Hindawi
Wireless Communications and Mobile Computing
Volume 2021, Article ID 9179100, 16 pages
https://doi.org/10.1155/2021/9179100

Research Article

Privacy Threats of Acoustic Covert Communication among Smart Mobile Devices

Li Duan,[1] **Kejia Zhang**,[2,3,4] **Bo Cheng**,[3] **and Bingfei Ren**[3]

[1]*Beijing Key Laboratory of Security and Privacy in Intelligent Transportation, Beijing Jiaotong University, Beijing 100044, China*
[2]*School of Mathematical Science, Heilongjiang University, Harbin 150080, China*
[3]*State Key of Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, 100876, China*
[4]*Cryptology and Cyberspace Security Laboratory of Heilongjiang University, Harbin 150080, China*

Correspondence should be addressed to Kejia Zhang; zhangkejia.bupt@gmail.com and Bingfei Ren; renbingfei0@gmail.com

Received 23 May 2021; Accepted 22 June 2021; Published 5 July 2021

Academic Editor: Weizhi Meng

The emerging, overclocking signal-based acoustic covert communication technique allows smart devices to communicate (without users' consent) utilizing their microphones and speakers in ultrasonic side channels, which offers users imperceptible and convenient personalized services, e.g., cross-device authentication and media tracking. However, microphones and speakers could be maliciously used and pose severe privacy threats to users. In this paper, we propose a novel high-frequency filtering- (HFF-) based protection model, named *UltraFilter*, which protects user privacy by enabling users to selectively filter out high-frequency signals from the metadata received by the device. We also analyze the feasibility of using audio frequencies (i.e., ≤18 kHz) to the acoustic covert communication and carry out the acoustic covert communication system by introducing the auditory masking effect. Experiments show that UltraFilter can prevent users' private information from leaking and reduce system load and that the audio frequencies can pose threats to user privacy.

1. Introduction

With the rapid development of the Internet of Things (IoT) and smart devices, a user may have more than a smart device [1] for communication and entertainment. There is an increasing need for cross-device authentication [2], and one of multiple devices can act as an identity to control the authorization of the other devices. Traditional cross-device authentication uses the network access or Bluetooth function of devices to complete the authorization task. Because of their long transmission distance, it cannot meet the demand of short-range cross-device authentication. Considering a situation that a user left the smart phone (i.e., the device to be authorized) in the office and a user is temporarily away, the user may use a smart watch (worn by the user and representing the user's identity) to complete an authorization process with the smart phone through wireless network or Bluetooth. At this time, attackers in the office can gain access to the user's smart phone.

Because the propagation distance of ultrasonic frequency (i.e., >18 kHz) is short and imperceptible to users, it has gradually become a research hotspot to use ultrasonic frequency for cross-device authentication or media tracking. Taking media tracking for example. India's Silverpush Company [3] conducts advertising push business by providing a software development kit (SDK). Cooperative clients integrate the advertising push function in their own shopping software by using SDK. The SDK has a built-in ultrasonic frequency signal detection module and a data reporting module. When the detection module detects a specific overclocking signal, the data reporting module starts to collect and report users' personal information to Silverpush's servers. In this way, Silverpush can acquire a large amount of private data and track user's personal trajectories.

Using microphones and speakers in Android devices, acoustic covert communication can provide users with personalized and convenient services, such as cross-device authentication [2] and media tracking-based advertisement push [3]. Covert communication refers to communication in which users do not perceive abnormality under normal circumstances [4]. However, a malicious use of microphones and speakers would bring potential privacy threats to users. Take the routing in an anonymous network [5] for an example. The source and destination addresses of a message are encrypted layer-by-layer, and the sender can access network resources anonymously.

By using acoustic covert communication, the anonymization operation in the anonymous network can be carried out the following four steps: (1) the attacker embeds overclocking signals in normal audio and anonymization video files as audio beacons and adds the synthesized audio files into the webpage of the anonymous network; (2) a target user sends the request to access the anonymous network by using a personal computer; (3) the target user plays the synthesized audio when browsing the webpage containing the synthesized audio and video files; (4) the target user does not perceive abnormality in the audio data, if the user carries an application with a decoding function or SDK integrated with a detection and decoding function (such as Silverpush); (5) the user receives the sound signal through the microphone and detects the hidden audio beacon; and (6) after detecting the special signal, the application begins to collect the user's personal information and send it to the attacker through the network. The detailed process is shown in Figure 1.

In order to protect microphones and speakers from malicious uses, the Android platform provides an authority system: Only when an application declares record authority (*RECORD_AUDIO*) in the configuration file and is authorized by the user, it can access the microphone of the device to collect sound, and only by declaring *MODIFY_AUDIO_SETTINGS*, the application can turn on the microphone or turn off the speaker of the device. However, the protection mechanism of microphone or speaker based on authority in Android systems can be easily bypassed by malicious software, such as collusion attack [6]. The devices can be turned to conduct covert sound wave communication, and malicious attacks can be executed without the user's awareness, causing serious privacy and security problems for users.

Existing studies typically focus on synthesizing confrontation samples, e.g., by using deep learning, to attack the Automatic Speech Recognition (ASR) systems of smart devices [7], such as Google Home, Apple's Siri, Amazon Echo, and Microsoft Cortana [8], and develop countermeasures. The existing studies have overlooked that attackers could use the microphone and speaker to achieve acoustic covert communication, compromising users' privacy (e.g., visiting anonymous networks). Despite microphone-based acoustic covert communication is analyzed in [9], yet no design or implementation of countermeasures is presented.

This paper designs and implements a new acoustic high-frequency signal filtering-based security protection mechanism, to address the privacy threats caused by acoustic covert communication attacks to Android devices. In particular, we carry out an in-depth study analysis of related works and reveal that acoustic covert communication between Android devices is primarily based on inaudible high-frequency signals (above 18 kHz). The inaudible high-frequency signals are embedded into audio files (e.g., music and advertisements) to generate a synthetic audio file and delivered by playing the synthetic audio file. Specialized applications can detect and recognize the inaudible high-frequency signals at target devices.

By using high-frequency filtering, our security mechanism erases near-ultrasonic signals that are inaudible to the users. Further, we study the feasibility of using audible frequency for acoustic covert communication and analyze whether acoustic covert communication imperceptible to users in the audible frequency band poses a potential threat to the user's privacy. The contributions of this paper are summarized, as follows.

(i) A new acoustic high-frequency filtering-based security framework, named *UltraFilter*, is proposed to address the privacy leakage issue of acoustic covert communication. Inaudible high-frequency signals (above 18 kHz) that do not affect the user's perception are filtered and suppressed, so as to protect user privacy

(ii) We reveal that acoustic covert communication can be achieved even in the audible spectrum; the acoustic covert communication without user's perception is finished by employing the auditory masking effect model

(iii) We design and implement two prototype systems. One is the security system based on the high-frequency filtering. The other is a prototype system of acoustic communication without user perception based on normal frequency. The functional verification and performance tests are conducted on the Android version 6.0 system (Xiaomi 4 devices)

The rest of this paper is organized, as follows. Section 2 reviews the related work. In Section 3, we elaborate on the proposed security mechanism, which erases near-ultrasonic signals and protects user privacy. In Section 4, we study the feasibility of using audible frequency for acoustic covert communication by introducing the auditory masking effect model. In Section 5, the prototype systems are designed, implemented, and tested. In Section 6, this paper is concluded.

2. Related Work

Existing studies on acoustic covert communication are focused primarily on three aspects: steganography on audio files, sound signals against sample generation for automatic speech recognition (ASR), and acoustic covert communication among smart devices based on inaudible high-frequency signals.

2.1. Steganography on Audio Files. Steganography [10] refers to the technical methods of hiding information in a harmless format. The communication parties attach the hidden information to normal carriers (such as text files) in a preagreed way and generate seemingly normal camouflage carriers

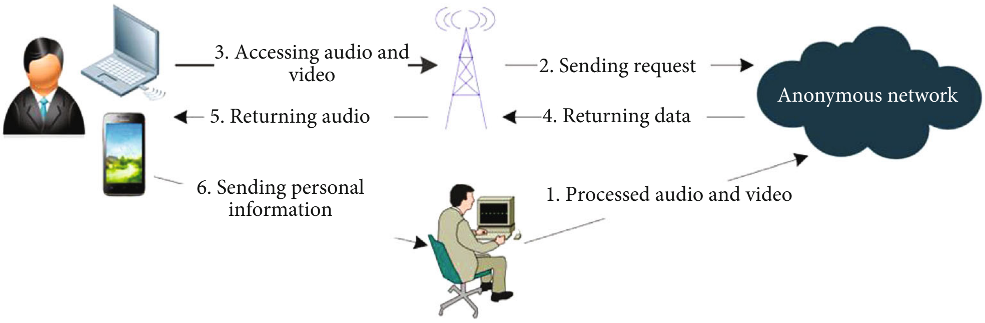

3. Accessing audio and video
2. Sending request
Anonymous network
5. Returning audio
4. Returning data
6. Sending personal information
1. Processed audio and video

Figure 1: De-anonymization of anonymous network users using acoustic covert communication.

and spread it. Only the communication parties can accurately detect and analyze the hidden information. The larger the data file is, the harder it is to find the hidden information. After the concept of modern steganography [11, 12] was put forward in 1985. With the popularity of mobile Internet and the wide application of digital media, steganography based on digital media has developed rapidly. Digital media, such as video, pictures, and audio, contain a large amount of data and widely spread through the Internet. Therefore, steganography using digital media (e.g., audio files) for hiding information and dissemination has attracted wide attention [13–16].

The main challenge of steganography is to deceive the Human Auditory System (HAS) [17]. In the application of audio-based steganography, information can be hidden in three ways: temporal domain, frequency domain, and code domain. Each approach has a corresponding technology to encode hidden information into the carrier audio without damaging it (that is, users cannot perceive it). For example, least significant bits (LSB) can be used to hide the hidden information in the temporal domain of audio [18]. Because no obvious noise is introduced into the audio file, the whole information transmission process is imperceptible to the user.

It is reported in [19] that Discrete Wavelet Transform (DWT) can transform sound signals in the temporal domain to/from the frequency domain and obtain the corresponding wavelet correlation coefficient. By writing the hidden information into the LSB of the correlation coefficient, the covert transmission of hidden information can be accomplished. However, steganography has high requirements for the receiver and the transmission channels. It is difficult for the receiver to extract hidden information accurately. LSB-based audio steganography has high transmission performance (i.e., more hidden information) and is relatively easy to implement. However, it has low anti-interference ability, and the hidden information data in the LSBs can be destroyed. A method is to add a small amount of noises to audio files or losslessly compress the files [20].

2.2. Sound Signal Countermeasure Sample Generation for Automatic Speech Recognition (ASR). In recent years, ASR has made remarkable progress. Its main working mode is to make the machine recognize and understand the speech signal and convert the speech signal into texts or commands [21]. ASR-based voice assistance is increasingly dominating the human-computer interaction, such as Google Assistant, Apple Siri, and Amazon voice assistant Alexa [8]. Voice assistants use voice classification models to detect voice commands, such as playing music, adding alarm clocks, making phone calls, inquiring weather, and controlling other smart devices in smart homes. The voice assistants also use the microphone of the devices to monitor the ambient sound continuously, so as to receive and recognize the voice commands quickly and provide timely services for users.

These automatic voice assistants are exposed to the risks of being maliciously controlled. Authors of [7] proposed that users' voice commands can be converted into ultrasonic frequencies by using ultrasonic devices. These overclocking voice commands, which are imperceptible to the user, can be used to control the speech recognition assistant. Compared with traditional speech recognition models based on hidden Markov chain, a deep learning-based ASR model generated by neural networks has greatly improved the recognition accuracy. However, neural networks are used by attackers to generate wrong targets or confrontation samples [22], so as to bypass the recognition of deep learning models or produce the results that attackers want from the models. Some researchers [23, 24] made use of this weakness of neural networks to generate confrontation samples of user voice commands to attack current ASR systems. The speech command countermeasure sample generation framework proposed in [25] can convert speech commands to any desired speech countermeasure samples. By processing a speech with the content of "without the dataset the article is useless", a corresponding confrontation sample can be generated and recognized by the ASR system as *"okay google browse to evil dot com"* [26]. The speech of this sample does not change to the user. At present, this kind of acoustic covert communication is mainly used to attack ASR systems. Because the attack target is clear, users can take precautions in advance.

2.3. High-Frequency Signal-Based Acoustic Covert Communication. With the rapid development of IoT technology and mobile intelligent devices, users are faced with the problem of continuous identity authentication for multiple mobile smart devices. Usually, it is necessary to use one of the devices as an identity to control the authorization of all other devices, such as cross-device authentication.

Traditional cross-device authentication technology often uses the network access or Bluetooth communication function in the devices to complete the authorization task. However, due to their long transmission distance, they cannot meet the requirements of short-distance cross-device authentication. In contrast, the high-frequency sound signal (above 18 kHz) has a short propagation distance and cannot be perceived by users. It has been increasingly considered for cross-device authentication [27, 28].

Yi et al. [29] proposed WakeLock, a smart phone security unlocking system based on acoustic communication. When the user unlocks its smart phone, the smart phone sends out a sound wave signal to verify the user's identity through the speaker of the device. The user receives the signal through the smart watch (which can be used as the user's identifier) and authorizes the request. Then, the mobile phone is unlocked for the user to use. When a stranger obtains the right to use the user's smart phone, the smart phone does not receive the authorization from the user's smart watch after sending the authorization request. The mobile phone remains locked to protect the user's device data from malicious access.

Mavroudis et al. [9] found that acoustic covert communication based on the microphones of smart phones can be used in media tracking. Shopkick [30] is location-based shopping software in the Android platform. When the user approaches a cooperative merchant, advertisement encoded into ultrasonic frequency signals is played at the door of the clients' shops. The ultrasonic signals that the user cannot perceive in the advertisement audio are detected by the Shopkick program in the user's mobile phone. Then, the merchant receives the notification of the user's arrival and sends a voucher to the user through the Shopkick application.

Covert communication based on inaudible high-frequency sound signals can provide users with convenient personalized services, such as identity authorization and shopping. However, it can also bring privacy threats to the users, such as the de-anonymization of anonymous network users. This paper focuses on how to mitigate the privacy threats imposed by the acoustic covert communication based on high-frequency signals and designs and implements the corresponding security model. In addition, the feasibility of using audible frequency for acoustic covert communication is studied, which can analyze whether the imperceptible acoustic covert communication in the audible frequency range can threat the users' privacy.

3. Analysis of Acoustic Covert Communication Based on High-Frequency Signal

This section first briefly summarizes some basic concepts of acoustic communication and then studies the characteristics of the temporal and frequency domains of synthesized sounds (normal audio carriers and high-frequency special signals) in acoustic covert communication. According to the characteristics, a new security model is proposed, designed, and implemented to address the privacy issue arising from the acoustic covert communication, e.g., de-anonymization in anonymous networks.

Sound is a wave phenomenon which is generated by vibration, transmitted through medium (e.g., air, liquid, and solid), and then perceived by human or animal auditory organs. Sound sources with different vibration frequencies produce sounds with different pitches. The distance between the sound source and the receiver also directly affects the loudness felt by the receiver. In the Human Auditory System (HAS) [17], sound waves can be divided into the following three categories, according to the frequency:

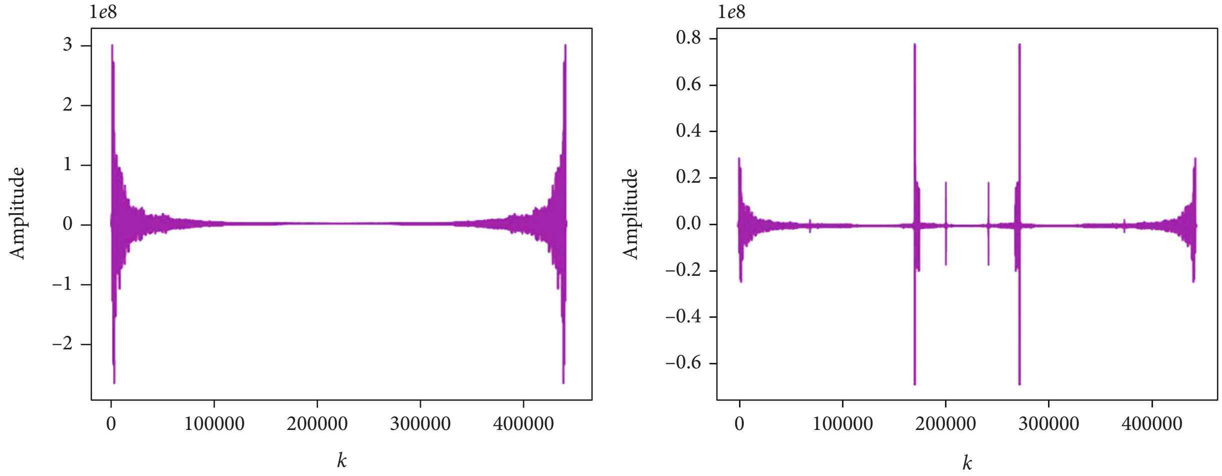
(1) *Infrasound*. Sound with a frequency lower than 20 Hz cannot be perceived by the human ears, and it is difficult to use infrasound to realize communication functions because the frequency is too low

(2) *Audible Sound Waves*. Sound waves with frequencies between 20 Hz and 20 kHz can be perceived by the auditory system of human ears. The range of human hearing is between 20 Hz and 20 kHz

(3) *Ultrasound*. sound wave, of which the frequency is higher than the upper limit of human hearing (i.e., 20 kHz), is called ultrasonic. Compared with infrasonic waves, ultrasonic waves have shorter wavelengths. Standard mobile devices can generate ultrasonic waves, and therefore, ultrasonic waves are suitable for short-distance acoustic communication

Due to the limitation of hardware configuration and human ear hearing system of smart mobile devices (such as smart phones and smart watches), only the frequency bandwidth from 18 kHz to 20 kHz can be used. Therefore, in order to realize acoustic covert communication without user's perception, FSK modulation technology is mainly used to modulate hidden information in acoustic covert communication. At the same time, in order to increase the transmission distance of high-frequency signals, the energy of high-frequency signals is set to a relatively high value (within this frequency bandwidth, users cannot perceive sound). As a result, the normal carrier audio (i.e., synthesized audio) synthesized with high-frequency signals is quite different from ordinary normal audio files in temporal domain, frequency domain, and spectrum characteristics. The normal carrier audio (synthesized audio) synthesized with high-frequency signals is quite different from ordinary normal audio files in temporal domain, frequency domain, and spectrum characteristics, as shown in Figures 2–4.

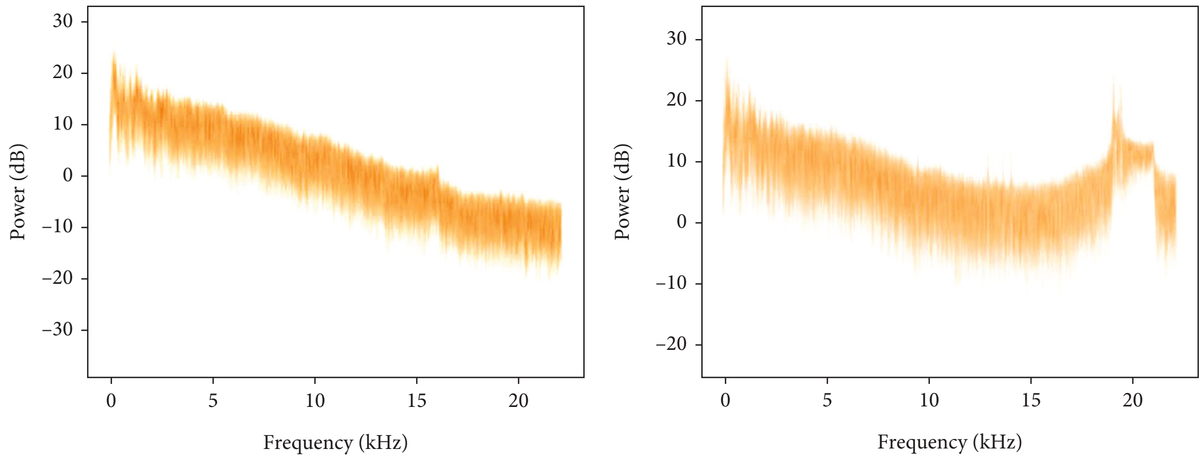
Because the high-frequency signal (18 kHz–20 kHz) is beyond the hearing range of normal people, the existing work makes use of this characteristic to realize acoustic covert communication based on high-frequency signal. By enhancing the energy of high-frequency signals, the long-distance transmission of high-frequency signals can be realized without causing users' doubts. Normal audio (such as music, advertisement, and conversation sounds) is in a frequency greater than 18 kHz, and its frequency energy is extremely low. However, the audio synthesized with high-frequency signal shows a strong energy distribution in the range of 18 kHz to 20 kHz (as shown in Figures 3 and 4). It is a great challenge to detect specific high-frequency signals in the sound



(a) Time domain of normal audio

(b) Time domain of synthetic audio

Figure 2: Comparison of time domain characteristics between synthetic audio and normal audio.



(a) Frequency domain of normal audio

(b) Frequency domain of synthetic audio

Figure 3: Comparison of frequency domain characteristics between synthesized audio and normal audio.

signals received by Android devices to identify potential acoustic covert communication. For example, it is difficult to master the modulation and demodulation scheme agreed by both parties and the specific high-frequency signal. Therefore, the corresponding security model can be designed and implemented in Android system by using the important characteristics of normal audio and synthesized audio, so as to enable users to protect personal privacy in a specific environment.

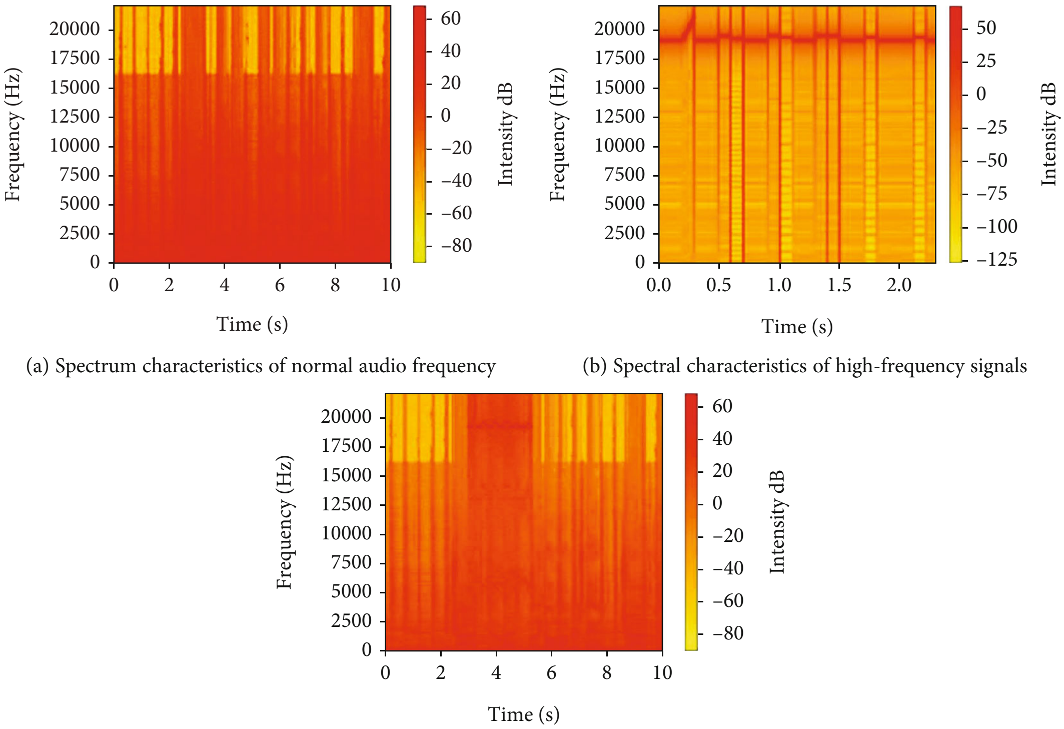
4. Security Model Based on High-Frequency Filtering of Sound Waves

4.1. Model Design. In a specific context, acoustic covert communication based on high-frequency signals, which is caused by the malicious use of equipment microphone and other resources, will bring potential privacy threats to users. Therefore, according to the analysis of the existing research work of acoustic covert communication based on high-frequency signals in the previous section, this section proposes a security model UltraFilter based on acoustic high-frequency filtering, which enable users to avoid privacy threats caused by high-frequency signal communication by controlling and protecting the Android system to obtain sound signals.

As shown in the work flow of security model in Figure 5, the upper part of the figure describes the work flow of Android device acquiring sound signal and Android system processing signal metadata under normal circumstances; the detailed description is as follows:

(1) *Receiving the Signal*. The normal audio synthesized with special high-frequency signals is played out through the speaker of the device, and then, the user samples the external sound signals through the microphone in the intelligent device

(2) *Sound Metadata*. Android system stores the sound signals sampled by microphone for subsequent processing



(a) Spectrum characteristics of normal audio frequency

(b) Spectral characteristics of high-frequency signals

(c) Spectral characteristics of synthesized audio

Figure 4: Comparison of spectrum characteristics between synthetic audio and normal audio.

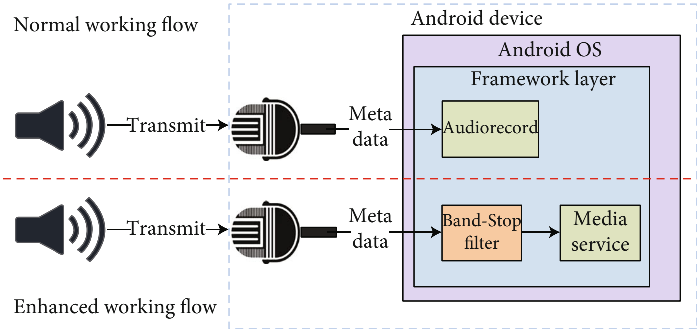


Figure 5: Security framework based on acoustic high-frequency filtering.

(3) *High-Frequency Signal*. Particular applications in Android platform (such as Silverpush and Shopkick) call the *AudioRecord* interface provided by the media component in the system framework layer to obtain sound signal metadata; the applications directly call the *MediaRecorder* and *AudioRecorder* interface, but both of them need to use *AudioRecord* object to obtain sound data to detect and extract special high-frequency signals in them, and then perform corresponding functions, such as issuing bonus coupons to stimulate store consumption, pushing advertisements in a targeted manner, or anonymizing anonymous network users

By adding a band-stop filter (BSF) before the *AudioRecord* reads the original sound signal, the user can control and filter out the high-frequency signals in the metadata, so as to ensure that personal privacy information is not leaked when accessing an anonymous network.

4.2. Model Implementation. This section introduces the technical details of the proposed UltraFilter, a security model based on high-frequency filtering of sound waves, which consists of the implementation of the BSF and the integration of the BSF into the Android system.

The purpose of the filter is to filter the signal in a specific frequency or frequency range. The commonly used filters [31] are Low-Pass Filter, High-Pass Filter, Band-Pass Filter, and band-stop filter (BSF). BSF is used in this paper. The key function of BSF is to attenuate the frequencies substantially in a specific range, so as to cut off the frequencies within a certain range in the signal data.

In this paper, a butterworth filter [32] is used to block the high-frequency part of the acoustic signal (from 18 kHz to 20 kHz). Band-stop filters based on various programming languages have similar structures. The following is our prototype.

$$\text{butter\_filter}(\text{lowcut}, \text{highcut}, \text{fs}, \text{order} = \text{order}), \quad (1)$$

where lowcut and hightcut are the upper and lower boundaries of the frequency range to be blocked, respectively, fs is the sampling frequency of the sound signal, order is the order of the filter, and the higher the order is, the faster the frequency attenuates in the specified range. For example, for a synthetic signal with frequencies of 600 Hz, 1200 Hz, and 1600 Hz is filtered with the band-stop filter. The noise data with frequency of 2 Hz is added to the synthetic signal for more obvious effect. The frequency to be suppressed ranges from 1000 Hz to 2000 Hz; i.e., the values of lowcut and highcut are 1000 and 2000, respectively. If the value of fs is consistent with the sampling frequency of the synthetic signal, the signal after passing through the band-stop filter should be a synthetic signal with a frequency of 600 Hz and the noise. The results shown in Figure 6 verify this reasoning and prove the effectiveness of the band-stop filter. Figure 7 shows the frequency attenuation characteristics in the range of 1000 Hz to 2000 Hz when order takes different values.

Because this paper needs to integrate the band-stop filter into the Android system, the third-party implementation library based on JAVA language [33] is the first choice in the implementation. The third-party library is integrated into the Android system, which is convenient for subsequent calling in the Android framework layer to realize the band-stop filter function. The integration steps are as follows:

(1) Create a custom folder (usually named after the library) in the *external* folder in the Android source directory and a new folder named *src* in the newly created directory

(2) Put the downloaded third-party library file or its source code file into the *src* directory created in the first step

(3) Open the *Androdi.mk* file under the path *frameworks/-base*, and add a new line "../../external/<your – dir > /src" after the definition of the variable ext_dirs

(4) Recompile Android source code or directly execute the "mmmframeworks/base" command to complete the integration of the third-party libraries

After integrating the third-party library into the Android system, the interface and function provided by the library file can be called directly in the Android framework code. In this paper, it is necessary to filter the metadata of sound signals at high frequency after the microphone obtains the metadata, to prevent the upper application from using the metadata through *MediaRecorder* or *AudioRecorder* interface. Through the detailed analysis of Android source code, it is found that both the *MediaRecorder* and *AudioRecorder* interfaces need to acquire sound signal data through the *AudioRecorder* interface [34] in the framework layer. Therefore, by customizing the interface, users can configure corresponding policies according to their needs, to dynamically control the high-frequency filtering of audio signal metadata.

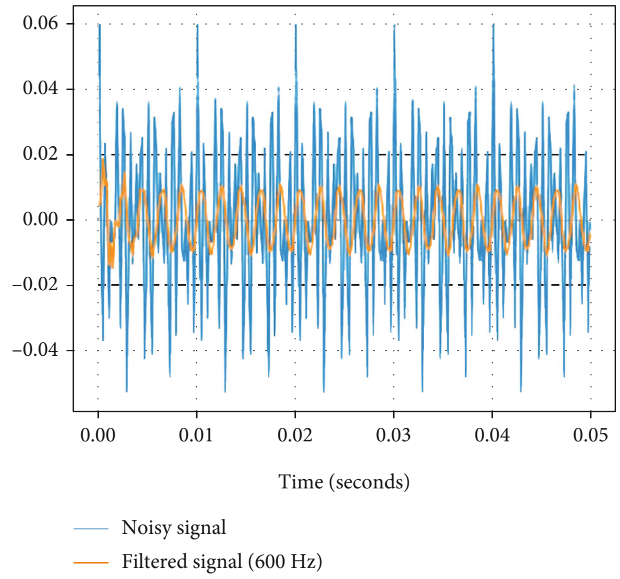


Figure 6: Working effect of band-stop filter.

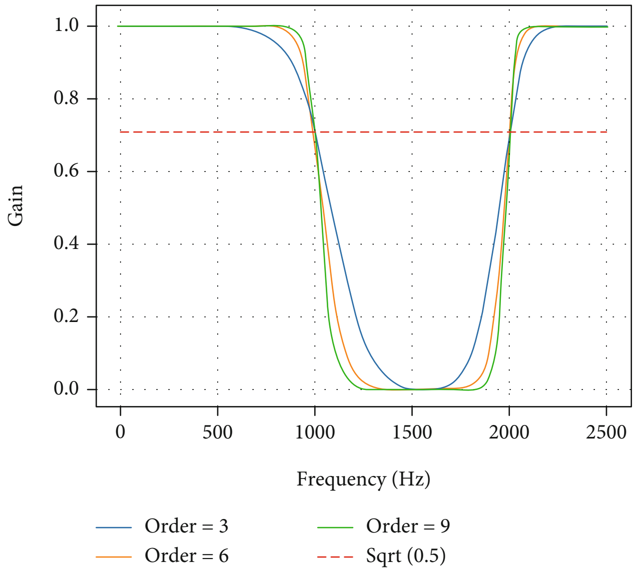


Figure 7: Influence of band-stop filters of different order on frequency attenuation speed.

5. Acoustic Covert Communication Based on Normal Frequency

In this section, we study the feasibility of the acoustic covert communication based on normal frequency and reveal that imperceptible acoustic covert communication in the normal frequency band can pose threats to users' privacy. We further put forward countermeasures to mitigate the threats. The prototype system of acoustic covert communication using the normal frequency faces four challenges: how to determine the available normal frequency bandwidth for information modulation, which carrier modulation scheme to be selected for information modulation, how to select the appropriate insertion point in normal audio to synthesize with special signals, and how to eliminate the sound and other noises of normal frequency signals so as not to arouse users' doubts.

5.1. Available Normal Frequency Bandwidth. Normal frequency is used as the carrier frequency for information modulation. Other factors, such as human ear perception, equipment capability, and environmental noise, are considered. The core functional organ of the human auditory perception system [35] is the human ear. Because of the special structure of the human ear, human beings have different sensitivity to sound signals at different frequencies. The typical audible frequency range of human beings is between 20 Hz and 20 kHz. Most people's auditory systems are insensitive to frequencies above 18 kHz. They are the most sensitive to sounds in the frequency range from 300 Hz to 4 kHz, which is also the frequency range of human speech sounds.

Extremely low signal energy can be perceived by human beings. The response frequency of microphones and speakers in smart mobile devices to sound signals is below 20 kHz. In addition, the frequencies of various noises are generally below 9 kHz. In this sense, it is reasonable to select carrier signals above 9 kHz to reduce the interference of environmental noises to carrier signals based on normal frequencies. According to the above factors, within the normal frequency range that humans can hear, the frequency suitable for acoustic covert communication ranges from 9 kHz to 18 kHz.

5.2. Selection of Modulation Technology. When sound wave is used as carrier to transmit information, the first problem to be considered is how to encode information (i.e., baseband signal); that is, information is modulated into the sound signal. Given the commonality between acoustic wave and electromagnetic wave, the modulation and demodulation techniques in electromagnetic wave communications can be applied to acoustic wave communications. In this paper, the FSK modulation is selected to implement the prototype system of the normal frequency-based acoustic covert communication.

5.3. Selecting the Insertion Point of the Carrier Signal. To realize acoustic covert communication using normal audio, it is necessary to modulate the covert information to generate signals at the carrier frequency. The covert information is synthesized with normal audio to reduce the possibility of the covert signals arousing the users' suspicions. We assume that the length of the signals is much shorter than the duration of normal audio. In order to find a suitable insertion point in the normal audio to synthesize the signals, the amplitude of the normal audio is analyzed to find the part with the strongest energy to place the signal. As shown in Figure 8, energy analysis is performed on the normal audio to simplify the implementation. The Python library pydub [36] is used to obtain the energy value of sound signals, and the part of the audio with the largest average energy is found. The covert signals are inserted from the start of the part of the audio.

5.4. Elimination of Sound and Other Noise of Special Signal. Because the carrier frequency of the covert signals is within the audible normal frequency range, it can produce audible noises. According to the masking effect of sound [37], we further process the signals to reduce the audible noises. The masking effect of sound [37] refers to the phenomenon that when masking sound and masked sound are played at the same time, the hearing threshold of masked sound increases due to the existence of the masking sound, and users can only perceive the existence of the masking sound. There are two typical types of masking techniques: frequency-domain masking and temporal-domain masking.

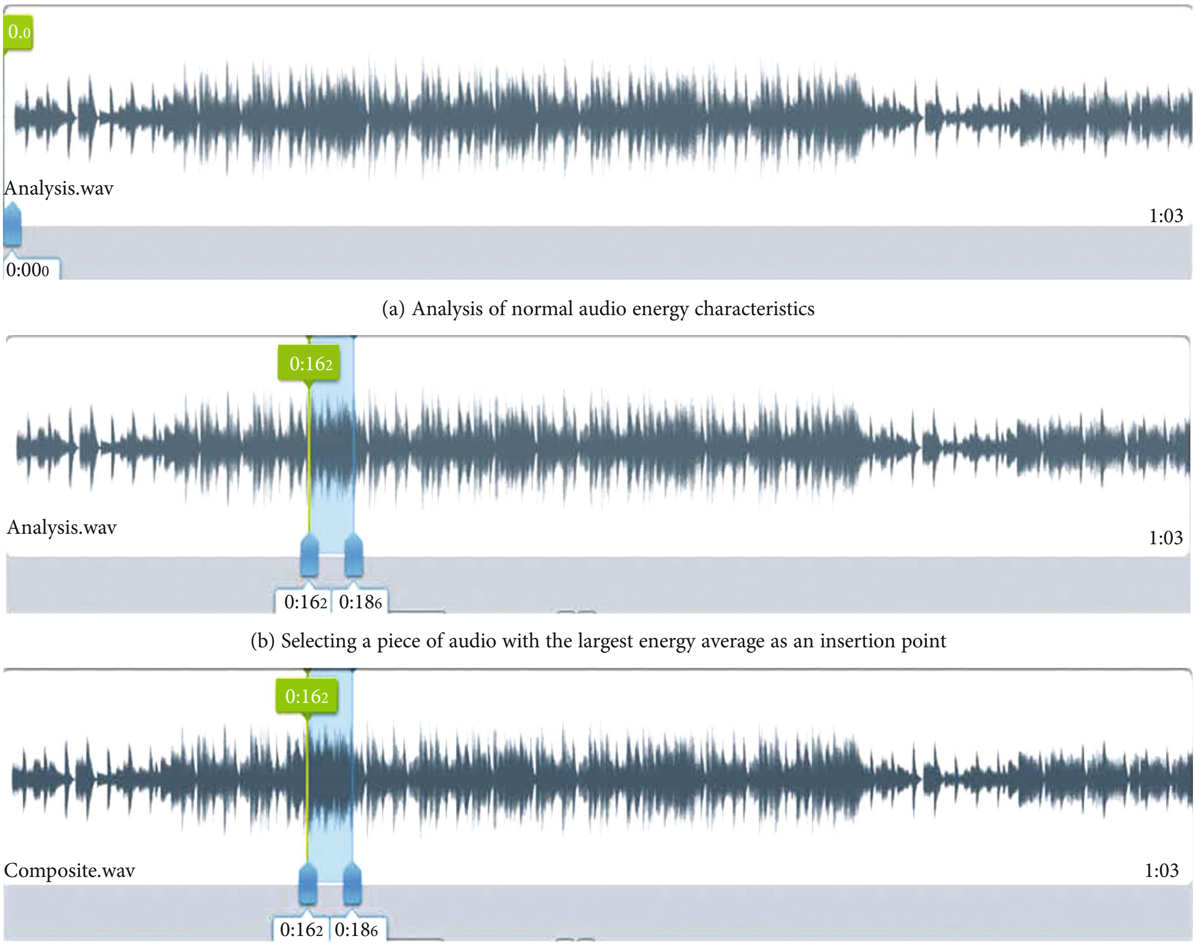
Frequency-domain masking enables sounds with different frequencies to mask each other. The basic rules are that low frequency signals mask high-frequency signals and strong sound masks weak sound. It is difficult to accurately calculate how much the energy of the masked sound is lower than that of the masking sound. When there is a big difference between two frequencies, it is impossible to mask completely. Some researchers put forward a calculation model [38], which roughly estimates the critical value of the masked sound (i.e., the covert signals), provided the knowledge of the energy intensity of the masked sound (i.e., the normal audio). In this paper, the critical value is evaluated and used as the decibel number of the masked sound.

Temporal-domain masking is constituted of premasking and postmasking, according to the time sequence of masking sound and masked sound. In premasking, masking sound is emitted first, followed by the masked sound (within 200 ms). In postmasking, the masking sound is emitted after the masked sound (within 50 ms). The masking effect is achieved within the corresponding time interval. Because the insertion point of the covert signals is within the normal audio frequency, temporal-domain masking can reduce the noises perceptible to the users.

When the carrier frequency is synthesized with normal audio frequency, the sharp change of frequency causes the raised tip in the sound wave or the discontinuity of the sound wave in a short time. This can make the noise easily perceivable to the users. In order to eliminate this influence, the signal at the insertion point can be faded in and out by using Sigmoid function [39]; that is, the signal strength of the normal audio is zero for a short time, and the strength of the covert signal increases from 0 to the normal value, as shown in Figure 9. In this way, the noise caused by the convex tip in the sound wave or the discontinuity of the sound wave in a short time can be eliminated.

6. Implementation of Prototype System

Based on the above analysis, we design a prototype system of acoustic covert communication based on normal audio

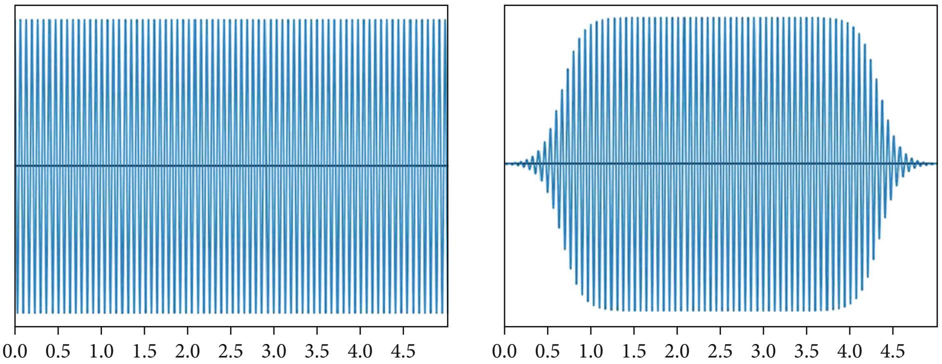


(a) Analysis of normal audio energy characteristics

(b) Selecting a piece of audio with the largest energy average as an insertion point

(c) A special signal is synthesized at the insertion point

Figure 8: Analysis of normal audio energy characteristics and selection of insertion points for special signal synthesis.



(a) Normal signal

(b) Signal after Sigmoid Function Smoothing Processing

Figure 9: Smooth processing of sound signal to eliminate noise.

frequency, in which frequencies are selected from 9 kHz to 18 kHz. The system architecture is shown in Figure 10.

The prototype system is developed for Android version 6.0 system, and the selected frequency range is from 12 kHz to 14 kHz. Because the main purpose of this paper is to verify the feasibility of normal audio-based acoustic covert communication, the system bit rate and other performance are not specially treated. 16-FSK modulation is used to modulate information. Each symbol has a duration of 0.1 s to transmit 80 bits. The synchronization signal is a chirp signal from the start frequency to the end frequency and has a length of 2 symbols. The signal synchronization between the sender and receiver is accomplished by using matched filters [40, 41].

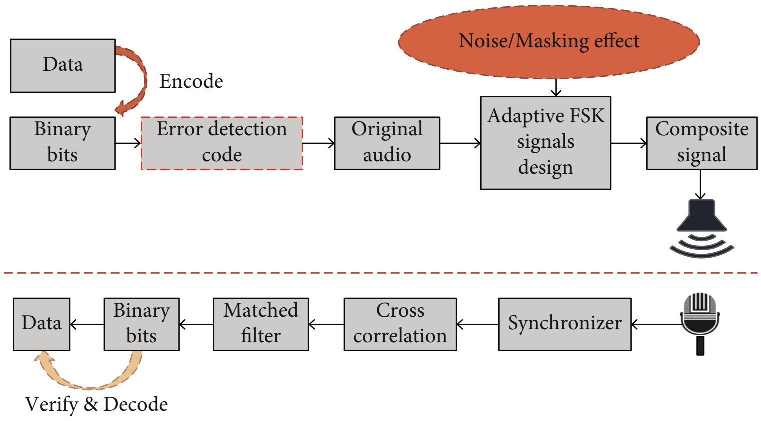
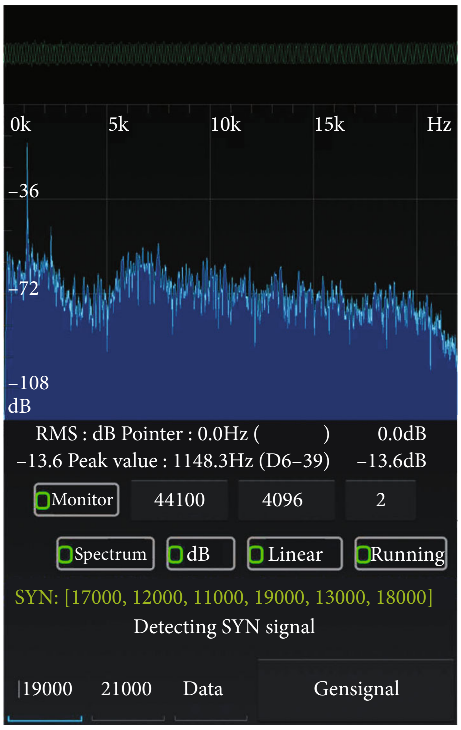


Figure 10: Prototype system architecture.



```
RMS : dB Pointer : 0.0Hz (          )     0.0dB
−13.6 Peak value : 1148.3Hz (D6−39)   −13.6dB
SYN: [17000, 12000, 11000, 19000, 13000, 18000]
Detecting SYN signal
```

Figure 11: A prototype system of acoustic covert communication based on normal audio frequency.

Our prototype of normal frequency-based acoustic covert communication is shown in Figure 11. The overall framework of the prototype program is based on open source projects (https://github.com/bewantbe/audio-analyzer-for-android.).

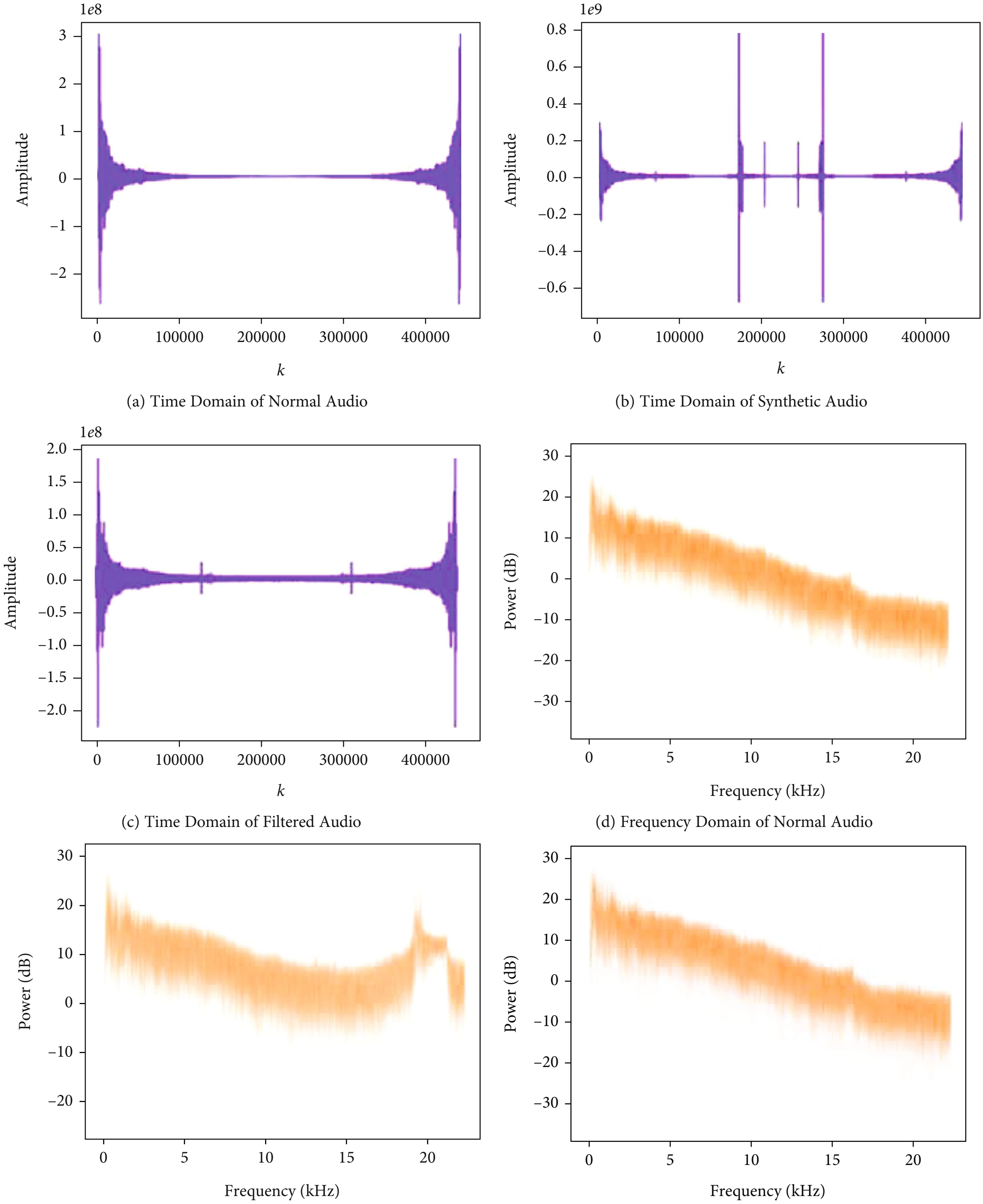
7. Function Test and Performance Analysis

In this section, we first design experiments to test the function and performance of the security model based on high-frequency filtering and the prototype system of normal audio-based acoustic covert communication. Then, the experimental results are discussed. Finally, we analyze the impact of this work on the execution performance of mobile smart phones.

7.1. Security Model Test with Acoustic High-Frequency Filtering. In order to increase the reliability of the experimental results, we first integrate the prototype system of security model based on high-frequency filtering into Android 6.0 version system. Then, we transplant the customized Android system (ROM) to Xiaomi 4 mobile phone for testing. The operation steps and related documents of Android source code compilation, ROM customization, and ROM transplantation to new devices involved in this paper were obtained from the relevant URLs [42].

7.1.1. Function Testing. The proposed security model based on acoustic high-frequency filtering is aimed at a specific context (i.e., visiting anonymous networks). Attackers use high-frequency signals above 18 kHz, which cannot be perceived by users and hidden in normal audio, to conduct covert acoustic communication and obtain users' private information (de-anonymization). To test the prototype system, we randomly select a music file and insert high-frequency covert signals and test the prototype system by observing the temporal domain, frequency domain, and spectrum characteristics of sound signals in three objects received by an Android device and filtered by the Android system 14. The three objects are the original file, the synthesized file, and the audio file. As shown in Figure 12, the prototype system successfully blocks the propagation of high-frequency signals (greater than 18 kHz) and effectively eliminates the temporal-domain, frequency-domain, and spectrum characteristics of the covert signals.

7.1.2. Performance Test. Because the security model based on high-frequency filtering is implemented in the framework layer of Android systems, the new band-stop filter filtering the metadata of received audio signals is consistent with the design purpose. As compared with the workflow of sound signal processing in the traditional Android system, the additional step of filtering metadata at high frequency can bring time delay for applications to use microphones. In order to test whether the delay has any impact on the operation of the applications, a recording application is issued separately, and the time consumed by the application to obtain sound signals through microphone in normal Android system and



(a) Time Domain of Normal Audio

(b) Time Domain of Synthetic Audio

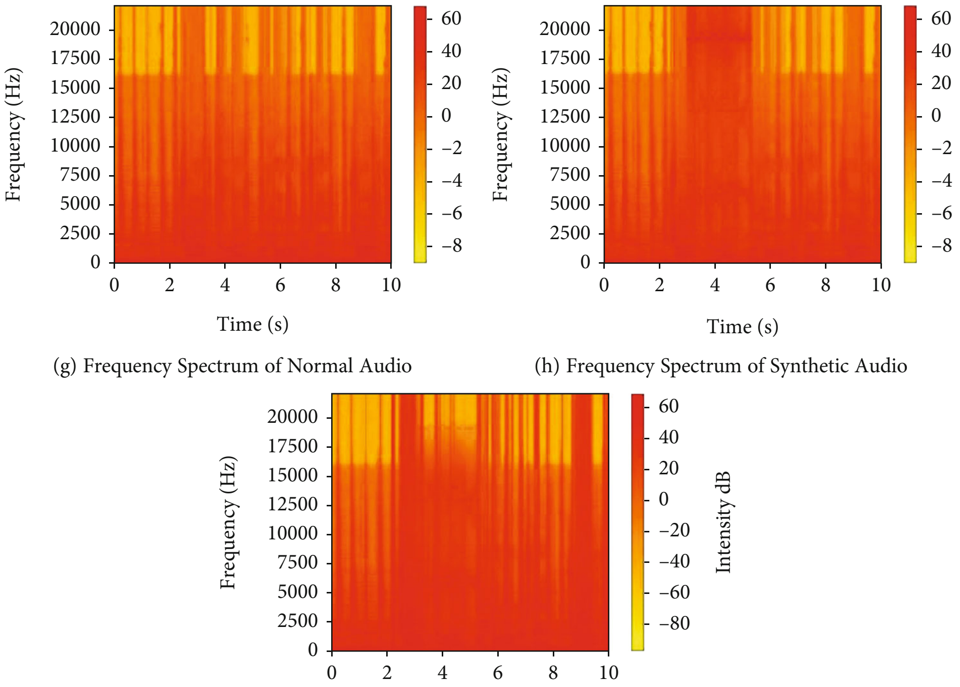
(c) Time Domain of Filtered Audio

(d) Frequency Domain of Normal Audio

(e) Frequency Domain of Synthetic Audio

(f) Frequency Domain of Filtered Audio

Figure 12: Continued.



(g) Frequency Spectrum of Normal Audio

(h) Frequency Spectrum of Synthetic Audio

(i) Frequency Spectrum of Filtered Audio

Figure 12: Functional test of security model based on high-frequency filtering.

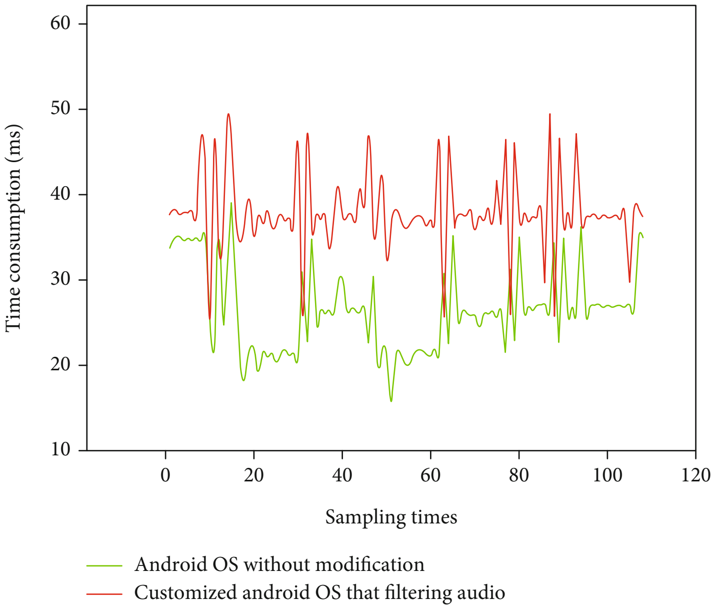


Figure 13: Performance test of the security model based on high-frequency filtering.

customized Android system (with high-frequency filtering function) is recorded by way of log. A comparison study is provided in Figure 13, which shows that the prototype system of the proposed security model only adds less than 20 ms delay, and meets the typical use requirements of protecting users' personal information from malicious theft.

7.2. Prototype System of Normal Frequency-Based Acoustic Covert Communication. This section tests the function and performance of normal frequency-based acoustic covert communication proposed in this paper. The receiver is tested comprehensively, and the sender is responsible for playing out the synthesized audio. The key functions and performances tested include the detection accuracy, the detection time/delay, and the evaluation of whether the user can perceive the covert signals in the audio.

7.2.1. Testing of Detection Accuracy. Two groups of comparative experiments are designed to test the detection accuracy of the prototype system in different environments. We fix the sound source (i.e., the equipment playing synthetic audio, such as PC) at one place and adjust the distance of the smart phone with the prototype application installed to the sound source: 0 m, 1 m, 2 m, 3 m, 4 m, and 5 m, so as to test the actual detection performance of the prototype system in various scenes at different distances. The prototype application is installed in Xiaomi 4 mobile phones, and the sound playing decibel value of the sound source is 70 dB.

The experimental results are shown in Figure 14. When the prototype system is in a noisy environment, due to the interference of environmental noise, the actual performance of receiving and detecting the covert signals decreases, and the detection accuracy also degrades with the increasing distance between the receiver and sound source. However, when the distance of the sound source and receiver is within 3 meters, the accuracy can still reach about 90%. The feasibility of using normal frequency as the carrier frequency to modulate covert information is verified.

7.2.2. Time Load Test. In the process of detecting covert signals in environmental sounds, a Xiaomi 4 mobile phone is used. There are three functional modules that need to perform calculations and introduce delays in the prototype application installed in the smart phone, namely, synchronous signal detection, original special signal recovery, and demodulation of covert information from covert signals. When the prototype application executes the detection task, the time consumed by each functional module is recorded in the form of log. The average time consumption of each functional module is shown in Table 1.

In Table 1, we can see that the prototype system consumes less than 1 millisecond when detecting the synchronization signal (chirp signal from the start frequency to the end frequency and with a length of 2 symbols) of the received sound signals. The reason is because the sender and receiver negotiated the format of the synchronization signal in advance, and the receiver only needs to apply a matched filter to the metadata of the received sound signal to complete the synchronization operation. The time consumption is short. In contrast, the recovery and demodulation functions take much longer, but they are still within the acceptable range (i.e., less than 2 seconds).

7.2.3. Concealment Satisfaction of Covert Communication. The core function of acoustic covert communication is to realize covert communication without arousing users' suspicion. In this paper, normal frequency is used as the carrier frequency to modulate the covert information. Because the frequency of the modulated signals is within the auditory range of the human ears, users can notice the signals. In order to reduce the perceivability of the covert signals and other noises at the users, the masking effect of sound is used to fade in and out the covert signals and normal audio insertion point signals.

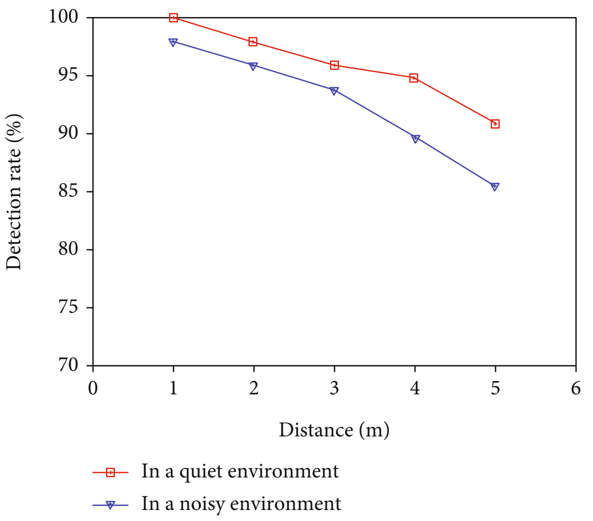


FIGURE 14: Testing accuracy of the acoustic covert communication based on normal frequency.

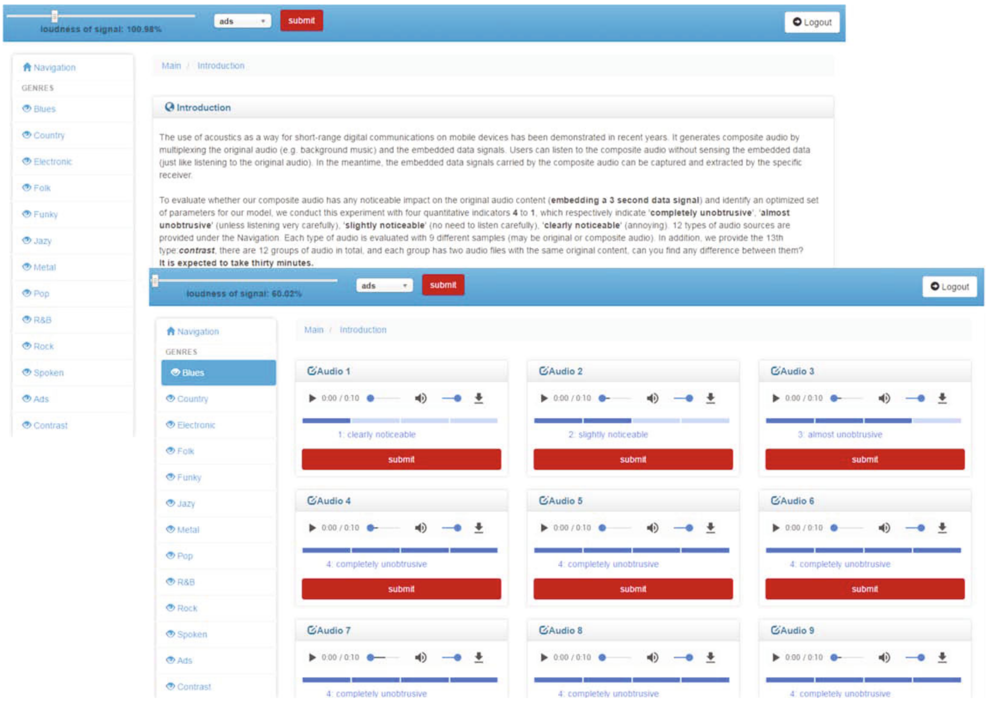
Fifty volunteers, aged between 20 and 50, participated in our satisfaction survey experiment. The participants (or users) were randomly divided into two groups A and B, with 25 people in each group. Those in group A were unaware of the existence of covert signals in the played audio files. Those in group B were informed that the audio files could be synthesized with covert signals before the audio files were played to them. In order to help the volunteers evaluate the quality of synthesized audio, four scores were set in this experiment: (1) 4 points, if no abnormality can be perceived; (2) 3 points, if it is difficult to perceive abnormal sounds unless extra attention is paid; (3) 2 points, if abnormality is perceived in the synthesized audio file without paying extra attention, but does not arouse suspicion; and (4) 1 point, if there are obvious abnormal noises arousing the user's suspicion. In addition, the experiment selects three types of audio as normal audio, i.e., light music (blues, country and folk, etc.), heavy metal music (electronic, rock, and metal), and human voice (advertisement and conversation), to assess whether different types of normal audio have impact on the perception of covert signals in the synthesized audio. These music files are free resources published on the Internet [43].

To obtain accurate feedback from the participants, we develop a satisfaction survey platform, as shown in Figure 15. By randomly generating and arranging normal audio and synthesized audio in each group of music, the interaction between participants is reduced. For example, when a participant P1 enters the blues music interface for the first time, the music labeled "Audio1" may be normal audio, but when P1 enters this page again or other participants enter the blues page, "Audio1" may be a synthesized audio. According to the experimental results in Figure 16, heavy metal music has higher high-frequency signal energy, and the covert sound signals can be well masked.

The experimental results of group B show that the users with a priori knowledge were particularly sensitive to

Table 1: Time load test results of prototype system.

Functional module	Synchronous signal detection	Special signal recovery	Demodulate special signals
Time consumption (milliseconds)	0.81	1063	431



The use of acoustics as a way for short-range digital communications on mobile devices has been demonstrated in recent years. It generates composite audio by multiplexing the original audio (e.g. background music) and the embedded data signals. Users can listen to the composite audio without sensing the embedded data (just like listening to the original audio). In the meantime, the embedded data signals carried by the composite audio can be captured and extracted by the specific receiver.

To evaluate whether our composite audio has any noticeable impact on the original audio content (**embedding a 3 second data signal**) and identify an optimized set of parameters for our model, we conduct this experiment with four quantitative indicators **4** to **1**, which respectively indicate '**completely unobtrusive**', '**almost unobtrusive**' (unless listening very carefully), '**slightly noticeable**' (no need to listen carefully), '**clearly noticeable**' (annoying). 12 types of audio sources are provided under the Navigation. Each type of audio is evaluated with 9 different samples (may be original or composite audio). In addition, we provide the 13th type:***contrast***, there are 12 groups of audio in total, and each group has two audio files with the same original content, can you find any difference between them?
It is expected to take thirty minutes.

Audio 1: 1: clearly noticeable
Audio 2: 2: slightly noticeable
Audio 3: 3: almost unobtrusive
Audio 4: 4: completely unobtrusive
Audio 5: 4: completely unobtrusive
Audio 6: 4: completely unobtrusive
Audio 7: 4: completely unobtrusive
Audio 8: 4: completely unobtrusive
Audio 9: 4: completely unobtrusive

Figure 15: Covert satisfaction research platform.

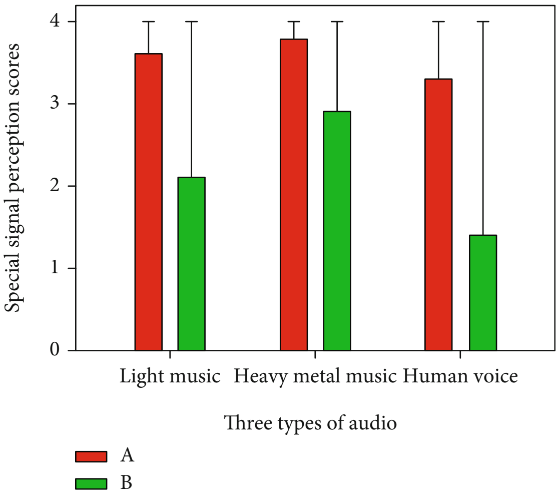


Figure 16: Investigation results of concealment satisfaction.

abnormal noises in the sound. However, if a user was not given the knowledge in advance, the noise caused by the covert signals was not easily perceivable. In most cases, it was considered as a short-term failure of the audio file or the playback device.

Based on the above analysis, acoustic covert communication based on normal frequency is feasible, and for ordinary users, it has the same or similar privacy threat as that based on high-frequency signals. Therefore, when users perform anonymous operations in a specific context, the mechanism based on high-frequency filtering can not cope with this attack. It is necessary to set up additional resource usage policies to control the application programs using microphones and speakers.

8. Conclusions

Applications conducting covert communication based on high-frequency sound waves threat users' privacy. We proposed a new security mechanism which uses high-frequency filtering to erase inaudible near-ultrasonic covert signals. We revealed that acoustic covert communication imperceptible to users in the normal frequency band is also threat. Two prototype systems were developed for the new security model and the normal-frequency acoustic covert communication. Their functions and performance were experimentally evaluated.

Despite the new security model can address the current privacy threats, our new study indicates that the normal frequency-based acoustic covert communication cannot be addressed by the model or other existing techniques. Our future work will focus on how to effectively detect the covert signals in audible normal frequency bands.

Data Availability

Previously reported data were used to support this study and are available at \url{https://github.com/bewantbe/audio-analyzer-for-android.}}.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work is supported by the National Natural Science Foundation of China under Grant No. 61902021, the Beijing Natural Science Foundation under Grant 4212008, the Open Foundation of State key Laboratory of Networking and Switching Technology (Beijing University of Posts and Telecommunications) under Grant SKLNST-2019-2-08, the National Natural Science Foundation of China under Grant No. 61802118, and the Natural Science Foundation of Heilongjiang Province under Grant No.YQ2020F013.

References

[1] Sophos, "Users weighed down by multiple gadgets and mobile devices [EB/OL]," 2013, https://www.sophos.com/en-us/press-office/pressreleases/2013/03/mobile-security-survey.aspx.

[2] P. Samangouei, V. M. Patel, and R. Chellappa, "Attribute-based continuous user authentication on mobile devices," in *Proceedings of the 7th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pp. 1–8, Arlington, VA, USA, 2015.

[3] Silverpush, "Artificial intelligence powered context detection marketing [EB/OL]," 2018, https://www.silverpush.co/.

[4] Z. Liu, J. Liu, Y. Zeng, and J. Ma, "Covert wireless communication in IoT network: from AWGN channel to THz Band," *IEEE Internet of Things Journal*, vol. 7, no. 4, pp. 3378–3388, 2020.

[5] D. McCoy, K. Bauer, D. Grunwald, T. Kohno, and D. Sicker, "Shining light in dark places: understanding the Tor network," in *Privacy Enhancing Technologies*, pp. 63–76, Springer, Berlin, Heidelberg, 2008.

[6] C. Marforio, A. Francillon, and S. Capkun, *Application Collusion Attack on the Permission-Based Security Model and its Implications for Modern Smartphone Systems*, ETH Zurich, 2011.

[7] G. Zhang, C. Yan, X. Ji, T. Zhang, T. Zhang, and W. Xu, "DolphinAttack: inaudible voice commands," in *Proceedings of the 2017 ACM SIGSAC conference on computer and communications security (CCS '17)*, pp. 103–117, New York, NY, USA, 2017.

[8] Forbes, "Voice assistants: this is what the future of technology looks like [EB/OL]," 2017, https://www.forbes.com/sites/herbertrsim/2017/11/01/voice-assistants-this-is-what-the-future-of-technology-looks-like/#7a2e4546523a.

[9] V. Mavroudis, S. Hao, Y. Fratantonio, F. Maggi, C. Kruegel, and G. Vigna, "On the privacy and security of the ultrasound ecosystem," *Proceedings on Privacy Enhancing Technologies*, vol. 2017, no. 2, pp. 95–112, 2017.

[10] N. F. Johnson and S. Jajodia, "Exploring steganography: seeing the unseen," *Computer*, vol. 31, no. 2, pp. 26–34, 1998.

[11] D. Kahn, "The history of steganography," *Proceedings of the International Workshop on Information Hiding*, , pp. 1–5, Springer, Berlin, Heidelberg, 1996.

[12] J. C. Judge, *Steganography: Past, Present, Future*, Lawrence Livermore National Lab., CA (US), 2001.

[13] K. Gopalan, "Audio steganography using bit modification," in *Proceedings of the International Conference on Multimedia and Expo. ICME'03. Proceedings (Cat. No. 03TH8698)*, vol. 1, pp. 1–629, Hong Kong, China, 2003.

[14] N. Cvejic and T. Seppanen, "Increasing the capacity of LSB-based audio steganography," in *Proceedings fo the IEEE Workshop on Multimedia Signal Processing*, pp. 336–338, St. Thomas, VI, USA, 2002.

[15] P. Jayaram, H. R. Ranganatha, and H. S. Anupama, "Information hiding using audio steganography - a survey," *The International Journal of Multimedia & Its Applications*, vol. 3, no. 3, pp. 86–96, 2011.

[16] M. Zamani, A. Manaf, R. B. Ahmad, F. Jaryani, H. Taherdoost, and A. M. Zeki, "A secure audio steganography approach," in *Proceedings of the 2009 International Conference for Internet Technology and Secured Transactions,(ICITST)*, pp. 1–6, London, UK, 2009.

[17] C. Alain, S. R. Arnott, S. Hevenor, S. Graham, and C. L. Grady, ""What" and "where" in the human auditory system," *Proceedings of the National Academy of Sciences*, vol. 98, no. 21, pp. 12301–12306, 2001.

[18] A. Chadha and N. Satam, "An efficient method for image and audio steganography using Least Significant Bit (LSB) substitution," 2013, https://arxiv.org/abs/1311.1083.

[19] N. Cvejic and T. Seppanen, "A wavelet domain LSB insertion algorithm for high capacity audio steganography," in *Proceedings of 10th Digital Signal Processing Workshop, 2002 and the 2nd Signal Processing Education Workshop*, pp. 53–55, Pine Mountain, GA, USA, 2002.

[20] F. Djebbar, B. Ayad, K. A. Meraim, and H. Hamam, "Comparative study of digital audio steganography techniques," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2012, no. 1, 2012.

[21] D. Yu and L. Deng, *Automatic Speech Recognition*, Springer london limited, 2016.

[22] S. M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "Deepfool: a simple and accurate method to fool deep neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2574–2582, Las Vegas, NV, USA, 2016.

[23] N. Carlini, P. Mishra, T. Vaidya et al., "Hidden voice commands," in *Proceedings of the 25th USENIX Security Symposium (USENIX Security 16)*, pp. 513–530, Austin, TX, USA, 2016.

[24] X. Lei, G.-H. Tu, A. X. Liu, K. Ali, C.-Y. Li, and T. Xie, "The insecurity of home digital voice assistants-Amazon Alexa as a case study," 2017, https://arxiv.org/abs/1712.03327.

[25] L. Schönherr, K. Kohls, S. Zeiler, T. Holz, and D. Kolossa, "Adversarial attacks against automatic speech recognition systems via psychoacoustic hiding," 2018, https://arxiv.org/abs/1808.05665.

[26] N. Carlini, "Audio adversarial examples [EB/OL]," 2018, https://nicholas.carlini.com/code/audio_adversarial_examples.

[27] N. Z. Gong, A. Ozen, Y. Wu et al., “Piano: proximity-based user authentication on voice-powered internet-of-things devices,” in *Proceedings of the 2017 IEEE 37th International Conference on Distributed Computing Systems (ICDCS)*, pp. 2212–2219, Atlanta, GA, USA, 2017.

[28] B. Zhang, Q. Zhan, S. Chen et al., “PriWhisper: enabling keyless secure acoustic communication for smartphones,” *IEEE Internet of Things Journal*, vol. 1, no. 1, pp. 33–45, 2014.

[29] S. Yi, Z. Qin, N. Carter, and Q. Li, “WearLock: unlocking your phone via acoustics using smartwatch,” in *Proceedings of the 2017 IEEE 37th International Conference on Distributed Computing Systems (ICDCS)*, pp. 469–479, Atlanta, GA, USA, 2017.

[30] G. Play, “Shopkick: shopping, rewards and deals [EB/OL],” 2019, https://play.google.com/store/apps/details?id=com.shopkick.app&hl=en_US.

[31] M. K. Mandal and P. Mondal, “Design of sharp-rejection, compact, wideband bandstop filters,” *IET Microwaves, Antennas and Propagation*, vol. 2, no. 4, pp. 389–393, 2008.

[32] I. W. Selesnick and C. S. Burrus, “Generalized digital Butterworth filter design,” *IEEE Transactions on Signal Processing*, vol. 46, no. 6, pp. 1688–1694, 1998.

[33] Source-Code, “A collection of java classes for digital signal processing [EB/OL],” 2018, http://www.source-code.biz/dsp/java/.

[34] AndroidXRef, “Android source code cross reference: AndroidRecord [EB/OL],” 2018, http://androidxref.com/9.0.0_r3/xref/frameworks/base/media/java/android/media/AudioRecord.java#native_read_in_short_array.

[35] E. Zwicker and U. T. Zwicker, “Audio engineering and psychoacoustics: matching signals to the final receiver, the human auditory system,” *Journal of the Audio Engineering Society*, vol. 39, no. 3, pp. 115–126, 1991.

[36] “pydub [OL],” https://pypi.org/project/pydub/.

[37] Y. C. Tung and K. G. Shin, “Exploiting sound masking for audio privacy in smartphones,” in *Proceedings of the 2019 ACM Asia Conference on Computer and Communications Security*, pp. 257–268, New York, NY, USA, 2019.

[38] “Decibels calculation [OL],” http://hyperphysics.phy-astr.gsu.edu/hbase/Sound/db.html.

[39] “Sigmoid function [OL],” https://mathworld.wolfram.com/SigmoidFunction.html.

[40] K. Suzuki, “Digital matched filter,” 1999, US Patent 5–903595.

[41] C. E. Wheatley III and J. E. Maloney, “Method and apparatus for pilot search using a matched filter,” 2004, US Patent 6–760366.

[42] “Build lineageos ROM [OL],” 2017, https://www.lineageosrom.com/2017/01/how-to-build-lineageos-rom-for-any.html.

[43] “Free music archive [OL],” http://freemusicarchive.org/.