Exploring Acoustic Reverse Nonlinearity Against Speech Forgery in Real-time Voice Applications

Ming Gao[1], Lingfeng Zhang[2,3], Yike Chen[2], Sifeng He[3], Feng Qian[3], Lei Yang[3], Fu Xiao[1]✉, Jinsong Han[2]
[1] School of Computer Science, Nanjing University of Posts and Telecommunications, China
[2] School of Computer Science and Technology, Zhejiang University, China
[3] Ant Group, China
{gaomingppm, xiaof}@njupt.edu.cn, {chenyike, hanjinsong}@zju.edu.cn
{zhanglingfeng.zlf, youzhi.qf}@antgroup.com, {hsf215kg, leiyang.ucsb}@gmail.com

***Abstract*—Unauthorized editing of speech recordings poses a significant threat to the security and authenticity of speeches, particularly in the forensic and legal fields. Even worse, the speech is increasingly at risk of being tampered with due to the development of AI techniques (e.g., Audio Deepfake). It is difficult for normal users to guarantee what they say has not been illegally changed. Audio watermark techniques are recognized as an active method against speech forgery. However, such techniques suffer from audio quality degradation and non-real-time insertion. Therefore, they cannot be adopted into real-time voice applications against forgery on remote recordings, e.g., phone calls, live broadcasts, and online meetings. Fortunately, high-definition (HD) audio techniques provide ultrasonic bands without distortion. Therefore, ultrasonic creditable factors can be utilized. We propose an audio tamper-proof system, named Aegis. It provides commodity mobile devices (e.g., smartphones) with an effective method of real-time insertion of inaudible creditable factors. Users can claim that audio with no or mismatched ultrasound is invalid and illegal. In particular, we explore the acoustic reverse-nonlinear phenomenon where audible signals can be modulated onto the ultrasonic spectrum. By emphasizing the correlation between speech signals and ultrasound, we realize effective defense against various tampering methods.**

***Index Terms*—Ultrasound, mobile sensing, acoustic nonlinearity, tamper-proof detection.**

I. Introduction

The authenticity of speeches is fundamental in various fields ranging from business to court. However, the speech is vulnerable to *forgery*. As illustrated in Fig. 1, the adversary (Bob) can tamper with the speech by replacing the 'lends to' segment with 'borrows from'. In this case, the fact is completely twisted. To make matters worse, AI-enabled techniques [1]–[3], e.g., Audio Deepfake [4], recently pose a new yet severe threat to speech security. It is hard for normal users to prove that edited recordings or synthetic audio are fake.

Existing active tamper-proof techniques rely essentially on trustworthy third parties (e.g., security organs) or factors (e.g., watermark [5]–[7]). However, they face practical limitations. Professional staff is equipped with site enforcement recorders (costing over $57.5 each) [8] to confirm speech authenticity in the prosecution and legal fields. However, normal users who cannot always carry such specialist devices can hardly perform forensics. Audio fragile watermark techniques [9] are deemed

✉ Fu Xiao is the corresponding author.

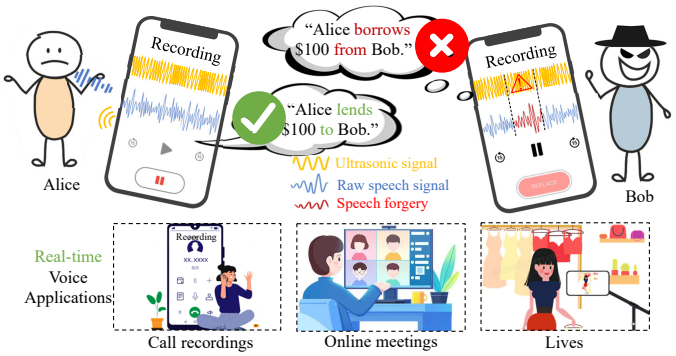


Fig. 1. An application scenario for Aegis, a tamper-proof technique for real-time voice applications. Adversaries, as remote/online participants, record the user's speeches and edit, tamper with, or forge them maliciously. Aegis inserts ultrasonic creditable factors into audio timely to defend speech authenticity.

an active solution to defend integrity and authenticity against speech forgery. If the inserted fragile watermarks are distorted or lost, the audio is recognized to be fake. However, watermark techniques [5]–[7], [10] are faced with several limitations in practice. Time-delayed manual operations are required for watermark insertion, along with audio quality degradation.

The issue remains unresolved for users who intentionally safeguard their speeches, especially in real-time voice applications. Remote participants might maliciously record and manipulate the users' speeches from phone calls, live broadcasts, online meetings, and other such interactions. Consequently, it becomes challenging to distinguish the authentic recording from the contradictory versions.

Benefiting from high-definition (HD) audio techniques, mobile devices (represented by smartphones with typical sampling rates of 44.1 kHz or 48 kHz) are able to record and transmit audio with undistorted ultrasonic features. The lossless audio compression has increasingly replaced the lossy compression (e.g., MP3) [11]. Popular audio compression techniques obtain high sampling rates of at least 44.1 kHz. For example, WAV, APE, and FLAC sample at 44.1 kHz, while advanced audio coding (AAC) supports a high sampling rate up to 96 kHz. Moreover, OPUS in the on-the-top (OTT) communication [12] and enhanced voice services (EVS) [13] in VoLTE offer a wide band of 48 kHz. Thus, mobile devices

are capable of acoustic (along with ultrasound) collection and transmission within 22.05 kHz or 24 kHz.

The ultrasonic creditable factors express the potential in real-time insertion against forgery. During speech recording, mobile devices play encoded ultrasonic signals. In this way, the ultrasonic creditable factors are inserted into audio timely. Such a method is inaudible to humans, along with high throughput [14]. Although low-pass filters can directly remove ultrasonic features, recordings with their ultrasound eliminated would be naturally invalid. Therefore, a commercial off-the-shelf (COTS) smartphone can serve as an audio enforcement recorder for normal users. In practice, the design of ultrasonic tamper-proof should overcome the two following challenges.

1) **Ultrasound-sound Correlation:** How to design ultrasonic creditable factors to relate with speeches in case of malicious editing, tampering, and forgery? The creditable factors should be capable of resisting sophisticated adversaries who might forge/replicate ultrasonic characteristics. If independent of speeches, the scheme is vulnerable to copy-move forgery [15]. As a countermeasure, multiple ultrasonic effects are exploited to characterize speeches, including Doppler Frequency Shift (DFS) and Time-of-Flight (TOF). In particular, we discover a novel phenomenon of acoustic reverse nonlinearity (RNL). Different from the wide belief that nonlinear harmonics occur merely among the band over 25 kHz [16]–[22], we observe that *low-frequency speech signals can be non-linearly modulated onto ultrasonic signals*, and such a phenomenon is common among COTS smartphones.

2) **Resilience:** How to resist various tampering attacks, especially those potentially emerging in the future? Instead of features introduced by forgery, our proposed method detects whether the ultrasonic creditable factors match the speeches. Therefore, it can effectively defend against various tampering methods, including both copy-move and Audio Deepfake, even for a future forgery technique.

We apply ultrasound to building a tamper-proof system, named Aegis. The basic idea is illustrated in Fig. 2. The ultrasonic carriers are encoded (upon predetermined information, e.g., device identification and time stamps) and emitted by the on-board speakers of mobile devices. In this way, ultrasonic creditable factors (including ultrasonic codes and ultrasound-sound correlation) are embedded into the audio. Aegis checks these factors for tamper-proof detection. We can detect the conflict between the ultrasound segments and the maliciously modified fragments of 'borrows from' in Fig. 1, which indicates the existence of forgery. Aegis provides normal users with an effective method for anti-forgery, with the characteristics of 'seamless integration' on COTS mobile devices. Extensive evaluations validate the defending effectiveness of Aegis in real-world scenarios.

Our contributions are summarized as follows:

- We realize Aegis, a new tamper-proof system, especially for real-time voice applications. It is the first to leverage the ultrasonic credible factors to defend against speech forgery. Aegis is resistant to diverse tampering methods and robust against future adversaries.

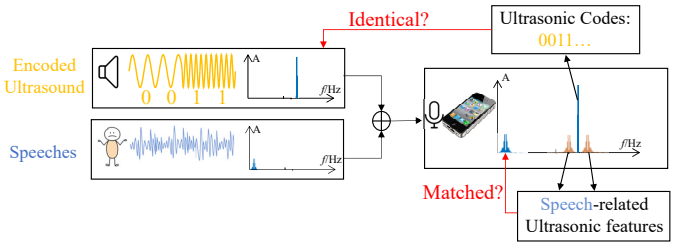


Fig. 2. Aegis modulates sounds upon encoded ultrasound using reverse nonlinearity. It checks the ultrasonic codes and detects the ultrasound-sound correlation for tamper-proof.

- Aegis enables 'seamless integration' deployment on COTS mobile devices, free from audio quality degradation. It presents a promising way to help normal users equipped merely with mobile devices to improve the authenticity of their speeches.
- We explore a new ultrasonic characteristic. Unlike the common sense on acoustic non-linearity, we find that audible signals can reversely convert into ultrasonic bands. We believe this reverse nonlinearity will significantly expand the acoustic application scope.

II. Background

Before presenting the system design, we describe a typical application scenario and analyze possible forgery threats.

A. Application Scenarios

Our main motivation is to resist editing, tampering, and forgery of audio recordings, especially in real-time voice applications. Some common scenarios include phone calls, voice messages, live broadcasts, and online meetings. Remote/online participants might record and forge the user's speeches, whereas traditional watermarks cannot provide real-time protection on remote recordings.

In practice, the user installs Aegis on his/her smartphone and holds it using the two most popular holding gestures: towards the microphone (keeping the bottom microphone about 2∼4 cm away from the mouth) or 'Phone Call'. While talking, Aegis emits modulated ultrasonic signals via on-board speakers and records them via microphones. In this way, ultrasonic creditable factors (including ultrasonic codes and ultrasound-sound correlation) are embedded into the audio.

Users or an impartial third party, such as blockchains, governments, courts, or arbitration associations, can assert that audio featuring mismatched or absent ultrasonic characteristics is invalid and illegal. Thereby, Aegis prevents economic losses and legal disputes caused by speech forgery. It has a wide applicable scope, ranging from individual daily recordings to digital law evidence, e.g., oral statements, voice signatures, and any other audio recordings.

Our paper aims at a forensics solution for users who actively protect their speeches. It is also suited for non-real-time applications. The cases with no measures taken are out of our scope. Besides, Aegis is not for live detection, user authentication, nor defending voice assistants against 'inaudible' commands [17], [22] and adversarial audio samples [23].

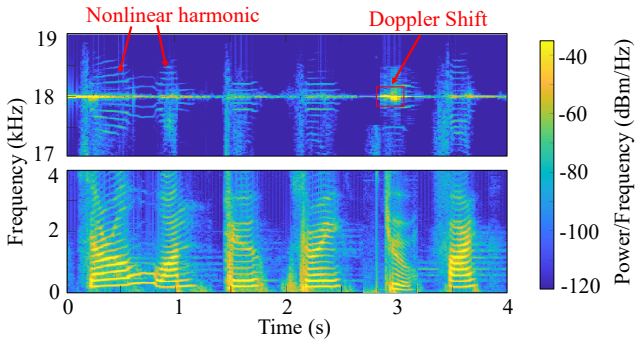


Fig. 3. An illustration of the T-F domain features of a speech segment.

B. Threat Model

We assume the adversary possesses sufficient knowledge about our mechanisms. The following attacks are considered to pose a threat to speech authenticity.

Audio Edition. Adversaries could leverage various traditional editing means, e.g., segment deletion and copy-move (splicing segments prerecorded from the same speaker). They can also copy-move or re-modulate the corresponding ultrasonic features to bypass Aegis.

Replay Attack. Adversaries might replay segments prerecorded from the same speaker.

Imposter Attack. Adversaries attempt to imitate the articulatory gestures of victims to forge the audio.

Synthetic Attack. Adversaries can utilize learning-based audio-generation techniques, of which the most popular ones are text to speech (TTS) [24] and voice conversion (VC) [25]. They also attempt to generate the ultrasonic features using generative adversarial network (GAN) [26].

Hybrid Attack. The adversaries may combine multiple means. For example, the adversary might execute a replay attack for RNL and an imposter attack for DFS and TOF.

III. Multiple Ultrasonic Effect Exploitation

We model the acoustic reverse nonlinearity and explore the simultaneous utilization of multiple ultrasonic effects.

A. Modeling Reverse Nonlinearity

In the literature, the common sense on the acoustic nonlinearity is in a direction from the ultrasound band to the low-frequency band [16]–[22]. On the contrary, our observed acoustic reverse nonlinearity shows a different direction, in which the low-frequency component that is below 1 kHz is migrated up to the ultrasonic bands above 17 kHz.

Due to the reverse linearity of microphones, tones in the speech signals are modulated into the ultrasonic carrier $U(t)$. A microphone will record the mixture of the inputs (i.e., $V(t)+U(t)$) with their quadratic terms as follows,

$$\begin{aligned} R(t) =& V(t)+U(t)+A_{nl}(V(t)+U(t))^2 \\ =& V(t)+U(t)+A_{nl}(V^2(t)+2V(t)U(t)+U^2(t)), \end{aligned} \tag{1}$$

where A_{nl} is the nonlinear gain. In the low-frequency band, the nonlinearity is insignificant, so the term of $V^2(t)$ can be ignored. Besides, the term of $U^2(t)$ will be removed by the low-pass filter in the microphone. Such characteristics induce the RNL harmonics $H_{nl}=2V(t)U(t)$ as follows,

$$\begin{aligned} H_{nl} &= 2\sum A_{nli}A_i cos(2\pi f_i t+\varphi_i) \ \cdot Ucos(2\pi f_u(t)t) \\ &= \sum A_{nli}A_i U cos(2\pi(f_u(t)\pm f_i)t-\varphi_i), \end{aligned} \tag{2}$$

where A_{nli} is the nonlinear gain of the i-th RNL harmonic and it is determined by f_i and f_u jointly. The $f_u(t)\pm f_i$ implies the double sideband nonlinear modulation where the RNL harmonics of f_is are symmetrically distributed around the ultrasonic carrier of $f_u(t)$.

Our finding enables a bi-directional acoustic nonlinearity effect, which is promising for future acoustic applications. This phenomenon lays the foundation for Aegis to comprehensively utilize multiple ultrasonic effects. Although these RNL harmonics can map the speech signals, we notice that audio played by a loudspeaker can also trigger reverse nonlinearity. In other words, RNL characteristics are vulnerable to a replay attack. Fortunately, the TOF and DFS can improve the resilience of the ultrasound-sound correlation to attacks. Therefore, we explore the feasibility of leveraging multiple ultrasonic effects together.

B. Simultaneous Utilization of Multiple Effects

Existing ultrasonic sensing approaches are mostly based on a single ultrasonic effect. That is, they employ either DFS in the continuous wave (CW) [27] or TOF in the frequency modulated continuous wave (FMCW) [28]. Nevertheless, DFS is also present in FMCW signals [29]. Meanwhile, the speech signals will be coupled with the ultrasound and generate RNL ultrasonic harmonics. This indicates that multiple ultrasonic effects would co-exist. This phenomenon motivates us to consider a novel sensing scheme that utilizes multiple effects to characterize speech signals.

We conduct a pilot experiment on a SAMSUNG Galaxy S8. The smartphone's bottom microphone is kept horizontally 2 cm away from the user's mouth. The user reads numbers from 'Zero' to 'Nine'. The voice volume is about 66 dB, during which the smartphone keeps playing the ultrasound of 18 kHz at its highest volume. The intensity of ambient noise is 46.1 dB. We measure the received signals, including the line-of-sight (LOS) and reflected signals. Their strength is 97 dB (flat weighting). After filtering out the LOS signals, we present the time-frequency (T-F) spectrogram in Fig. 3. Besides the DFS of within $\pm$150 Hz, it can be obviously observed that the low-frequency speech signals are coupled with the ultrasound. The high-frequency RNL harmonics are symmetrically distributed on either side around 18 kHz within a region of $\pm$1 kHz. After being processed by a square-law demodulator [30], the cross-correlation coefficients [31] between the RNL harmonics and their corresponding speech segments are more than 0.7, while those with the other speeches are below 0.12. Empirically, the reverse nonlinearity occurs, if the ultrasonic intensity is above 60 dB SPL (sound pressure level).

The reverse nonlinearity is common among 18 tested mobile

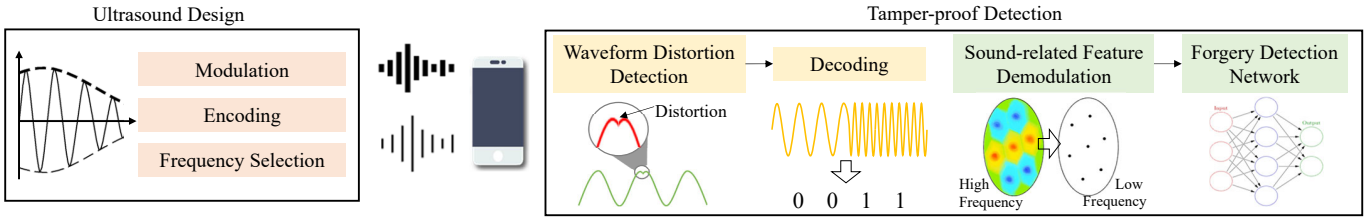


Fig. 4. System Overview. Against speech forgery, Aegis emits well-designed ultrasonic signals, which are easily accessed on mobile devices. The ultrasounds are recorded synchronously with speeches. The ultrasonic waveform with codes and the ultrasound-sound correlation jointly enable tamper-proof detection.

phones (including Samsung, HUAWEI, OPPO, etc.), along always with DFS and TOF features. With this insight, Aegis could obtain abundant characteristic information merely using ultrasound from COTS smartphones.

IV. System Overview

We propose Aegis, an audio tamper-proof system on mobile devices. It leverages the multiple ultrasonic effects on COTS smartphones. The system design is illustrated in Fig. 4.

In ultrasound transmission, after modulation, encoding, and frequency selection, Aegis emits ultrasonic signals using the on-board speaker, carrying the ultrasonic codes. Such codes will be inserted into the recorded audio timely. The ultrasonic codes could be composed of device identification, time stamps, etc, under a predetermined rule.

In the tamper-proof detection, Aegis analyzes the received audio (including speech signals and reflected ultrasound). It defends speech forgery from two perspectives. On the one hand, Aegis detects the waveform distortion and decodes the ultrasonic codes in case of audio editing. On the other hand, Aegis matches the reflected ultrasound and speech signals using a network against other tampering methods.

V. Ultrasonic Signal Design

The design of ultrasonic signals aims at the following goals.

(i) Robustness (Security): In reality, a sophisticated adversary is likely to modify the speeches as well as the ultrasonic components to bypass Aegis. For example, CW is widely adopted in ultrasound-based applications [32]–[36]. However, it is vulnerable to the copy-move forgery due to its constant ultrasonic carriers. Therefore, it is necessary to explore a resilient modulation method for better defense. In particular, though the adversaries could remove the high-pass ultrasonic features using low-pass filters or lossy compression (e.g., MP3), the users could claim such audio is invalid.

(ii) Throughput: The ultrasonic signals should be encoded with a high data rate and a low error rate.

(iii) Imperceptibility (Effectiveness): The selected inaudible signals should make ultrasonic effects significant enough. Thereby, the correlation is distinctly exhibited. These effects, especially RNL, are sensitive to the ultrasonic frequency. It is necessary to select appropriate bands beyond human hearing.

A. Modulation Mode: sinusoidal FMCW

An FMCW signal can simultaneously provide the DFS, TOF, and RNL characteristics. A simple scheme is to utilize a linear FMCW signal. As a representative, a chirp signal sweeps linearly over time within a specific band. However, this scheme is vulnerable. Homologous ultrasonic carriers (corresponding to identical speech segments) keep a fixed frequency difference, based on which these carriers can be mutually transformed using a square-law demodulator [30]. Hence, an adversary can easily bypass detectors based on linear FMCW by copy-move with ultrasound re-modulation.

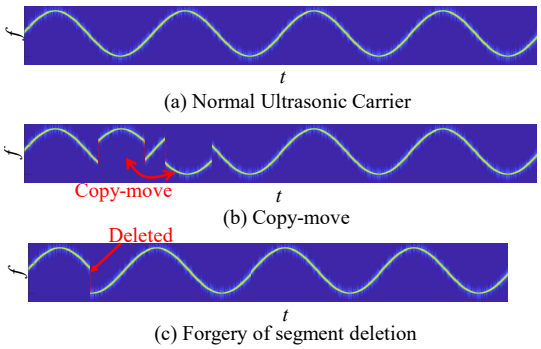


Fig. 5. Illustration of the sinusoidal FMCW modulation of the ultrasonic carrier and the distortions caused by the forgery of copy-move and deletion.

To resist the audio edition, we employ a sinusoidal FMCW signal. Its frequency varies sinusoidally across a bandwidth B with a duration τ on a frequency bias F_{bias}, i.e., $f_u = \frac{B}{2}cos(2\pi\frac{1}{\tau}t) + F_{bias}$. Thus, we have

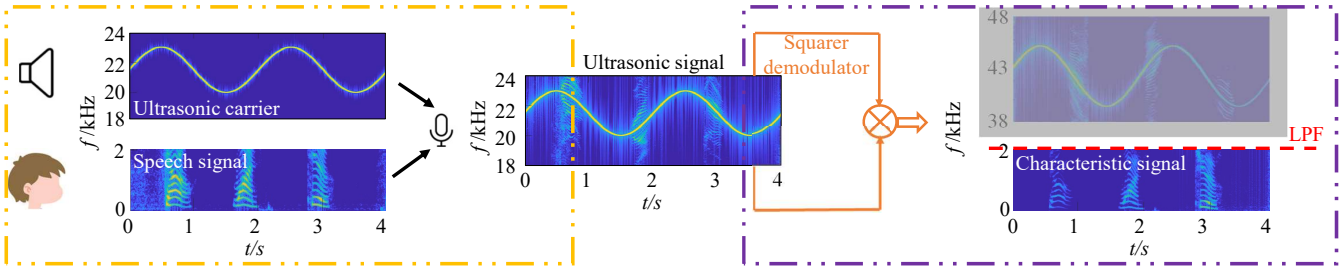
$$\begin{aligned} U(t) &= A_u cos(2\pi \int_0^t f_u dt) \\ &= A_u cos[\frac{B\tau}{2} sin(2\pi\frac{t}{\tau}) + 2\pi F_{bias} t]. \end{aligned} \tag{3}$$

Comparing the Fig. 5(a) with 5(b)&(c), the audio edition would introduce waveform distortion in the frequency domain.

In particular, the sinusoidal FMCW modulation resists the re-modulation of ultrasound. To edit the nonlinear FMCW, advanced tools require a high sampling rate, more than sixfold the signal frequencies [37]. Otherwise, they would distort the signals. The requirement can hardly be satisfied by acoustic sampling rates in most modern smartphones and audio compression methods [27]. Therefore, we defend against the audio edition by detecting these distortions.

B. Encoding: AM

We encode the ultrasonic signals using amplitude modulation (AM). The ultrasonic codes act as an additional credible



(a) Modulation (b) Demodulation

Fig. 6. Illustration of ultrasonic modulation (sinusoidal FMCW) and demodulation (square-law demodulator).

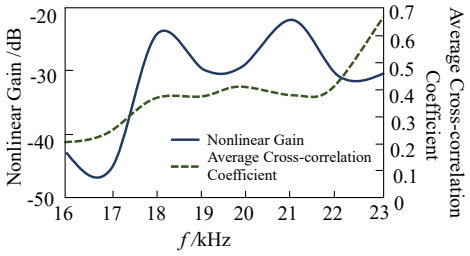


Fig. 7. Nonlinearity of a Samsung Galaxy S8.

factor. Although frequency modulation (FM) is more popular [14], it conflicts with our adopted sinusoidal FMCW.

In the ultrasonic codes, the lower amplitude (empirically set as 60∼63 dB SPL, diverse among devices) is encoded as '0', and the higher (at approximately 68 dB SPL) is as '1'. The ultrasonic codes follow a synchronization code, which is used for watermark location. The requirement for ultrasonic intensity can be satisfied by mobile devices, which usually support up to 90 dB SPL. Such a setup also meets the human-exposure limit of below 70 dB SPL, suggested by the International Non-Ionizing Radiation Committee [38].

C. Frequency Selection

The carrier frequency should be inaudible to humans and capable of fully exhibiting each ultrasonic characteristic.

The intensity of RNL characteristics depends on the ultrasonic carrier frequency. As a representative, Fig. 7 presents the nonlinear gains of a Samsung Galaxy S8 to different ultrasonic carrier frequencies and their average cross-correlation coefficients with the corresponding speech segments. RNL performs significantly over 18 kHz. To remain two adjacent sidebands for the RNL characteristics (with a bandwidth of approximately ±1 kHz), we select the band of 19∼21 kHz (i.e., F_{bias}=20 kHz and B=2 kHz in Eq. 3).

The selected band is inaudible. Along with the RNL sidebands, the ultrasonic features are distributed among 18∼22 kHz. If being below 18 kHz, RNL characteristics are probably audible and thus degrade audio quality. If being over 22 kHz, the aliasing distortion occurs under the sampling rate of 44.1 kHz according to the Nyquist sampling theorem. In addition, we consider the ringing effect [16]. It results in audible noise, due to impulses caused by discrete frequency changes. Hence, we set a slow frequency change of τ=2 s to avoid the ringing effect.

The selected band is also effective for TOF and DFS. Typically, an articulatory gesture lasts 100∼700 ms [39] at a speed of approximately ±80 cm/s [27]. The corresponding DFS ranges up to 100 Hz at carriers of over 18 kHz. About 5,000 samples can be utilized to represent each articulatory gesture. A wider bandwidth of 2 kHz also supports an acceptable distance resolution in the TOF [40].

Note that the above design applies to various acoustic sampling rates, including 44.1 kHz, 48 kHz, 96 kHz, etc.

VI. Tamper-proof Detection

We detect the waveform distortion and ultrasonic codes in case of the audio edition. Then, the ultrasound-sound correlation is emphasized against the other tampering methods.

A. Distortion Detection

We detect the integrity of ultrasonic signals. Audio with no or distorted ultrasound would be recognized to be invalid.

We utilize the slope of the curve (i.e., the derivative) to check the sinusoidal continuity of waveforms in the frequency domain. This method detects the discontinuity points caused by the audio editing. In detail, we transform the temporal audio into the T-F spectrogram via the short-time Fourier transform (STFT). Here, the STFT is computed using a Hann window of 85 ms, a hop length of 10 ms, and an FFT size of 4096. Based on an adaptive threshold using the maximum entropy method [41], the T-F spectrogram is binarized. In the binarized sinusoidal curve, the derivative of each point should follow a cosine pattern. On the contrary, there would exist a sudden change in the derivatives at a discontinuity point. Accordingly, the modified signal is distinguished from the normal ones.

In addition, microphone fingerprint [42] can be utilized. It can detect whether the sound and ultrasound are from an identical microphone. This method helps to avoid fake ultrasonic characteristics that are illegally recorded via malicious microphones or forged.

B. Decoding

We decode the ultrasonic codes as a supplementary way against the audio edition. The maximum entropy method [41] is used to obtain an adaptive threshold to distinguish '0' with low ultrasonic intensities and '1' with high ones. Benefiting from advanced ultrasonic communication techniques, the throughput of the ultrasonic codes is high up to 1 kbps.

The above means jointly defend against the audio edition. In case adversaries conduct replay, imposter, or synthetic attacks, we correlate the ultrasonic characteristics with speeches against forgery in the following.

C. Ultrasonic Feature Demodulation

When leveraging a learning-based method for audio-related applications, an intuitive and simple scheme is to directly input the received audio (mixing ultrasonic and speech signals) into the network. However, this scheme underutilizes these ultrasonic effects, especially in the frequency domain. Separating the mixed temporal audio into speech segments and ultrasonic signals and transforming them into T-F spectrograms [35] benefits the network in extracting the correlation. Before the STFT, we extract the ultrasonic features from the mixed audio using a square-law demodulator.

A square-law demodulator can migrate modulated signals from high-frequency carriers to low-frequency bands [27], [30]. It produces an output proportional to the square of the input, and then removes high-frequency components using a low-pass filter (LPF), with an example illustrated in Fig. 6(b). To be specific, we separate the ultrasonic signal using a high-pass filter (HPF), of which the cut-off frequency is 18 kHz. The modulated signal is denoted as $U(t) \cdot F(t)$, where $F(t)$ is the characteristic signal modulated on the ultrasonic carrier $U(t) = Ucos(2\pi \int_0^t f_u dt)$. Through a squarer (i.e., multiplying $U(t) \cdot F(t)$ by itself), we obtain

$$[U(t) \cdot F(t)]^2 = \frac{1}{2}F^2(t) + \frac{1}{2}U^2F^2(t)cos(4\pi \int_0^t f_u dt). \quad (4)$$

An LPF with a cut-off frequency of 2 kHz is utilized to filter out the high-frequency components (i.e., the second term in Eq. 4). Therefore, we extract the characteristic signals $F^2(t)$ (composed of the DFS, TOF, and RNL characteristics). Here, we cannot directly distinguish or separate each ultrasonic characteristic from the others due to the complexity of an articulatory gesture [43], especially from the square of the mixed characteristic signals, i.e., $F^2(t)$. Therefore, we propose a learning-based method for correlating the characteristic signals with the speech signals in case of forgery.

D. Forgery Detection Network

We leverage a multi-modal framework to emphasize the correlation between the ultrasonic characteristics and speech signals. Thus, it facilitates detecting audio forgery.

1) Network Input: We pre-process the speech signals $V(t)$ and the characteristic signals $F^2(t)$ via three steps. (1) Filter: We maintain the audible bands below 2 kHz using an LPF. Such a band covers the fundamental band of human voice [27], [35]. (2) Segment: We locate cutting points based on the spectral entropy [44] of speech signals. In particular, the speech signals and the characteristic signals share the same cutting points due to the natural alignment. Accordingly, we divide the two signals, and each pair of segments corresponds to a single word. (3) STFT: To utilize the audio pattern in both the time and frequency domains, we transform the two signals into T-F spectrograms using the STFT with the identical parameters in Sec. VI-A. We combine the two processed T-F spectrograms as a two-channel input.

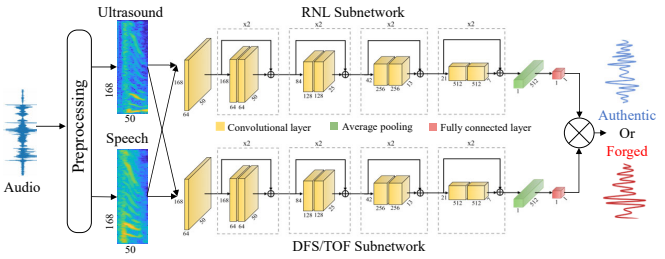


Fig. 8. Network structure of parallel training mechanism. The two subnetworks are focused on RNL and DFS/TOF respectively. This mechanism reduces training cost, compared to joint optimization on all ultrasonic characteristics, which are distributed among different bands (150 Hz vs. 1 kHz).

2) Network Design: RNL and DFS/TOF characteristics are distributed in different bands, in which the former ranges within 1 kHz and the latter is mainly below 150 Hz. If directly combining the loss of all characteristics with significant frequency difference for joint optimization, it would be hard to train the network with our limited dataset, not to mention the required excessive computation resources. To well balance the computational cost and performance, we leverage a parallel training mechanism. The network structure is shown in Fig. 8.

In each parallel subnetwork, we utilize a ResNet18 [45]. The binary cross entropy loss function is adopted as follows,

$$BCE = -\frac{1}{N}\sum_{i}^{N} y_i \cdot log(p(y_i)) + (1-y_i) \cdot log(1-p(y_i)) \quad (5)$$

where y_i is the binary index of the sample i, in which $y_i = 1$ if the sample i is positive, otherwise $y_i = 0$, and $p(y_i)$ is the output prediction. This training mechanism is resilient against adversaries who tamper with one or several characteristics independently. An AND gate combines the predictive values of the two subnetworks with the same weight, of which the samples predicted to be both positive (authentic) are recognized to be forgery-free.

Combining the above efforts together, Aegis realizes a 'seamless integration' tamper-proof detection on COTS mobile devices, especially in real-time scenes.

VII. Evaluation

We implement and evaluate Aegis with COTS smartphones. All experiments follow the approved IRB protocol. We will release our datasets to facilitate the ultrasonic sensing research after necessary data desensitization.

A. Experimental Setup

Dataset. To train and evaluate Aegis, we construct the first multi-effect ultrasound-speech dataset. We recruit 28 volunteers (14 males and 14 females, aged from 20 to 50). Each volunteer is asked to read two texts: the most common 3000 words [46] and 200 sentences in the TIMIT speech corpus [47]. Data are collected from five smartphones (SAMSUNG Galaxy S8, Mi 10, HUAWEI P20 Pro, iQOO 3, and OnePlus 9). The volunteers hold the smartphone horizontally

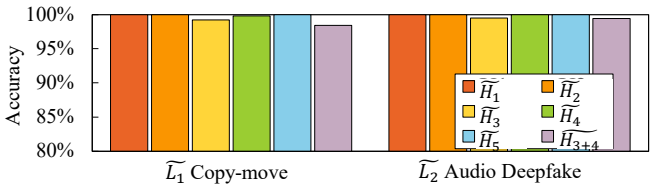


Fig. 9. Overall performance against forgery.

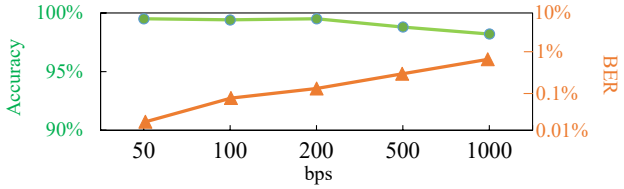


Fig. 10. Throughput.

or in the posture of the phone call. Their voice volumes are at about 66 dB in a quiet room with ambient noise of about 45.7 dB. The bottom microphone keeps 2∼4 cm away from volunteers' mouths. The smartphones are sampled at 44.1 kHz.

Attack means. Guided by Sec. II-B, we design the following attack means to cover the most of possible forgery. In the audible band, we consider two means:

- $\widetilde{L_1}$ *Cope-move*: We splice or delete segments.
- $\widetilde{L_2}$ *Audio Deepfake*: We consider the top two popular learning-based generation techniques: TTS represented by SV2TTS [24] and VC represented by AutoVC [25].

According to the knowledge about our system, adversaries could take the following actions in the ultrasonic band.

- $\widetilde{H_1}$ *Non-treatment*: Adversaries merely tamper with the audible segments and ignore or remove the ultrasonic band.
- $\widetilde{H_2}$ *Re-modulation*: We consider re-modulation attacks using two advanced demodulators: a square-law demodulator [30] and the three wave mixing (TWM) [37].
- $\widetilde{H_3}$ *Replay*: Adversaries record and replay the speeches.
- $\widetilde{H_4}$ *Impostor*: Adversaries imitate the victim's articulatory gestures to obtain the corresponding ultrasonic segments. We consider two cases, in which some of the speeches from the impostor are within or beyond the training dataset.
- $\widetilde{H_5}$ *Generation*: Adversaries directly generate ultrasound by learning-based techniques. Here we use the GAN [26].
- $\widetilde{H_{3+4}}$ *Replay+Impostor*: The adversaries may combine multiple means. Here we consider the most dangerous combined forgery as a representative, i.e., replaying for RNL and imitating for DFS and TOF simultaneously.

Combining the above means, we obtain 12 (=2×6) types of negative samples. We denote $\widetilde{L_iH_j}$ as the tampering method that adversaries conduct. For example, $\widetilde{L_2H_5}$ means that adversaries directly generate the whole audio including both the audible speech and the ultrasound by AI techniques.

Network Implementation. We implement the detection network using PyTorch. The network contains a total of 33.6 M parameters. The batch size is 64. The model is trained in a server with Intel(R) Xeon(R) Silver 4210R CPU@2.40GHz and two Nvidia GeForce RTX 3090. For training the network, we use Adam optimizer with a 1e-04 initial learning rate, dropping by 25% every 5 epochs for a total of 40 epochs. The collected ultrasound-speech data are randomly divided into two parts: 80% for training and 20% for testing.

Metrics. We evaluate the tamper-proof performance of Aegis under three frequently used metrics, i.e., accuracy, precision, and recall. The accuracy is defined as the ratio of correctly-predicted samples to the total samples. The precision and recall are the ratio of correctly-predicted positive samples (i.e., authentic audio segments) to the total positive samples and the total samples to be predicted as positive ones respectively. To describe the capacity of the ultrasonic codes, we use the metrics of bits per second (bps) and bit error rate (BER).

B. Overall Performance

1) Effectiveness against forgery: We first evaluate the effectiveness of Aegis defending against various audio forgery methods. To demonstrate the threat from the attack means in Sec. VII-A, we mix the modified audio (i.e., the negative samples) with the original ones. The other 20 volunteers are asked to distinguish these audios with a human-predicting accuracy of below 40%. On the contrary, Aegis presents better non-repudiation in detecting forgery. As shown in Fig. 9, the average accuracy reaches up to 99.5% on average.

In the attack of $\widetilde{H_4}$ *Impostor*, Aegis can accurately detect modified audio, no matter whether the impostors are included in the training dataset. Among the total 12 impostors, the accuracy is 99.82% against the six seen impostors and it is 99.78% against the six unseen ones. Aegis is able to work effectively without any knowledge about the adversary.

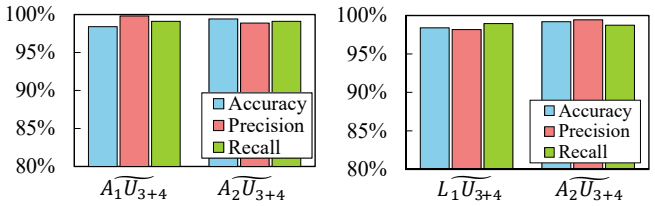
In the attack of $\widetilde{H_5}$ *Generation*, Aegis can resist generative attacks with an accuracy of 99.81%. Although advanced attacks can generate two-dimensional images of the mouth movement according to speeches [48], they can hardly obtain the corresponding TOF/DFS characteristics. Overviews, the TOF/DFS-based authentication systems [36], [40] would be vulnerable. Besides, the parameters (the acoustic intensity and frequency, which would determine the TOF/DFS characteristics) are time-varying in our proposed ultrasonic signals. This measure increases the difficulty and cost of GAN-based forgery in data collection and training.

Among these tampering methods, audio signals are most vulnerable to $L_i\widetilde{H_{3+4}}$. Nevertheless, Aegis still maintains an accuracy of 98.9%. Few modified segments could bypass Aegis with a high recall of 99.1%, while almost all authentic audios are under effective protection with a high precision of 99.3%, as shown in Fig. 11(a). These results indicate that Aegis resists various forgery techniques.

2) Throughput of ultrasonic codes: As shown in Fig. 10, the capacity of the ultrasonic codes reaches high up to 1 kbps. BERs maintains below 1%, which is acceptable in watermark [5]–[7]. The accuracy always exceeds 98%. In our experiments, we transmit the ultrasonic codes at 200 bps by default. The above results imply that Aegis could work under different requirements of capability and error rates.

TABLE I
PERFORMANCE OF AEGIS AGAINST VARIOUS UNSEEN ATTACK METHODS IN AUDIO DEEPFAKE

Accuracy (Train \ Test)	SV2TTS [24]	AutoVC [25]	[49]	[50]
No Generated	99.0%	99.2%	99.0%	99.6%
SV2TTS [24]	99.2%	99.2%	99.0%	99.6%
AutoVC [25]	99.0%	99.6%	98.8%	99.4%
SV2TTS+AutoVC	99.2%	99.4%	99.4%	99.6%



(a) On known words (b) On new words

Fig. 11. Performance comparison between known and new words.

3) Ablation study: In an ablation study, the accuracy is 92.8% if we leverage a network trained on the samples merely with DFS/TOF features. The accuracy will be 96.8% if on the samples only with RNL features (except the cases of defending against replay attacks because the RNL is vulnerable with a low accuracy of 28.8%). Moreover, the accuracy drops to merely 47.1% if we use a ResNet18 [45] trained on speech-only signals without ultrasound. In the ablation of network design, the average accuracy drops to 92.8%, demonstrating the necessity of the parallel training mechanism. The above results show the effectiveness of the Aegis design.

C. Proactivity against New Attacks

We evaluate the proactivity of Aegis, that is, the potential defending effectiveness against unseen and future attacks.

Various Audio Deepfake techniques keep emerging. Here, we conduct two other Audio Deepfake (with open-source codes or datasets) [49], [50]. The results are listed in Tab. I. Obviously, Aegis keeps a high detection accuracy even under the attack of unseen Audio Deepfake techniques. In particular, we train a model on a dataset excluding any AI-generated speeches (i.e., trained merely on $\widetilde{L_1}$ Copy-move). Even under such unfavorable conditions, Aegis still holds the ability of a proactive defense and keeps its accuracy as 99.2% (See the second row in Tab. I). On the other hand, if the negative samples in the training datasets consist of merely under $\widetilde{L_2}$, Aegis still exhibits a similar performance of 99.2% against the forgery from $\widetilde{L_1}$ Copy-move. In comparison, the traditional anti Audio Deepfake method [51] obtains 24.1%∼96.6% against unseen attacks. Hence, Aegis is able to resist unseen or future attacks proactively.

D. Robustness Evaluation

1) Text-free Performance: Aegis is independent of words and texts. We select 20% words as the testing set and the rest as the training set. As shown in Fig. 11(b), the accuracy maintains 98.9% on new words against the most dangerous tampering method $L_i\widetilde{H_{3+4}}$. The performance on new words degrades very slightly compared to the results in Fig. 11(a), with a drop of below 1%.

Furthermore, we consider the generalization of Aegis to other languages. Fortunately, the relationship between voices and articulatory gestures is similar across different languages [35], [43]. Thus, the DFS and TOF are independent of languages. Moreover, the RNL effect is language-free [17]. Thus, Aegis can be applied to any language if being trained on the corresponding datasets. An experiment on Chinese Mandarin from three users presents an accuracy of 98.8% against forgery, which verifies its applicability among languages.

2) Impact of Acoustic Sampling Rate: Aegis can support a wide range of sampling rates, including 96kHz, 48 kHz, and 44.1 kHz. Experimentally, Aegis achieves an average accuracy of 99.6%, 99.2%, and 99.5%, respectively. There is no significant difference in terms of detection accuracy among devices with different sampling rates.

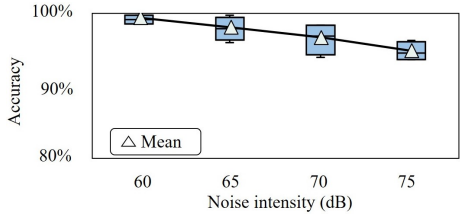


Fig. 12. Ambient noise impact.

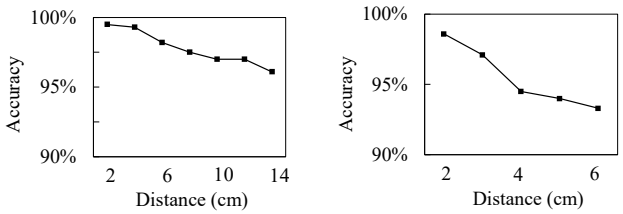
3) Impact of Environmental Noise: We evaluate Aegis in four common scenes, including a seminar room, an office, a crowded cafe, and a street side. The ambient noises measure approximately 56.1 dB, 50.1 dB, 63.9 dB, and 71.2 dB respectively. Aegis is robust with an average accuracy of over 98.8%. Even on the noisiest scene, i.e., the street side, Aegis keeps an accuracy of 96.2%. To further test its anti-noise performance, we manually add noises up to 75 dB. As shown in Fig. 12, Aegis maintains an accuracy of approximately 95% under the noise level of 75 dB, considering that 76 dB is the upper bound of community noise by current regulations [52]. Aegis is able to perform robustly in various realistic scenarios. In the future, we can adopt noise removal techniques (e.g., Wiener filtering [53] and speech enhancement [27], [35]) to enable a higher detection accuracy in extremely noisy environments.

E. Impact of Position

1) Holding Styles: We consider two common holding styles when users speak to smartphones, i.e., the 'Towards Mic'

TABLE II
IMPACT OF HOLDING STYLES.

Accuracy (Train \ Test)	Towards Mic	Phone Call	Mixture
Towards Mic	**99.3%**	71.6%	85.5%
Phone Call	72.9%	**99.3%**	85.4%
Mixture	96.8%	97.5%	97.2



(a) Towards Mic. (b) Phone Call.

Fig. 13. Impact of distance.

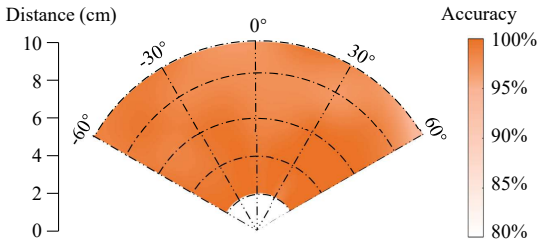


Fig. 14. Impact of angle.

and 'Phone Call' modes. Though it performs poorly when trained on data from merely one set but tested on the data from the other, Aegis maintains a high accuracy of over 97% when trained on both sets as shown in Tab. II. In practice, we can recognize the holding style by leveraging proximity sensors and motion sensors, which are widely embedded in both Android and iOS smartphones, and thus maintain a performance of over 99%.

2) Distance Between Mouth and Device: To be more general, we evaluate Aegis under different mouth-phone distances in the two holding styles. In our training data, the distance from the mouth to the smartphone's microphone ranges from 2 cm to 4 cm. Figure 13 demonstrates its robustness to distance variations. Although the ultrasonic signals will attenuate as the distance increases, the accuracy keeps over 98% at 6 cm, 97% at 10 cm, and 96% at 14 cm in the 'Towards Mic' mode. It also achieves an acceptable sensing distance of 4 cm with an accuracy of 96.4% in the 'Phone Call' mode in which the users' mouths prefer to approach their smartphones yet the microphones' reception directions are not pointing at the mouths. Aegis performs consistently and reliably with effective coverage.

3) Angle of Mouth toward Device: We consider the cases where the users do not always keep the standard 'Towards Mic' position, where the microphones' reception directions do not directly face the mouths, or the smartphones are placed non-horizontally. We vary the angle between the mouth and the smartphone's microphone at each distance, with the results shown in Fig. 14. In an opening angle of $\pm 60°$, Aegis performs robustly with an accuracy of over 93% within 10 cm and of over 98% within 8 cm. In addition, when the smartphone is not horizontal, Aegis maintains a high accuracy of over 97.8%.Therefore, the users can act in their habitual way when operating Aegis with satisfactory protection effectiveness. Note that we still suggest that users approach their smartphones for better protection in practice.

4) Motion Interference: Users' body movements would distort the DFS characteristics and cover up the ones related to speeches. We test Aegis when the users are speaking during walking, running, and driving. Aegis achieves an acceptable accuracy of 98.6%, 97.4%, and 98.7% respectively. The results demonstrate the motion robustness of Aegis.

F. Perceptual Quality Study

We recruit 26 volunteers (13 males and 13 females, aged 18 to 55) to perform a subjective study of the user experience. Following the ITU BS.1387 procedure [54], we ask the volunteers to rate audio under Aegis for the objective difference grade (ODG) on a 5-point scale ranging from -4 to 0 (where a high score refers to little influence on audio). The average ODG is -0.13, where an ODG of over -1 means the perceptual quality degradation is imperceptible [10]. The users report hearing no additional acoustics during use. The results imply that Aegis is imperceptible, barely affecting the audio quality.

VIII. Related Work

Ultrasound has been widely exploited for sensing due to its availability on COTS mobile devices, e.g., smartphones. Ultrasonic schemes are able to reach the sensing resolution of the millimeter scale [32], [33]. Thus, the movement of nearby users' vocal tract [39] can be characterized using ultrasound for lip reading [34], speech enhancement [27], [35], and biometric authentication [36], [40].

Existing ultrasonic sensing methods focus merely on one of TOF/DFS characteristics [36], [40]. Instead, we combine the two features with acoustic reverse nonlinearity (different from common senses [16]–[22], in which ultrasound is converted into audible sound). Exploiting multiple effects presents great potential for speech forgery detection, and we believe reverse nonlinearity will facilitate acoustic applications.

IX. Conclusion

To resist speech forgery, especially in real-time voice applications, we realize a novel audio tamper-proof system, named Aegis. We are the first to probe the acoustic reverse nonlinearity effect that converts audible signals into ultrasonic bands, which has not been explored in the literature. We establish the ultrasound-sound correlation by exploiting multiple acoustic effects. Accordingly, Aegis enables effective detection against various tampering methods.

Acknowledgment

This paper is partially supported by the National Science Fund for Distinguished Young Scholars of China (Grant No. 62125203), National Natural Science Foundation of China (Grant No. U21A20462, No. 62372400), "Pioneer" and "Leading Goose" R&D Program of Zhejiang (Grant No. 2024C03287), Natural Science Foundation of Jiangsu Province of China (Grant No. BK20240615), Natural Science Research Start-up Foundation of Recruiting Talents of Nanjing University of Posts and Telecommunications (Grant No. NY224030), and Ant Group.

REFERENCES

[1] J. H. Engel, K. K. Agrawal, S. Chen, I. Gulrajani, C. Donahue, and A. Roberts, "Gansynth: Adversarial neural audio synthesis," in *ICLR*, 2019.

[2] T. Liu, D. Yan, N. Yan, and G. Chen, "Anti-forensics of fake stereo audio using generative adversarial network," *Multimedia Tools and Applications*, vol. 81, no. 12, pp. 17155–17167, 2022.

[3] Y. Jiang and D. Ye, "Black-box adversarial attacks against audio forensics models," *Security and Communication Networks*, vol. 2022, pp. 6410478:1–6410478:8, 2022.

[4] Y. Mirsky and W. Lee, "The creation and detection of deepfakes," *ACM Computing Surveys*, vol. 54, pp. 1–41, jan 2022.

[5] B. Laurence, T. A. H., and H. K. N., "Digital watermarks for audio signals," in *IEEE ICMCS*, 1996.

[6] I. Natgunanathan, P. Praitheeshan, L. Gao, Y. Xiang, and L. Pan, "Blockchain-based audio watermarking technique for multimedia copyright protection in distribution networks," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 18, no. 3, pp. 86:1–86:23, 2022.

[7] J. Zhao, T. Zong, Y. Xiang, L. Gao, W. Zhou, and G. Beliakov, "Desynchronization attacks resilient watermarking method based on frequency singular value coefficient modification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 2282–2295, 2021.

[8] SENKEN Inc., "Site enforcement recorder." https://www.senkencorp.com/product/Site-enforcement-Recorder.htm, 2016.

[9] Y. Hu, W. Lu, M. Ma, Q. Sun, and J. Wei, "A semi fragile watermarking algorithm based on compressed sensing applied for audio tampering detection and recovery," *Multim. Tools Appl.*, vol. 81, no. 13, pp. 17729–17746, 2022.

[10] V. B. K., I. Sengupta, and A. Das, "An audio watermarking scheme using singular value decomposition and dither-modulation quantization," *Multimedia Tools and Applications*, vol. 52, no. 2-3, pp. 369–383, 2011.

[11] M. Harris, "What is lossless audio compression?." https://www.lifewire.com/what-makes-an-audio-format-lossless-2438560, 2020.

[12] OPUS, "Opus interactive audio codec." https://opus-codec.org/, 2021.

[13] Fraunhofer IIS, "Enhanced voice services." https://www.iis.fraunhofer.de/en/ff/amm/communication/evs.html, 2016.

[14] K. Qian, Y. Lu, Z. Yang, K. Zhang, K. Huang, X. Cai, C. Wu, and Y. Liu, "AIRCODE: hidden screen-camera communication on an invisible and inaudible dual channel," in *USENIX NSDI*, 2021.

[15] X. Yang, X. Wu, and M. Zhang, "Audio digital signature algorithm with tamper detection," in *IEEE IAS*, 2009.

[16] N. Roy, H. Hassanieh, and R. Roy Choudhury, "Backdoor: Making microphones hear inaudible sounds," in *ACM MobiSys*, 2017.

[17] G. Zhang, C. Yan, X. Ji, T. Zhang, T. Zhang, and W. Xu, "Dolphinattack: Inaudible voice commands," in *ACM CCS*, 2017.

[18] Y. He, J. Bian, X. Tong, Z. Qian, W. Zhu, X. Tian, and X. Wang, "Canceling inaudible voice commands against voice control systems," in *ACM MobiCom*, 2019.

[19] K. Sun, C. Chen, and X. Zhang, "'Alexa, stop spying on me!': Speech privacy protection against voice assistants," in *ACM SenSys*, 2020.

[20] L. Li, M. Liu, Y. Yao, F. Dang, Z. Cao, and Y. Liu, "Patronus: Preventing unauthorized speech recordings with support for selective unscrambling," in *ACM SenSys*, 2020.

[21] Y. Chen, H. Li, S.-Y. Teng, S. Nagels, Z. Li, P. Lopes, B. Y. Zhao, and H. Zheng, "Wearable microphone jamming," in *ACM CHI*, 2020.

[22] N. Roy, S. Shen, H. Hassanieh, and R. R. Choudhury, "Inaudible voice commands: The long-range attack and defense," in *USENIX NSDI*, 2018.

[23] W. Zong, Y. Chow, W. Susilo, K. Do, and S. Venkatesh, "Trojanmodel: A practical trojan attack against automatic speech recognition systems," in *IEEE S&P*, 2023.

[24] Y. Jia, Y. Zhang, R. J. Weiss, Q. Wang, J. Shen, F. Ren, Z. Chen, P. Nguyen, R. Pang, I. Lopez-Moreno, and Y. Wu, "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," in *NeurIPS*, 2018.

[25] K. Qian, Y. Zhang, S. Chang, X. Yang, and M. Hasegawa-Johnson, "Autovc: Zero-shot voice style transfer with only autoencoder loss," in *ICML*, 2019.

[26] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio, "Generative adversarial nets," in *NeurIPS*, 2014.

[27] K. Sun and X. Zhang, "Ultrase: single-channel speech enhancement using ultrasound," in *ACM MobiCom*, 2021.

[28] W. Mao, J. He, and L. Qiu, "CAT: high-precision acoustic motion tracking," in *ACM MobiCom*, 2016.

[29] V. C. Chen, F. Li, s. s. Ho, and H. Wechsler, "Micro-doppler effect in radar: phenomenon, model, and simulation study," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 42, no. 1, pp. 2–21, 2006.

[30] D. Chen and J. Laneman, "Modulation and demodulation for cooperative diversity in wireless systems," *IEEE Transactions on Wireless Communications*, vol. 5, no. 7, pp. 1785–1794, 2006.

[31] D. M. Etter and S. D. Stearns, "Adaptive estimation of time delays in sampled data systems," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 29, no. 3, pp. 582–587, 1981.

[32] W. Wang, A. X. Liu, and K. Sun, "Device-free gesture tracking using acoustic signals," in *ACM MobiCom*, 2016.

[33] K. Sun, T. Zhao, W. Wang, and L. Xie, "Vskin: Sensing touch gestures on surfaces of mobile devices using acoustic signals," in *ACM MobiCom*, 2018.

[34] J. Tan, C. Nguyen, and X. Wang, "Silenttalk: Lip reading through ultrasonic sensing on mobile phones," in *IEEE INFOCOM*, 2017.

[35] Q. Zhang, D. Wang, R. Zhao, Y. Yu, and J. Shen, "Sensing to hear: Speech enhancement for mobile devices using acoustic signals," *IMWUT/Ubicomp*, vol. 5, no. 3, pp. 137:1–137:30, 2021.

[36] L. Lu, J. Yu, Y. Chen, H. Liu, Y. Zhu, Y. Liu, and M. Li, "Lippass: Lip reading-based user authentication on smartphones leveraging acoustic signals," in *IEEE INFOCOM*, 2018.

[37] C. Sun, C. Chen, J. Wei, and P. Li, "Efficient three-process frequency conversion based on straddling stimulated raman adiabatic passage," *IEEE Photonics Journal*, vol. 6, no. 6, pp. 1–10, 2014.

[38] F. A. Duck, "Medical and non-medical protection standards for ultrasound and infrasound," *Progress in biophysics and molecular biology*, vol. 93, no. 1-3, pp. 176–191, 2007.

[39] L. Zhang, S. Tan, and J. Yang, "Hearing your voice is not enough: An articulatory gesture based liveness detection for voice authentication," in *ACM CCS*, 2017.

[40] L. Lu, J. Yu, Y. Chen, and Y. Wang, "Vocallock: Sensing vocal tract for passphrase-independent user authentication leveraging acoustic signals on smartphones," *IMWUT/Ubicomp*, vol. 4, no. 2, pp. 51:1–51:24, 2020.

[41] J. Kapur, P. Sahoo, and A. Wong, "A new method for gray-level picture thresholding using the entropy of the histogram," *Compute Vision, Graphics, and Image Processing*, vol. 29, pp. 273–285, 1980.

[42] Y. Jiang and F. H. F. Leung, "Source microphone recognition aided by a kernel-based projection method," *IEEE Trans. Inf. Forensics Secur.*, vol. 14, no. 11, pp. 2875–2886, 2019.

[43] A. Gafos, *The articulatory basis of locality in phonology*. Garland, 1996.

[44] R. Liang, L. Zhao, and X. Wei, *Experimental Course on Speech Signal Processing*. China Machine Press, 2018.

[45] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE/CVF CVPR*, 2016.

[46] Lextutor, "Longman communication 3000." https://lextutor.ca/freq/lists_download/longman_3000_list.pdf/, 2022.

[47] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and N. L. Dahlgren, "Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1," *NASA STI/Recon Technical Report N*, vol. 93, pp. 1–85, 1993.

[48] J. Wang, X. Qian, M. Zhang, R. T. Tan, and H. Li, "Seeing what you said: Talking face generation guided by a lip reading expert," in *IEEE/CVF CVPR*, IEEE, 2023.

[49] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Ryan, R. A. Saurous, Y. Agiomyrgiannakis, and Y. Wu, "Natural TTS synthesis by conditioning wavenet on MEL spectrogram predictions," in *IEEE ICASSP*, 2018.

[50] N. Kalchbrenner, E. Elsen, K. Simonyan, S. Noury, N. Casagrande, E. Lockhart, F. Stimberg, A. van den Oord, S. Dieleman, and K. Kavukcuoglu, "Efficient neural audio synthesis," in *ICML*, 2018.

[51] A. Lieto, D. Moro, F. Devoti, C. Parera, V. Lipari, P. Bestagini, and S. Tubaro, ""hello? who am I talking to?" A shallow CNN approach for human vs. bot speech classification," in *IEEE ICASSP*, 2019.

[52] W. LI, Y. XU, C. ZHANG, Y. TIAN, M. LIU, and J. HUANG, "Multi-frequency-ranging positioning algorithm for 5g ofdm communication systems," *Chinese Journal of Electronics*, vol. 32, pp. 773–784, 2023.

[53] L. Gui, W. Yuan, and F. Xiao, "Csi-based passive intrusion detection bound estimation in indoor nlos scenario," *Fundamental Research*, vol. 3, no. 6, pp. 988–996, 2023.

[54] International Telecommunication Union, "Recommendation bs.1387-0." https://www.itu.int/rec/R-REC-BS.1387-0-199812-S/en, 1998.