SHANGHAI JIAO TONG UNIVERSITY

Inaudible Voice Attack Cancellation

Course Report for Wireless Communication Technology

Zihui Qian

May 28, 2018

1 INTRODUCTION

Voice controllable systems(VCS) allow people to control the smart device by voice command with speech recognition(SR) technologies. It has been an increasingly popular human-computer interaction mechanism because of its accessibility, efficiency, and recent advances in recognition accuracy. However, as the rapid progress in the application of VCS, the security problem of it should also be considered. Backdoor[2] showed how hardware non-linearities in microphones can be exploited, such that *inaudible ultrasound signals* can become audible to any microphones. DolphinAttack[3] developed on Backdoor to demonstrate that no software is needed at the microphone, i.e., a voice enabled device like Amazon Echo can be made to respond to inaudible voice commands. These attacks are becoming increasingly relevant, particularly with the proliferation of voice enabled devices including Amazon Echo, Google Home, Apple Home Pod, Samsung Bixby, etc.

This report aims at exploiting the ability of ultrasonic attack and proposing an practical solution against this kind of attack. We propose a practical defense method and design the protecting system for actual situation.

The rest of the report is organized as follows. Section 2 discuss the background. Section 3 state our attack cancellation scheme and section 4 evaluates our model and the report concludes with section 5. My personal work is interpreted in section 6.

2 BACKGROUND

2.1 NON-LINEARITY IN MICROPHONE SYSTEM

In audio recording system, microphone is indispensible part. They are in general case designed to be linear systems, meaning that the output signals are linear combination of the

input. In the case of power amplifiers(PA) inside the microphones, if the input sound signal is $s(t)$, then the output should ideally be:

$$s_{\text{out}}(t) = A_1 s(t)$$

where A_1 is the amplifier gain. In practice, however, acoustic components in microphones are linear only in the audible frequency range(<20 kHz). In Ultrasound bands(>25 kHz), the responses exhibit non-linearity [1]. Thus, for ultrasound signals, the output of the PA becomes:

$$\begin{aligned} s_{\text{out}}(t) &= \sum_{i=1}^{\infty} A_i s^i(t) = A_1 s(t) + A_2 s^2(t) + A_3 s^3(t) + ... \\ &\approx A_1 s(t) + A_2 s^2(t) \end{aligned} \tag{2.1}$$

Higher order terms are typically extremely weak since $A_{4+} \ll A_3 \ll A_2$ and hence can be ignored.
As above shows the non-linearity can be modeled as a simple power series polynomial. To validate this non-linearity, we inject $m(t)$ which contain a voice command to a ultrasonic signal and then constructed $s(t)$ input to microphone. Mathematically,

$$s(t) = (m(t) + \alpha) \times \sin(w_c t)$$

After passing through the PA of the microphone, the output $s_{\text{out}}(t)$ can be calculated as follow:

$$\begin{aligned} s_{\text{out}}(t) &= A_1 s(t) + A_2 s^2(t) \\ &= \frac{A_2}{2}\alpha^2 + A_2 \alpha m(t) + \frac{A_2}{2} m^2(t) - \frac{A_2}{2}\alpha^2 \cos(2 w_c t) \\ &- A_2 \alpha m(t) \cos(2 w_c t) - \frac{A_2}{2} m^2(t) cos(2 w_c t) \\ &+ A_1 m(t) \sin(w_c t) + A_1 \alpha \sin(w_c t) \end{aligned} \tag{2.2}$$

Filtering by low pass filter(LPF) and anti-DC:

$$s_{\text{out}}(t) = A_2 \alpha m(t) + \frac{A_2}{2} m^2(t)$$

The result is shown in Figure 2.1.
This is essentially a tone which will be recorded and interpreted by the microphones.

3 Attack Cancelation

3.1 Attack Cancellation for existing devices

For the limitation that we cannot modify the existing hardware, we propose to add external tones to modulate the attack signal to lower frequency baseband. With carefully design, our cancellation scheme processes in two step, to move attack signal and to cancel attack signal. Figure 3.1 indicates the process of our scheme.

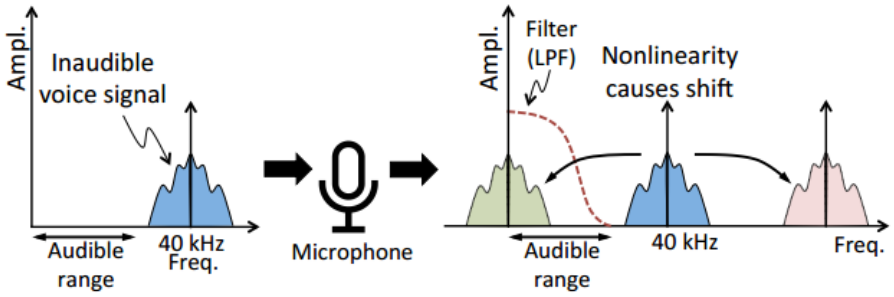


Figure 2.1: Hardware non-linearity creates frequency shift. Voice commands transmitted over inaudible ultrasound frequencies get shifted into the lower audible bands after passing through the non-linear microphone hardware

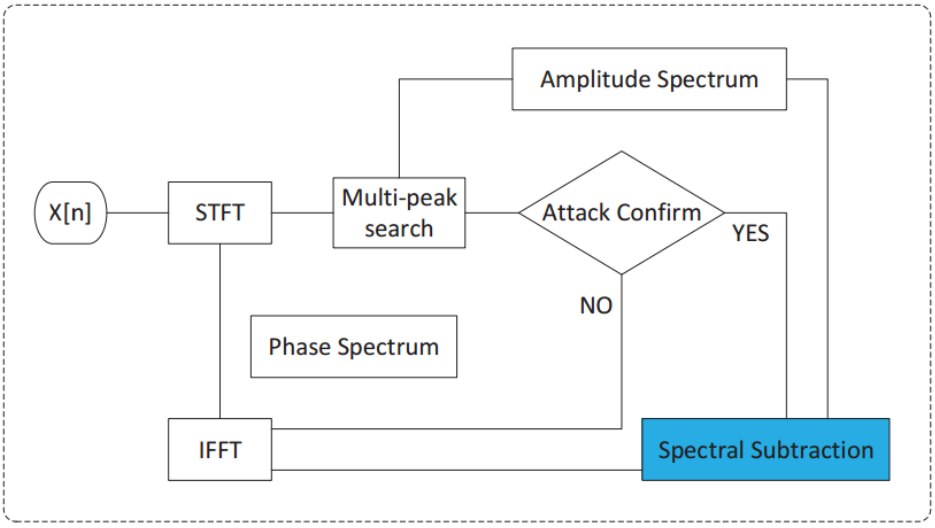


Figure 3.1: The Process of Attack Cancellation Scheme

1. **Multi-tones scheme**: We assign two specifically selected frequencies to deal with the attacking signal ranging from 20 kHz to 40 kHz, which is 23 kHz and 43 kHz. In this way, we can modulated the attacking signal to two lower frequency f_{w1} and f_{w2}. Specifically,

$$f_{w1} + f_{w2} = 20\,\text{kHz}$$

   if the attacking signal is above 40 kHz, we can add more frequencies, such as 64 kHz, to deal with it.

2. **STFT and IFFT**: Because the voice control device is a continuous sensing device. We have to achieve the time efficiency. We consider to compute the FFT of attack signal and modulate it on frequency domain. Using short time Fourier transform(STFT) with windows can achieve a fewer spectrum leakage and be time-efficient. Besides, due to the fact that human ear is insensitive to phase information, we can recover the processed signal by original phase spectrum.

3. **Attack detection**: STFT calculates the amplitude and phase spectrum for each frame in which we determine where the attack signal is. Inside each frame, we can search the peak of spectrum and sum the pair of frequencies in which the peak occurs. If the sum is equal to 20 kHz, we judge it as attack signal. In practice, because of the limitation of spectrum resolution, we can make the criterion less strict, allowing a Δf error when searching the peak.

4. **Spectral substraction**: To apply the spectral substraction in our situation, we can simply and rapidly substract the attack recovered signal by moved signal in frequency domain. Let us suppose $|X_{\text{amp}}(k)|$ denotes the amplitude at kHz, P_h denotes peak in [10 kHz,20 kHz]. Then:

$$|\hat{X_{\text{sub}}}(k)| = \begin{cases} ||X_{\text{amp}}(k)| - a|X_{\text{amp}}(\text{sum} - k)|| & (k < sum/2) \\ ||X_{\text{amp}}(k)| - a|X_{\text{amp}}(P_h + k)|| & (k > sum/2) \end{cases}$$

3.2 Attack Cancellation for New Devices

Recent years, many researchers focus on acoustic application for its advantages such as low power consumption and easy implementation, etc. 20 kHz cannot meet the need of increasing acoustic applications. We are convinced that wider waveband will be used in the future. On the other hand, with the reduction on ADC and DAC, small smart devices are more likely to be equipped with high sampling rate chips. Motivated by these two consideration ,we design an attack cancellation scheme for high-sampling-rate devices, as shown in Figure 3.2.

As the flowchart shows, we first seperate the signal into high frequency part and low frequency part. For the former one, we use Hilbert transform to demodulate and derive the envelope of signal. Then we impose FT on both the envelope and low frequency part. Ultimately, we substract them on amplitude and reconstruct the original signal by IFT.
To calculate the Helbert transform of the signal, we have to clarify the non-linearity model of PA and other non-linear devices. It is achieved by using Minimum Mean Square Error(MMSE).

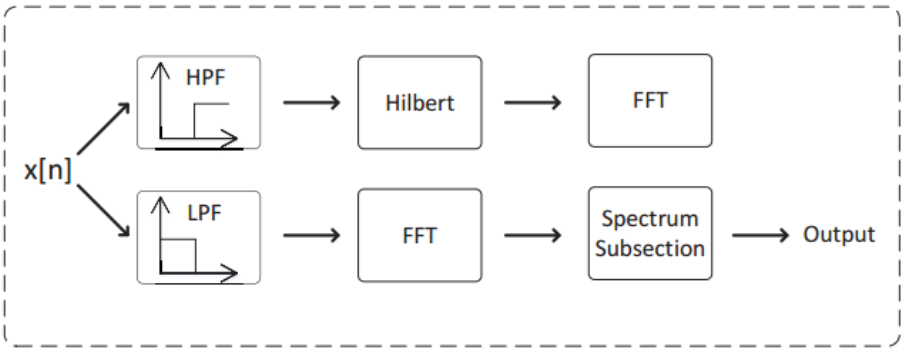


Figure 3.2: Attack Cancellation for New Devices

Step 1: Obtain the testing data vectors of input power and output power.

$$\mathbf{p}_{\text{in}} = [p_{\text{in}}(0) \quad p_{\text{in}}(1) \quad \cdots \quad p_{\text{in}}(L-1)]^T$$

$$\mathbf{p}_{\text{out}} = [p_{\text{out}}(0) \quad p_{\text{out}}(1) \quad \cdots \quad p_{\text{out}}(L-1)]^T$$

Step 2 Defining the polynomial model with coefficients vectors **a**

$$\mathbf{a} = [a_0 \quad a_1 \quad \cdots \quad a_N]^T$$

And the observation matrix **U** of power is:

$$\mathbf{U} = \begin{bmatrix} 1 & p_{in}(0) & p_{in}^2(0) & \cdots & p_{in}^N(0) \\ 1 & p_{in}(1) & p_{in}^2(1) & \cdots & p_{in}^N(1) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & p_{in}(L-1) & p_{in}^2(L-1) & \cdots & p_{in}^N(L-1) \end{bmatrix} \tag{3.1}$$

Then we get polynomial model:

$$g(\mathbf{p}_{\text{in}}) = \mathbf{Ua}$$

The minimize object function:

$$E(\mathbf{a}) = (p_{\text{out}} - \mathbf{Ua})^T (p_{\text{out}} - \mathbf{Ua})$$

Solved above equation by differentiating $E(\mathbf{a})$ with **a** and set it to 0, we can get:

$$\hat{\mathbf{a}} = (\mathbf{U}^T\mathbf{U})^{-1}(\mathbf{U}^T\mathbf{p}_{\text{out}})$$

Thus, we can derive the coefficients of this non-linear model.

4 EVALUATION

Currently the evaluation is moving forwards. The Evaluation parts is divided into three parts.

1. **Simulation on computer**: We have finished this part, Figure 4.1 has shown the simulation result in which we can clearly observe that the attacking signal are filtered out. The result is shown in Figure 4.1.

2. **Practical application**: This part is currently moving forwards. There are several circuit to build. One is to generate the exact frequency as mentioned in section 3.1. Also, there are some recording and signal processing curcuits that require to be built and tested.

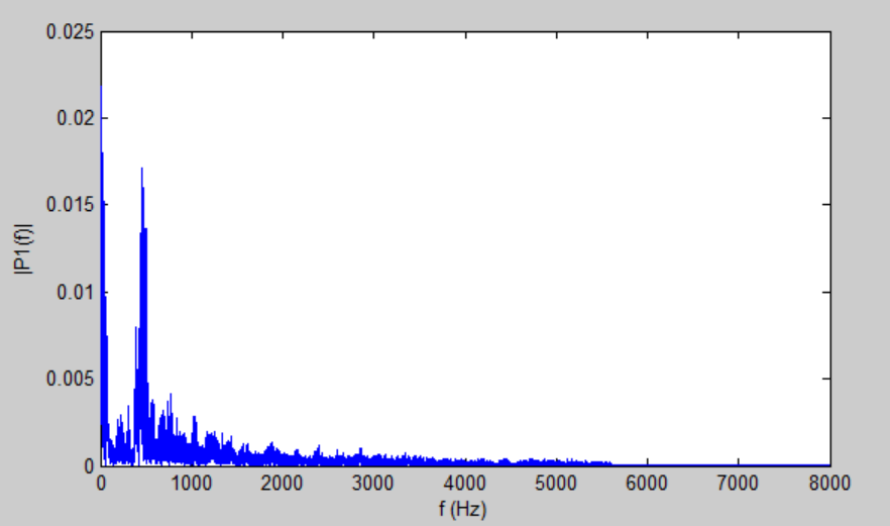
3. **Experiment and Results**: Waiting for the previous step to finish.

5 CONCLUSION

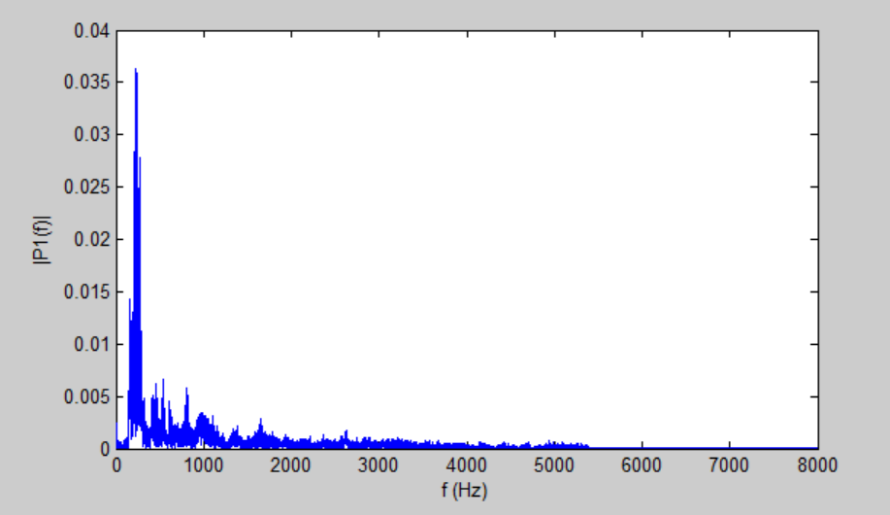
This report builds on existing works to show that inaudible voice commands are potential information security threat and proposes a new method to inspect the existence of attacking signals and eliminate them. Through simulations on computer, we can confirm the feasibility of this scheme, which can successfully find and remove the attacking signal in most circumstances. The further implementation practical verifications are moving forwards.

6 MY WORKS

In this project, our junior provided us with his idea and the main part of works. And my works is first, to simulate the scheme on computer to check it feasibility and estimate its performance with several external inference to test its robustness.
Then I, with my teammate, am designing and building curcuits to implement it and to obtain the actual testing data. It is currently moving forwards and I think we will finish it until the end of this semester.



(a) Before Processing



(b) After Processing

Figure 4.1: Simulation of Scheme

REFERENCES

[1] Gilda Gabriela Gámez González and Instructor M Sc Viktor Nässi. Measurements for modelling of wideband nonlinear power amplifiers for wireless communications. *Department of Electrical and Communications Engineering, Helsinki University of Technology*, 2004.

[2] Nirupam Roy, Haitham Hassanieh, and Romit Roy Choudhury. Backdoor: Making microphones hear inaudible sounds. pages 2–14, 06 2017.

[3] Guoming Zhang, Chen Yan, Xiaoyu Ji, Tianchen Zhang, Taimin Zhang, and Wenyuan Xu. Dolphinattack: Inaudible voice commands. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, CCS '17, pages 103–117, New York, NY, USA, 2017. ACM.