPad Pro | 24.8 × 17.9 × 0.6 |
| 7 | Laptop | Lenovo | ThinkPad X1 | 32.4 × 21.7 × 1.6 |
| 8 | Loudspeaker | Xiaomi | AI Speaker | 8.8 × 8.8 × 21.2 |
| 9 | Loudspeaker | Amazon | Echo Dot | 9.9 × 9.9 × 4.3 |

For spoofing data collection, we replayed all collected human speech samples using different spoofing devices to simulate loudspeaker-based spoofing attacks. Specifically, we employed eight spoofing devices with diverse form factors, sizes, and acoustic characteristics, as summarized in Table IV, in order to introduce variations in replay-based acquisition conditions. Since the proposed method focuses on the physical characteristics induced by loudspeaker sound emission, we do not explicitly distinguish between replay, speech synthesis, and voice conversion attacks at the signal level, as they all involve loudspeaker-based sound generation during the acquisition stage. In total, we collected 8000 voice spoofing commands (equivalent to 48000 words) and the corresponding magnetometer data.

In addition to the primary dataset, we collected several additional small-scale datasets by intentionally modifying acquisition-related factors, such as recording environments, user postures, voice volumes, and device configurations. These auxiliary datasets correspond to distinct acquisition protocols and are used to evaluate the robustness and generalizability of MagLive under protocol variations. Details of these datasets and evaluations are provided in the corresponding sections.

*2) Privacy and Ethical Considerations:* All experiments are conducted under the approval of the institutional review board (IRB) of our university, with informed consent obtained from

Authorized licensed use limited to: Nanyang Technological University Library. Downloaded on April 02,2026 at 16:02:23 UTC from IEEE Xplore. Restrictions apply.

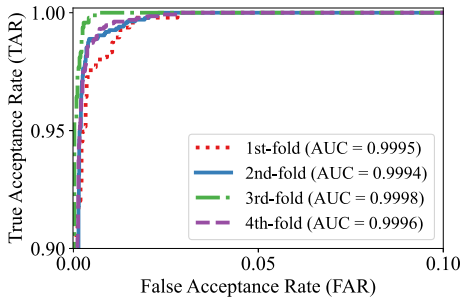



Fig. 15. ROC curves of the feature extractor under a 4-fold cross-validation.

TABLE V
BAC (%), FAR (%), FRR (%) AND EER (%) FOR THE FEATURE EXTRACTOR UNDER A 4-FOLD CROSS-VALIDATION

| Metrics | 1st-fold | 2nd-fold | 3rd-fold | 4th-fold | Average |
|---|---|---|---|---|---|
| BAC (%) | 98.41 | 99.05 | 99.55 | 99.02 | **99.01** |
| FAR (%) | 0.44 | 0.35 | 0.28 | 0.38 | **0.36** |
| FRR (%) | 2.73 | 1.55 | 0.63 | 1.58 | **1.62** |
| EER (%) | 1.08 | 0.87 | 0.41 | 0.71 | **0.77** |

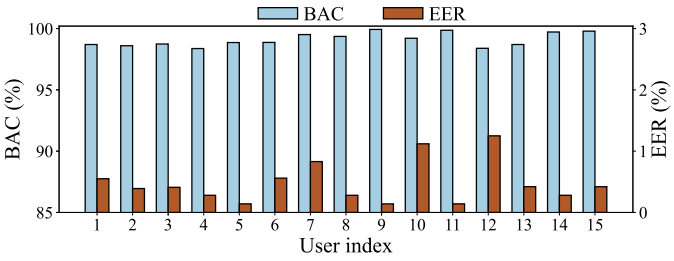



Fig. 16. The BAC and EER performance for each user.

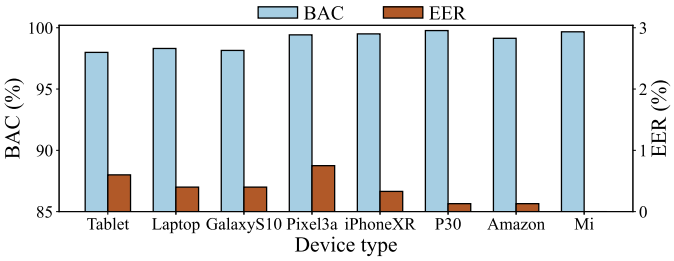



Fig. 17. The BAC and EER performance for different spoofing devices.

all participants. Audio data required for voice authentication and MagLive's additional magnetometer data are collected solely for research purposes and are not linked to participants' real-world identities. All collected data are anonymized at the subject level and stored in encrypted form on secure institutional servers. During processing, raw sensor data are used only for feature extraction and model training, and are not released or shared with third parties. All data handling strictly follows institutional data protection guidelines, which helps minimize potential privacy and ethical risks for participants.

*3) Evaluation Metrics:* We use the following metrics for evaluation. False Acceptance Rate (FAR) represents the rate at which attack samples are wrongly accepted and classified as bona fide samples, while False Rejection Rate (FRR) represents the rate at which bona fide samples are wrongly rejected. In the context of presentation attack detection (PAD), FAR and FRR correspond to the Attack Presentation Classification Error Rate (APCER) and the Bona-Fide Presentation Classification Error Rate (BPCER) defined in ISO/IEC 30107-3, respectively. Balanced accuracy (BAC) measures the overall probability that the system accepts bona fide samples and rejects attack samples [53]. Equal Error Rate (EER) [54] shows a balanced view of FAR and FRR, where FAR is equal to FRR. The Receiver Operating Characteristic (ROC) curve shows the relationship between the True Acceptance Rate and the FAR at various thresholds [55]. The Area Under the ROC Curve (AUC) is used to measure the probability that prediction scores of bona fide samples are higher than those of attack samples.

## B. Overall Performance

In this subsection, we evaluate the overall performance of our voice liveness detection system in detecting various attacks. Specifically, we assess how effectively MagLive can differentiate between authentic human speakers and attackers (i.e., spoofing devices).

*1) Performance of Feature Extractor:* We utilize a 4-fold cross-validation to evaluate the performance of the feature extractor. Specifically, the 20 participants are divided into four groups, each containing five users. In each fold, we use the data of 5 users to train the feature extractor and then test it with the remaining 15 users. This process is repeated four times. Fig. 15 shows the ROC curves of the feature extractor under the 4-fold cross-validation. The AUCs are 0.9995, 0.9994, 0.9998, and 0.9996, respectively. A higher AUC value indicates better system performance. Table V presents the BAC, FAR (equivalent to APCER defined in ISO/IEC 30107-3), FRR (equivalent to BPCER), and EER for the 4-fold performance. MagLive achieves an average BAC of 99.01% and an EER of 0.77%. We use data from the 5 users in the 3rd-fold to train the feature extractor for later evaluation. Despite being trained on limited data, the feature extractor demonstrates effectiveness to a wide range of users.

*2) Per-User Breakdown Analysis:* This analysis is conducted as a diagnostic study to examine inter-user variability, rather than as a practical deployment strategy. Under a controlled setting, we evaluate each user independently to analyze user performance variations. Specifically, a separate classifier is trained and evaluated for each of the 15 users using that user's bona fide and spoofing samples. Fig. 16 reports the BAC and EER for each user. The best case (user 9) achieves a BAC of 99.93% and an EER of 0.14%, while even the worst case (user 12) maintains a BAC above 98.3% with an EER below 1.3%. Despite the performance variations across users, the consistently high BAC and low EER values indicate that MagLive remains effective in distinguishing humans from loudspeaker-based attacks across a diverse user population.

*3) Performance for Different Voice Content:* We consider data from 15 participants and train the classifier separately for different types of voice content, including single digits, single words, and entire voice commands. Similar to Voshield [7], which utilizes short voice clips for voice liveness detection, we first report the performance when using single-digit and single-word utterances. We then evaluate the performance on

Authorized licensed use limited to: Nanyang Technological University Library. Downloaded on April 02,2026 at 16:02:23 UTC from IEEE Xplore. Restrictions apply.

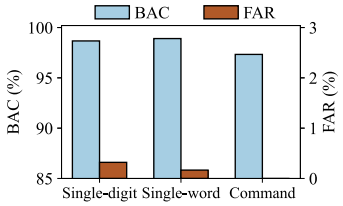



Fig. 18. The BAC and FAR performance for different voice content.

TABLE VI
ESTIMATED SOUND SOURCE DISTANCES (MEAN ± STD, CM) AT DIFFERENT ACTUAL DISTANCES (CM)

| Actual (cm) | 3 | 5 | 7 | 10 |
|---|---|---|---|---|
| Estimated (cm) | 3.02 ± 0.12 | 5.07 ± 0.30 | 6.98 ± 0.13 | 9.98 ± 0.56 |

entire voice commands [56], as shown in Table III. We assume that once any word within a voice command is identified as a spoofing sample, the entire command is regarded as spoofing. Fig. 18 shows the BACs and FARs for different voice content. For single-digit, single-word, and voice-command inputs, the BACs are 98.67%, 98.91%, and 97.33%, respectively, with FARs of 0.32%, 0.17%, and 0%. Although there is a slight decrease in BAC for entire voice commands, MagLive successfully detects all spoofing samples.

*4) Defending Against Advanced and Unseen Attacks:* In addition to conventional spoofing attacks, we further evaluate the robustness of MagLive against more advanced attack variants including modulated attacks [26] and adversarial attacks [35]. Modulated attacks [26] aim to alleviate loudspeaker-induced spectral distortion by modifying the voice spectrum, while adversarial attacks [35] preserve the semantic content and perceptual naturalness of human speech to evade detection. In this experiment, we use human speech from one participant as bona fide data and employ a Huawei P30 smartphone as the spoofing device to generate 200 spoofing samples for each attack type. MagLive achieves BACs of 99.5% and 100%, respectively, with FARs of 0%. These results indicate that the effectiveness of MagLive against advanced and unseen attacks stems from a fundamental physical constraint: loudspeaker actuation inevitably induces distinctive magnetic disturbances during sound playback, which remain observable regardless of how the speech signal is manipulated.

*5) Performance of Sound Source Distance Detection:* To evaluate the effectiveness of our sound source distance detection, we conduct experiments with varying actual distances between the authentication and spoofing devices. We use a Huawei P30 as the authentication smartphone and a Galaxy S10 as the spoofing device. The Galaxy S10 is placed at four distances from the Huawei P30: 3 cm, 5 cm, 7 cm, and 10 cm. At each distance, we replay 20 voice samples using the Galaxy S10 and record them with the Huawei P30. For each sample, we apply our proposed distance detection method (see Section V-A). Table VI reports the mean and standard deviation of the estimated distances at each ground truth. The results demonstrate that our method provides reliable distance estimations, effectively supporting the proposed detection mechanism.

TABLE VII
COMPARISON OF BAC (%), FAR (%), FRR (%) AND EER (%) UNDER DIFFERENT CONDITIONS

| Condition | BAC (%) | FAR (%) | FRR (%) | EER (%) |
|---|---|---|---|---|
| W/O denoising | 93.96 | 0.35 | 11.73 | 4.16 |
| W/O voice-seg | 94.72 | 4.27 | 6.29 | 4.95 |
| **Ours (full)** | **99.01** | **0.36** | **1.62** | **0.77** |

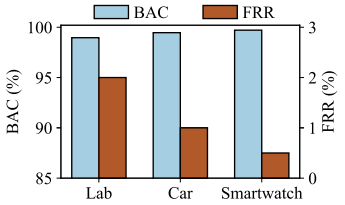



Fig. 19. The BAC and FRR performance for different environments.

*6) Performance of the Denoising Method:* Magnetic changes induced by speakers are weak and can be easily submerged by noise in the magnetometer. To suppress such interference, we apply a high-pass Butterworth filter, as described in Section V-B. To evaluate its effectiveness, we compare the system performance with and without applying the denoising method, keeping all other components unchanged. As shown in Table VII, the denoising method significantly improves BAC and reduces both FRR and EER, demonstrating its effectiveness in enhancing magnetic signal quality.

*7) Performance of the Voice-Assisted Segmentation Method:* We propose a segmentation method that leverages synchronized voice data to locate the speech-induced magnetic changes in the magnetometer signals. To assess its benefit, we compare it against a baseline that segments the magnetometer data using a fixed-size sliding window. As reported in Table VII, our voice-assisted segmentation method substantially improves BAC, FAR, FRR, and EER, confirming its superiority in isolating meaningful magnetic signal segments.

### C. Robustness of MagLive

In this subsection, we evaluate the performance of MagLive under different settings and factors, which demonstrate the robustness and effectiveness of MagLive.

*1) Impact of Spoofing Devices:* We investigate the liveness detection performance when the attacker uses different types of spoofing devices. Table IV lists 8 spoofing devices used in this study, including four smartphones, a tablet, a laptop, and two smart loudspeakers. These devices represent common types that an attacker might employ. To evaluate the performance of a spoofing device, we train the classifier using data from the remaining 7 spoofing devices. Fig. 17 illustrates the BAC and EER of MagLive for each device in this case. Among the 8 devices, smart loudspeakers are more easily detected, possibly due to their stronger magnetic effect. However, even the worst-performing device achieves a BAC of over 98% and an EER of less than 0.75%. Overall, MagLive demonstrates robustness to various spoofing devices and achieves device-irrelevance.

*2) Impact of Cross-User Training:* To evaluate the cross-user performance of MagLive, we train the classifier with

Authorized licensed use limited to: Nanyang Technological University Library. Downloaded on April 02,2026 at 16:02:23 UTC from IEEE Xplore. Restrictions apply.

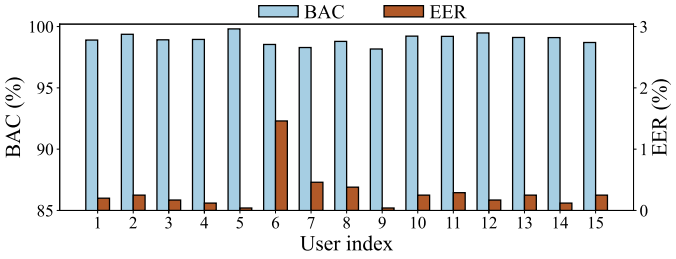



Fig. 20. The BAC and EER performance of cross-user training.

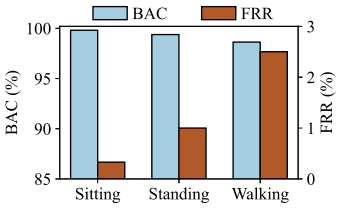



Fig. 21. The BAC and FRR performance under different postures.

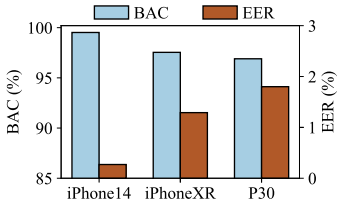



Fig. 22. The BAC and EER for cross-device evaluation.

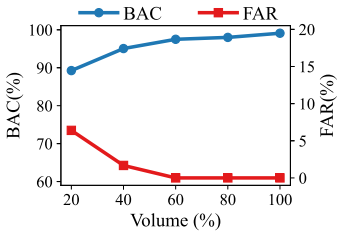



Fig. 23. The BAC and FAR at different volume levels.

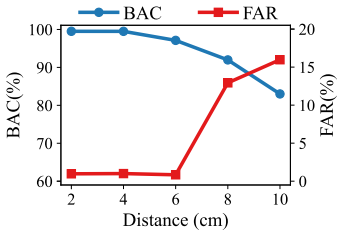



Fig. 24. The BAC and FAR at different distances.

legitimate data and spoofing data from 14 users, and then test it with the data from a remaining unseen user. Fig. 20 shows the BACs and EERs for different users. All users exhibit good performance, achieving a BAC of over 98% and an EER of less than 1.5%. In summary, MagLive demonstrates high user-irrelevant performance.

*3) Impact of Speech Type:* Our dataset consists of two types of speech: digits and voice commands. Initially, we train the classifier using the entire data of digits and test it on the voice command data. The resulting BAC is 99.36%, with an EER of 0.45%. Subsequently, we train the classifier using the voice command data and test it on the digit data. In this case, the BAC is 98.47%, with an EER of 0.63%. These results demonstrate that MagLive can achieve content-irrelevant performance.

*4) Impact of Environments:* To evaluate the impact of different magnetic environments, we adopt a cross-environment evaluation protocol. Specifically, we use the model trained on the primary training dataset, which is collected in an office environment, and test it on data collected from three additional scenarios. First, we evaluate our method in a laboratory environment surrounded by various electrical devices, such as computers, wireless keyboards, and wireless mice. Second, we conduct experiments in an in-vehicle scenario using a Volkswagen Bora. Finally, we examine a wearable-assisted scenario in which the user wears a smart wristband. Fig. 19 shows the BACs and FRRs obtained under these different magnetic field environments. The BACs for the three test environments are 98.96%, 99.46%, and 99.71%, respectively, while all FRRs remain below 2%. These results indicate that MagLive generalizes well across environments and remains robust under diverse magnetic field conditions.

*5) Impact of User Posture:* In the overall performance evaluation, the participants adopt sitting postures. In this experiment, we also consider the standing and walking postures to evaluate the impact of different user postures. For this analysis, we assess one participant and 8 spoofing devices. As shown in Fig. 21, MagLive exhibits better performance in sitting and standing postures compared to walking, suggesting potential interference from user movement. To mitigate this fluctuation, we intend to develop a model that learns movement patterns to effectively filter out noise caused by motion.

*6) Impact of Voice Volume:* We evaluate the performance of the smartphone's magnetometer at different volume levels. Specifically, we use Huawei P30 as the spoofing device to replay recordings at 20%, 40%, 60%, 80%, and 100% of the maximum volume supported by the smartphone. As shown in Fig. 23, as the volume increases from 20% to 100%, the BAC improves from 89.25% to 99.1%, while the FAR decreases from 6.4% to 0%. A higher voice volume could lead to larger magnetic changes. Typically, human normal conversation exceeds 40 dB [56], which corresponds to over 40% of the maximum smartphone volume. These results present that MagLive is available for various volume levels.

*7) Impact of Distance Threshold:* We study the impact of the distance between the sound source and the authentication device (i.e. smartphone) to determine an appropriate distance threshold. We recruit a participant to speak from various distances, and the Huawei P30 replays these voice samples accordingly. Fig. 24 shows the BACs and FARs of MagLive at different authentication distances. Notably, for better performance (e.g., BAC over 97%), the results are most satisfactory within 6 cm. The performance degradation beyond 6 cm can be mainly attributed to the attenuation of the magnetic field generated by the loudspeaker with increasing distance. To

Authorized licensed use limited to: Nanyang Technological University Library. Downloaded on April 02,2026 at 16:02:23 UTC from IEEE Xplore. Restrictions apply.

balance security and user-friendliness, we ultimately set the threshold at 6 cm for optimal performance.

*8) Impact of Authentication Smartphones:* To evaluate the generalization ability of our method across different authentication devices, we conduct experiments using three representative smartphones (iPhone 14 Pro, iPhone XR, and Huawei P30) from different hardware configurations. Three participants are recruited, and on each smartphone, each participant is asked to speak digits from 0 to 9 for 20 repetitions. The Galaxy S10 is used as the replay device to simulate spoofing attacks. We train the classifier solely using data from the iPhone 14 Pro, and test it on all three smartphones without any fine-tuning. Fig. 22 shows the BAC and EER under each test condition. The BACs for iPhone 14 Pro, iPhone XR, and Huawei P30 are 99.52%, 97.54%, and 96.9%, respectively, with corresponding EERs of 0.27%, 1.29%, and 1.8%. These results demonstrate that our model maintains high authentication accuracy even when applied to unseen authentication devices, indicating cross-device generalization despite differences in magnetometer hardware.

### D. Overhead

We implemented a prototype of MagLive to evaluate its authentication latency and computational overhead on Huawei P30. After capturing the data, we evaluated the processing time for a single data segment and calculated the average latency over 10 trials. MagLive achieves an average authentication latency of 0.1 seconds. Additionally, we used the Android Profiler tool to assess the computational overhead of MagLive. The results show that MagLive requires an average memory usage of around 59.4 MB during the authentication process.

## VII. Discussion

In this section, we discuss some limitations in our work and experiments, and provide an outlook for potential improvements in future work.

MagLive targets a realistic voice spoofing attack scenario on smartphones, where attackers record and manipulate bona fide speech samples and replay them through loudspeakers to bypass voice authentication systems. MagLive exploits magnetic pattern variations induced by speakers during speech generation for liveness detection and thereby improving robustness against spoofing attacks. Compared with existing visual- and audio-based liveness detection approaches, MagLive does not rely on favorable lighting conditions or explicit user cooperation, and can operate passively using built-in sensors in smartphones, making it suitable for practical smartphone authentication scenarios. To achieve reliable performance, MagLive estimates the distance between the sound source and the authentication smartphone and determines whether the source lies within a short-range region (e.g., within 6 cm). Our method for detecting sound source distance is lightweight and operates on a 2D plane. To enhance accuracy and universality, future work could explore incorporating deep learning technologies to develop a more sophisticated detection scheme.

While MagLive demonstrates robust performance across a wide range of practical conditions, its robustness under more extreme and rapidly changing scenarios could be further enhanced by incorporating adaptive filtering, movement-aware modeling, and dynamic decision threshold adjustment. Previous studies have shown that magnetic fields can never be completely eliminated, and simple magnetic shielding measures have limited effectiveness in evading detection [25], [57]. In future work, we aim to conduct additional experiments to explore the impact of magnetic field shielding on MagLive.

Although we have collected datasets from a group of students and devices in our experiments, the authentication-side cross-device evaluation is conducted on a limited number of smartphone models. Future work will extend the evaluation to a broader range of smartphones and participants to further validate in real-world deployments.

## VIII. Conclusion

In this paper, we propose MagLive, a robust voice liveness detection system on smartphones. MagLive utilizes the smartphone's built-in magnetometer to capture magnetic pattern changes associated with speech, eliminating the need for active sensing. By employing TF-CNN-based deep learning, self-attention-based fusion mechanisms, and supervised contrastive learning, MagLive extracts effective and robust magnetic features, achieving consistent performance in various environments. MagLive requires no specialized hardware, imposes minimal usage constraints, and demonstrates resilience against attacks. Overall, MagLive presents a promising security enhancement for existing voice authentication systems on smartphones.

## References

[1] F. Alonso-Fernandez et al., "A comparative study of fingerprint image-quality estimation methods," *IEEE Trans. Inf. Forensics Security*, vol. 2, no. 4, pp. 734–743, Dec. 2007.

[2] N. Poh et al., "Benchmarking quality-dependent and cost-sensitive score-level multimodal biometric fusion algorithms," *IEEE Trans. Inf. Forensics Security*, vol. 4, no. 4, pp. 849–866, Dec. 2009.

[3] (2015). *Voiceprint: The New WeChat Password*. [Online]. Available: https://blog.wechat.com/2015/05/21/voiceprint-the-new-wechat-password/

[4] (2016). *Citi Uses Voice Prints To Authenticate Customers Quickly And Effortlessly*. [Online]. Available: https://www.forbes.com/sites/tomgroenfeldt/2016/06/27/citi-uses-voice-prints-to-authenticate-customers-quickly-and-effortlessly/

[5] Z. Li et al., "Robust detection of machine-induced audio attacks in intelligent audio systems with microphone array," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, Nov. 2021, pp. 1884–1899.

[6] Y. Meng et al., "Your microphone array retains your identity: A robust voice liveness detection system for smart speakers," in *Proc. USENIX Secur. Symp.*, 2022, pp. 1077–1094.

[7] Q. Yang, K. Cui, and Y. Zheng, "VoShield: Voice liveness detection with sound field dynamics," in *Proc. IEEE Conf. Comput. Commun.*, May 2023, pp. 1–10.

[8] Y. Lee et al., "Using sonar for liveness detection to protect smart speakers against remote attackers," *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 4, no. 1, pp. 1–28, Mar. 2020.

[9] Y. Meng et al., "WiVo: Enhancing the security of voice control system via wireless signal in IoT environment," in *Proc. 18th ACM Int. Symp. Mobile Ad Hoc Netw. Comput.*, Jun. 2018, pp. 81–90.

[10] C. Zhao, Z. Li, H. Ding, W. Xi, G. Wang, and J. Zhao, "Anti-spoofing voice commands: A generic wireless assisted design," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 5, no. 3, pp. 1–22, 2021.

[11] S. Pradhan, W. Sun, G. Baig, and L. Qiu, "Combating replay attacks against voice assistants," *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 3, no. 3, pp. 1–26, Sep. 2019.

Authorized licensed use limited to: Nanyang Technological University Library. Downloaded on April 02,2026 at 16:02:23 UTC from IEEE Xplore. Restrictions apply.

[12] H. Li et al., "VocalPrint: A mmWave-based unmediated vocal sensing system for secure authentication," *IEEE Trans. Mobile Comput.*, vol. 22, no. 1, pp. 589–606, Jan. 2023.

[13] C. Shi, Y. Wang, Y. Chen, N. Saxena, and C. Wang, "WearID: Low-effort wearable-assisted authentication of voice commands via cross-domain comparison without training," in *Proc. Annu. Comput. Secur. Appl. Conf.*, Dec. 2020, pp. 829–842.

[14] H. Feng, K. Fawaz, and K. G. Shin, "Continuous authentication for voice assistants," in *Proc. 23rd Annu. Int. Conf. Mobile Comput. Netw.*, Oct. 2017, pp. 343–355.

[15] L. Blue, H. Abdullah, L. Vargas, and P. Traynor, "2MA: Verifying voice commands via two microphone authentication," in *Proc. Asia Conf. Comput. Commun. Secur.*, May 2018, pp. 89–100.

[16] C. Yan, Y. Long, X. Ji, and W. Xu, "The catcher in the field: A fieldprint based spoofing detection for text-independent speaker verification," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, Nov. 2019, pp. 1215–1229.

[17] L. Zhang, S. Tan, and J. Yang, "Hearing your voice is not enough: An articulatory gesture based liveness detection for voice authentication," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, Dallas, TX, USA, Oct. 2017, pp. 57–71.

[18] L. Lu et al., "LipPass: Lip reading-based user authentication on smartphones leveraging acoustic signals," in *Proc. IEEE INFOCOM - IEEE Conf. Comput. Commun.*, Apr. 2018, pp. 1466–1474.

[19] L. Wu, J. Yang, M. Zhou, Y. Chen, and Q. Wang, "LVID: A multimodal biometrics authentication system on smartphones," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 1572–1585, 2020.

[20] Y. Chen et al., "Chestlive: Fortifying voice-based authentication with chest motion biometric on smart devices," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 5, no. 4, pp. 148:1–148:25, 2022.

[21] Y. Wang, W. Cai, T. Gu, W. Shao, Y. Li, and Y. Yu, "Secure your voice: An oral airflow-based continuous liveness detection for voice assistants," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 3, no. 4, pp. 1–28, 2019.

[22] L. Zhang, S. Tan, J. Yang, and Y. Chen, "VoiceLive: A phoneme localization based liveness detection for voice authentication on smartphones," in *Proc. ACM SIGSAC Conf. Comput. Commun. Security*, New York, NY, USA, 2016, pp. 1080–1091.

[23] L. Blue, L. Vargas, and P. Traynor, "Hello, is it me You're looking for?: Differentiating between human and electronic speakers for voice interface security," in *Proc. 11th ACM Conf. Secur. Privacy Wireless Mobile Netw.*, Jun. 2018, pp. 123–133.

[24] M. E. Ahmed, I. Kwak, J. H. Huh, I. Kim, T. Oh, and H. Kim, "Void: A fast and light voice liveness detection system," in *Proc. USENIX Secur. Symp.*, 2020, pp. 2685–2702.

[25] S. Chen et al., "You can hear but you cannot steal: Defending against voice impersonation attacks on smartphones," in *Proc. IEEE 37th Int. Conf. Distrib. Comput. Syst. (ICDCS)*, Jun. 2017, pp. 183–195.

[26] S. Wang, J. Cao, X. He, K. Sun, and Q. Li, "When the differences in frequency domain are compensated: Understanding and defeating modulated replay attacks on automatic speech recognition," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur. (CCS)*, 2020, pp. 1103–1119.

[27] H. Cao et al., "LiveProbe: Exploring continuous voice liveness detection via phonemic energy response patterns," *IEEE Internet Things J.*, vol. 10, no. 8, pp. 7215–7228, Apr. 2023.

[28] Q. Wang et al., "VoicePop: A pop noise based anti-spoofing system for voice authentication on smartphones," in *Proc. IEEE Conf. Comput. Commun. (INFOCOM)*, Apr. 2019, pp. 2062–2070.

[29] L. Blue et al., "Who are you (I really wanna know)? Detecting audio DeepFakes through vocal tract reconstruction," in *Proc. 31st USENIX Secur. Symp.*, 2022, pp. 2691–2708.

[30] T. Liu et al., "MagBackdoor: Beware of your loudspeaker as a backdoor for magnetic injection attacks," in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2023, pp. 3416–3431.

[31] A. Kassis and U. Hengartner, "Breaking security-critical voice authentication," in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2023, pp. 951–968.

[32] C. Yan, X. Ji, K. Wang, Q. Jiang, Z. Jin, and W. Xu, "A survey on voice assistant security: Attacks and countermeasures," *ACM Comput. Surveys*, vol. 55, no. 4, pp. 1–36, Apr. 2023.

[33] E. Wenger et al., "'Hello, It's Me': Deep learning-based speech synthesis attacks in the real world," in *Proc. ACM Conf. Comput. Commun. Secur. (CCS)*, 2021, pp. 235–251.

[34] J. Deng, Y. Chen, Y. Zhong, Q. Miao, X. Gong, and W. Xu, "Catch you and i can: Revealing source voiceprint against voice conversion," in *Proc. USENIX Secur. Symp.*, 2023, pp. 5163–5180.

[35] Z. Yu, Y. Chang, N. Zhang, and C. Xiao, "SMACK: Semantically meaningful adversarial audio attack," in *Proc. USENIX Secur. Symp.*, 2023, pp. 3799–3816.

[36] Q. Liao, Y. Huang, Y. Huang, Y. Zhong, H. Jin, and K. Wu, "MagEar: Eavesdropping via audio recovery using magnetic side channel," in *Proc. 20th Annu. Int. Conf. Mobile Syst., Appl. Services*, Jun. 2022, pp. 371–383.

[37] C. Wu, J. Chen, K. He, Z. Zhao, R. Du, and C. Zhang, "EchoHand: High accuracy and presentation attack resistant hand authentication on commodity mobile devices," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, Nov. 2022, pp. 2931–2945.

[38] Q. Yang and Y. Zheng, "DeepEar: Sound localization with binaural microphones," *IEEE Trans. Mobile Comput.*, vol. 23, no. 1, pp. 359–375, Jan. 2024.

[39] C. Cai, H. Pu, L. Ye, H. Jiang, and J. Luo, "Active acoustic sensing for 'hearing' temperature under acoustic interference," *IEEE Trans. Mob. Comput.*, vol. 22, no. 2, pp. 661–673, 2023.

[40] T. Tao et al., "Sound localization and speech enhancement algorithm based on dual-microphone," *Sensors*, vol. 22, no. 3, p. 715, Jan. 2022.

[41] M. Wang et al., "Automatic calibration of magnetic tracking," in *Proc. 28th Annu. Int. Conf. Mobile Comput. Netw.*, Oct. 2022, pp. 391–404.

[42] L. Wang, M. Chen, L. Lu, Z. Ba, F. Lin, and K. Ren, "Voicelistener: A training-free and universal eavesdropping attack on built-in speakers of mobile devices," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 7, no. 1, pp. 1–22, 2023.

[43] H. Pan et al., "MagDefender: Detecting eavesdropping on mobile devices using the built-in magnetometer," in *Proc. 19th Annu. IEEE Int. Conf. Sens., Commun., Netw. (SECON)*, Sep. 2022, pp. 28–36.

[44] Y. Cao, F. Li, H. Chen, X. Liu, C. Duan, and Y. Wang, "I can hear you without a microphone: Live speech eavesdropping from earphone motion sensors," in *Proc. IEEE Conf. Comput. Commun.*, May 2023, pp. 1–10.

[45] K. Nguyen, H. Proença, and F. Alonso-Fernandez, "Deep learning for iris recognition: A survey," *ACM Comput. Surveys*, vol. 56, no. 9, pp. 1–35, Oct. 2024.

[46] K. Hernandez-Diaz, F. Alonso-Fernandez, and J. Bigun, "One-shot learning for periocular recognition: Exploring the effect of domain adaptation and data bias on deep representations," *IEEE Access*, vol. 11, pp. 100396–100413, 2023.

[47] C. Wu, K. He, J. Chen, Z. Zhao, and R. Du, "Liveness is not enough: Enhancing fingerprint authentication with behavioral biometrics to defeat puppet attacks," in *Proc. USENIX Secur. Symp.*, 2020, pp. 2219–2236.

[48] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2018, pp. 7132–7141.

[49] P. Khosla et al., "Supervised contrastive learning," in *Proc. Conf. Neural Inf. Process. Syst.*, 2020, pp. 18661–18673.

[50] T.-S. Ng, J. C. L. Chai, C.-Y. Low, and A. B. J. Teoh, "Self-attentive contrastive learning for conditioned periocular and face biometrics," *IEEE Trans. Inf. Forensics Security*, vol. 19, pp. 3251–3264, 2024.

[51] H. Cao et al., "HandKey: Knocking-triggered robust vibration signature for keyless unlocking," *IEEE Trans. Mobile Comput.*, vol. 23, no. 1, pp. 520–534, Jan. 2024.

[52] *Sensor Logger*. Accessed: JAN. 26, 2024. [Online]. Available: https://www.tszheichoi.com/sensorlogger

[53] C. Wu, K. He, J. Chen, Z. Zhao, and R. Du, "Toward robust detection of puppet attacks via characterizing fingertip-touch behaviors," *IEEE Trans. Dependable Secure Comput.*, vol. 19, no. 6, pp. 4002–4018, Nov. 2022.

[54] E. Gonzalez-Sosa, J. Fierrez, R. Vera-Rodriguez, and F. Alonso-Fernandez, "Facial soft biometrics for recognition in the wild: Recent works, annotation, and COTS evaluation," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 8, pp. 2001–2014, Aug. 2018.

[55] C. Wu et al., "It's all in the touch: Authenticating users with HOST gestures on multi-touch screen devices," *IEEE Trans. Mobile Comput.*, vol. 23, no. 10, pp. 10016–10030, Oct. 2024.

[56] Y. Chen, J. Yu, L. Kong, H. Kong, Y. Zhu, and Y.-C. Chen, "RF-Mic: Live voice eavesdropping via capturing subtle facial speech dynamics leveraging RFID," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 7, no. 2, pp. 1–25, 2023.

[57] Z. Liu et al., "Camradar: Hidden camera detection leveraging amplitude-modulated sensor images embedded in electromagnetic emanations," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 6, no. 4, pp. 1–25, 2023.


Authorized licensed use limited to: Nanyang Technological University Library. Downloaded on April 02,2026 at 16:02:23 UTC from IEEE Xplore. Restrictions apply.

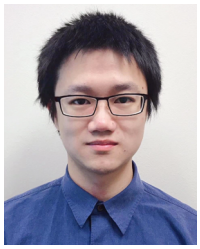
**Xiping Sun** received the B.E. degree in information security from Wuhan University, China, in 2019, where she is currently pursuing the Ph.D. degree with the School of Cyber Science and Engineering. Her research interests include system and mobile security.

**Jing Chen** (Senior Member, IEEE) received the Ph.D. degree in computer science from the Huazhong University of Science and Technology, Wuhan. He is currently a Full Professor with the School of Cyber Science and Engineering, Wuhan University. He has published more than 150 research papers in many international journals and conferences, including USENIX Security, ACM CCS, INFOCOM, IEEE TRANSACTIONS ON DEPENDABLE AND SECURE COMPUTING, IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, IEEE TRANSACTIONS ON MOBILE COMPUTING, IEEE TRANSACTIONS ON COMPUTERS, IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, and IEEE TRANSACTIONS ON SERVICES COMPUTING. His research interests include the areas of network security, cloud security, and mobile security. He was the Runner-Up for the Best Paper at INFOCOM 2018 and INFOCOM 2021. He has served as the Vice Chair for ACM Turing Award Celebration Conference (TURC) 2023.

**Cong Wu** received the B.E. degree from Xidian University and the Ph.D. degree from Wuhan University. He is currently a Full Professor with Wuhan University. Before that, he was a Post-Doctoral Researcher with The University of Hong Kong and a Research Fellow with Nanyang Technological University. His research interests focus on the security and privacy of intelligent systems. He received seven best paper awards of international flagship conferences. He is an Associate Editor of IMWUT, SPY, and IJCS.

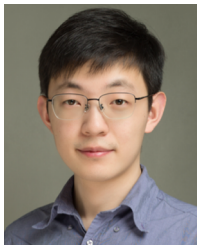
**Kun He** (Member, IEEE) received the Ph.D. degree from Wuhan University, Wuhan, China. He is currently an Associate Professor with Wuhan University. He has published more than 70 research papers in various conferences and journals, such as S&P, USENIX Security, CCS, NDSS, INFOCOM, IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, IEEE TRANSACTIONS ON DEPENDABLE AND SECURE COMPUTING, IEEE TRANSACTIONS ON NETWORKING, and IEEE TRANSACTIONS ON MOBILE COMPUTING. His research interests include cryptography and data security.

**Haozhe Xu** received the B.E. degree in information security from Wuhan University, China, in 2022, where he is currently pursuing the M.S. degree with the School of Cyber Science and Engineering. His research interests include authentication and liveness detection.



**Yebo Feng** received the Ph.D. degree in computer science from the University of Oregon (UO) in 2023. He is a Research Fellow with the School of Computer Science and Engineering (SCSE), Nanyang Technological University (NTU). His research interests include network security, blockchain security, and anomaly detection. He was a recipient of the Best Paper Award of 2019 IEEE CNS, the Gurdeep Pall Graduate Student Fellowship of UO, and the Ripple Research Fellowship. He has served as the reviewer for IEEE TRANSACTIONS ON DEPENDABLE AND SECURE COMPUTING, IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, ACM TKDD, IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS, and IEEE COMMUNICATIONS SURVEYS AND TUTORIALS. He has been a member of the program committees for international conferences, including SDM, CIKM, and CYBER; and has also served on the Artifact Evaluation (AE) Committee for USENIX OSDI and USENIX ATC.

**Ruiying Du** received the B.S., M.S., and Ph.D. degrees in computer science from Wuhan University, Wuhan, China, in 1987, 1994, and 2008, respectively. She is a Professor with the School of Cyber Science and Engineering, Wuhan University. She has published more than 80 research papers in many international journals and conferences, such as IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, USENIX Security, CCS, INFOCOM, SECON, TrustCom, and NSS. Her research interests include network security, wireless networks, cloud computing, and mobile computing.



**Xianhao Chen** (Member, IEEE) received the B.Eng. degree in electronic information from Southwest Jiaotong University in 2017 and the Ph.D. degree in electrical and computer engineering from the University of Florida in 2022. He is currently an Assistant Professor with the Department of Electrical and Electronic Engineering, The University of Hong Kong. His research interests include wireless networking, edge intelligence, and machine learning. He received the 2022 ECE Graduate Excellence Award for research from the University of Florida. He serves as an Associate Editor for *ACM Computing Surveys*.

Authorized licensed use limited to: Nanyang Technological University Library. Downloaded on April 02,2026 at 16:02:23 UTC from IEEE Xplore. Restrictions apply.