InfoMasker: Preventing Eavesdropping Using Phoneme-Based Noise

Peng Huang[†‡], Yao Wei[†‡], Peng Cheng[†‡], Zhongjie Ba[†‡§], Li Lu[†‡], Feng Lin[†‡], Fan Zhang[†], and Kui Ren[†‡]
[†]Zhejiang University, Hangzhou, China
[‡]ZJU-Hangzhou Global Scientific and Technological Innovation Center, Hangzhou, China
{penghuang, weiy, peng_cheng, zhongjieba, li.lu, flin, fanzhang, kuiren}@zju.edu.cn

***Abstract*—With the wide deployment of microphone-equipped smart devices, more and more users have concerns that their voices would be secretly recorded. Recent studies show that microphones have nonlinearity and can be jammed by inaudible ultrasound, which leads to the emergence of ultrasonic-based anti-eavesdropping research. However, existing solutions are implemented through energetic masking and require high energy to disturb human voice. Since ultrasonic noise can only remain inaudible at limited energy, such noise can merely cover a short distance and can be easily removed by adversaries, which makes these solutions impractical. In this paper, we explore the idea of informational masking, study the transmission and coverage constraints of ultrasonic jamming, and implement a highly effective anti-eavesdropping system, named InfoMasker. Specifically, we design a phoneme-based noise that is robust against denoising methods and can effectively prevent both humans and machines from understanding the jammed signals. We optimize the ultrasonic transmission method to achieve higher transmission energy and lower signal distortion, then implement a prototype of our system. Experimental results show that InfoMasker can effectively reduce the accuracy of all tested speech recognition systems to below 50% even at low energies (SNR=0), which is much better than existing noise designs.**

I. Introduction

Microphones are commonly seen in many kinds of electric devices nowadays, which keeps raising concerns over voice privacy. It is not exaggerating that an individual is constantly surrounded by several microphones wherever he/she is. As shown in Figure 1, the microphones, embedded in smart devices such as smart speakers, smart TV, and people's smartphones, can be exploited, compromised or even misconfigured to eavesdrop on the conversations happened in the environment [10]. The resulting speech recordings, once interpreted either by human or an automatic speech recognition (ASR) system, could leak large amount of victim's private information thus violating personal privacy. Lots of news report the risk of being eavesdropped all the time by smart home devices including Siri, Alexa, and Google Assistant [37], [31], [51]. Several highly visible news have shown the consequences caused by eavesdropping: In 2013, NSA was reported to routinely monitors calls of world leaders [34]; In 2018, the defense secretary of the UK was interrupted by voice assistant during Commons statement because Siri listens constantly to seek the wake word [33], which can be treated as a new form of eavesdropping. In 2020, the Ukraine prime minister submitted his resignation because a leaked recording suggesting he had criticized the president [18].

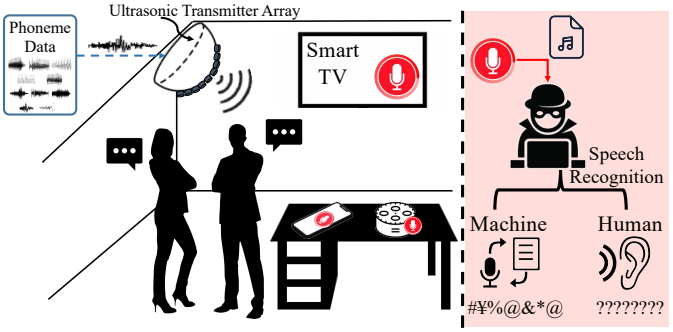


Fig. 1: InfoMasker prevents eavesdropping.

Anti-eavesdropping, although not a new problem, still needs reliable solutions desperately with the proliferation of microphone-equipped smart devices. However, even the latest solutions in industry are far from mature. Some manufacturers have developed products utilizing audible noise jamming. Project Alias [23] and Paranoid Home Wave inject white/chatter noise into a smart speaker to jam the potential sound monitoring and lifts the jamming if a custom wake word is detected [36], but they require a special jammer attached to the microphones of smart speakers, which limits their application scenarios. In addition, studies have shown it is possible to recover the original signal with speech processing techniques, even under white-noise jamming [49].

Audio injection via ultrasound [55], [40] is a trendy research topic in recent years, and the technique have been applied in various applications including anti-jamming. One one hand, a lot of works try to address the formidable threat brought by ultrasound injection, such as EarArray (NDSS'21) [54] and AIC [21]. On the other hand, works utilizing the technique keep emerging, and jamming microphone with ultrasound is one of these popular research strands. The procedure is as follows: jamming signals are first transmitted on an ultrasonic frequency area inaudible to human auditory system; the signals would be leaked into the audible frequency range automatically due to the non-linearity of microphones. Based on this, Chen et al. implemented a bracelet-like wearable

[§]Corresponding author

Network and Distributed System Security (NDSS) Symposium 2023
27 February - 3 March 2023, San Diego, CA, USA
ISBN 1-891562-83-5
https://dx.doi.org/10.14722/ndss.2023.24457
www.ndss-symposium.org

to emit special ultrasound signals, achieving good jamming coverage [10]. However, it depends on users' irregular arm movements to achieve omni-directional jamming, thus limiting its usability and being ineffective in other scenarios (e.g., meeting). Li et al. propose a system emitting their noise to disturb unauthorized microphones' recording while allowing the recording of legitimate devices utilizing the noise pattern knowledge [28]. Sun et al. design MicShield [46], a selective jamming mechanism to jam smart speakers with ultrasonic waves carrying white noise while passing authentic audio commands utilizing fast wake word prediction. The above ultrasonic jamming approaches rely on high energy to interfere with speech signals and make the recordings uninterpretable. Nevertheless, these studies have obvious limitations. To be specific, they only work in a short distance (i.e., shorter than 1 m), lack of human interpretation test on the jammed signals, and do not evaluate the robustness of their noise against denoising algorithms (see the experiments in Section IV). A determined eavesdropper can easily leverage these vulnerabilities to bypass these jamming methods.

We systematically analyze existing studies, and find all of them realize jamming based on energetic masking which is one of the two types of masking effects determining the obscuring performance in jamming. According to research in speech perception, the energetic and informational masking accumulatively resolve the level of interference. Since existing methods solely depend on energetic masking for jamming, a relatively high power is required to guarantee their performance. Meanwhile, long-range jamming also requires boosting the noise energy. Such a dual requirement for high energy constraint creates a central dilemma between jamming distance and inaudibility for these works. Because non-linearity also exists in speakers and power higher than certain threshold makes the jamming noise audible at the ultrasound transmitters [41], these ultrasonic jamming approaches all trade off jamming distance for noise inaudibility. Another major shortcoming of these methods is that they overlook the importance of informational masking thereafter do not design sophisticated noise form, which makes their noises easily removable by speech enhancement methods (e.g., noise removal algorithms). This is devastating for an anti-eavesdropping application as an adversary is very likely to apply denoising before extracting the semantic contents.

In this paper, we aim to propose an anti-eavesdropping system to achieve effective and reliable jamming in real-world privacy preservation scenarios. We explore the idea of informational masking and compose our jamming noise with multi-layers of phoneme sequences. As a result, the phoneme structure of the original speech signal can be greatly obscured by our phoneme-based noise (highlighted in red in Figure 1). since phoneme is the basic elements of sound used to distinguish words, the chaotic phoneme pattern makes the obscured speech signals intelligible. Moreover, our noise shows inherent robustness against noise reduction algorithms. The basic units of our noise are genuine phonemes, and such characteristic causes noise removal methods fail to disentangle the speech elements from our distracting ones due to the high similarity between them. We encourage readers to listen to audio examples at the demo website [1].With the utilization of informational masking, the requirement on high energy is relaxed, and we further address the challenge on increasing jamming range without losing inaudibility. We study the limitations of transmitting ultrasonic sound in the physical world, then optimize the noise transmission including designing a unique transmitter, choosing appropriate modulation, and tailoring noise shape to compensate device distortions, etc. In general, we summarize our contributions as follows:

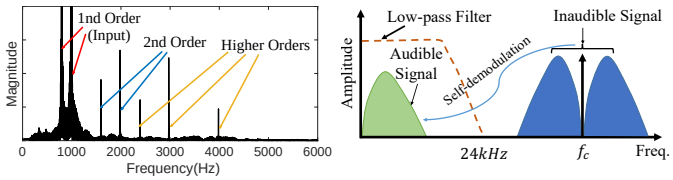


Fig. 2: Nonlinearity in microphone. The data in the left figure is recorded by a Huawei P10 smartphone and the inputs are two single tones with frequencies of 800Hz and 1000Hz.

- We propose a new type of noise, named Phoneme-Based Audio Jamming Noise, based on the idea of informational masking.
- We conduct an in-depth study of the ultrasound transmission method, and then optimize several aspects to make it practical in real-world scenarios and more suitable for our noise.
- Based on our noise and the optimized transmission method, we design and implement a system named InfoMasker. Through extensive experiments, our system is proved to have high effectiveness and high security in real-world scenarios.

II. Preliminary

In this section, we introduce the nonlinearity effect in the microphone, then we describe the linguistic structure of speech signals and how humans and machines understand them.

A. Nonlinearity in Microphone

A microphone is a type of transducer which converts acoustic signals into electrical signals. Previous studies show that the preamplifier in most types of microphones, including Electret Condenser Microphones (ECMs) and Micro Electro Mechanical Systems (MEMS) Microphone, involves nonlinear operations which cause inter-modulation distortion in its output [8]. As a result, the microphone's output contains both the frequency components of the input and all possible linear combinations of them [26]. To illustrate, suppose the microphone's inputs are two single tones with frequencies of f_1 and f_2, the nonlinearity makes the output contains not only components with frequencies of f_1, f_2, but also $f_1+f_2, f_1-f_2, 2f_1, 2f_2...$ etc., as shown in the left of Figure 2.

To inject a human audible noise signal $n(t)$ into a microphone stealthily, we first modulate $n(t)$ onto a high-frequency

[1]https://github.com/desperado1999/InfoMasker. Relevant codes will also be released here.

carrier $c(t) = cos(2\pi f_c t)$ via amplitude modulation and then transmit the modulated signal along with the carrier at the same time. When these signals arrive at a microphone, the non-linearity produces distortion that are harmonics and cross-products of the carrier and the modulated noise [55], generating a low-frequency shadow signal and other high frequency components. The shadow signal is the same as $n(t)$ and other components will be filtered out by the low-pass filter in the microphone, as shown in the right of Figure 2.

B. Informational Masking

Informational masking, which is first defined in [38], describes the degradation of the auditory detection threshold in the human brain when the target sound is embedded in other interferers with similar characteristics. Informational masking is usually associated with its complementary term: energetic masking, which occurs when interferers are present at the same time and frequency bands [27]. Unlike energetic masking which mainly depends on the relative energy between the target and the interferer in each frequency band, the degree of information masking mainly depends on the similarity between the target and the interferer. Generally speaking, these two types of masking are not independent and they always affect the auditory detection threshold simultaneously.

C. Human Auditory System and ASR

One of the main tasks of the human auditory and ASR systems is extracting semantic information from speech signals. In order to improve speech intelligibility, both systems need to first eliminate the noise in the signals, then extract phoneme series, the primary component of a speech signal, and decode the phoneme sequence into meaningful content.

The attention mechanism in the human auditory system, as known as "Cocktail Party Effect", has a strong noise reduction effect [14], [27], [7], [29]. It allows a listener to distinguish signals from different sources and then eliminates the influence of uninterested noises, thus achieving the effect of denoising. To be more vivid, imagine you are whispering to a friend in a crowded cocktail bar where there are many people talking loudly. Although other people's voices may be louder than your friend's, you are still able to focus on his/her voice and understand what he/she is talking about. This mechanism also helps human beings to reduce the impact of masking effect caused by interferences. The effectiveness of the attention mechanism mainly depends on the degree of difference between the target signal and the noise in three aspects: fundamental frequency, temporal properties, and spatial distribution. A larger difference means better discrimination with the help of the attention mechanism.

A typical ASR system works similarly to the human auditory system to "comprehend" speech signals. Most of these ASR systems will first extract voice features, such as mel-frequency cepstral coefficients (MFCC), from the audio segments and then recognize the phoneme series from these features using an acoustic model. Then with the help of the pronunciation model and language model, the phoneme series will be decoded into normal text information. To improve recognition accuracy, noise reduction methods are always applied before recognition. The widely used speech enhancement methods for ASR systems can be roughly divided into two types: noise reduction and noise separation. The former, such as spectral subtraction and wiener filter, targets on reducing the noise from the signals. This type of method relies on the invariance of statistical characteristics of noise and the differences in temporal properties between the noise and the target speech signal. The noise separation, such as blind signal separation, targets on separating source signals from mixed signals. Such methods rely on the assumption that each source signal is independent.

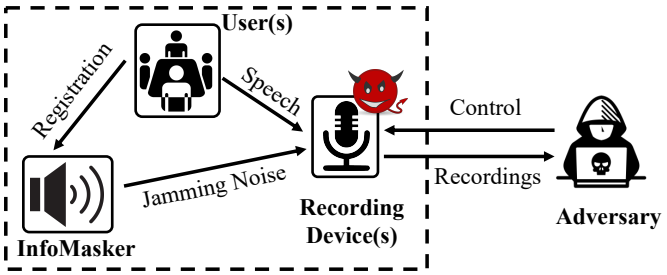


Fig. 3: System Model.

In this work, we aim to prevent both humans and machines from extracting semantic information embedded in speech signals by injecting noise signals into the recordings. To achieve this target, the injected noise should interfere with human and ASRs' understanding of semantics. Equally critical, the injected noise should be difficult to remove by both auditory attention mechanism and noise reduction algorithms. Therefore, we design our noise that is highly correlated with speech signals to affect the phoneme structure of the target speech signal, which disturbs human and ASR understanding of the semantics, and improve the robustness of noise against speech enhancement methods.

III. Problem Formulation

In this section, we first introduce the system and threat model and then discuss the design goals of our system.

A. System Model

We consider scenarios where people want to protect their voice privacy in a common indoor environment such as a conference room, a dining room and an office. Without losing generality, we use the office as an example in this paper. As shown in Figure 3, the system involves three entities:

User(s): The users are the people who want to prevent their voices from being eavesdropped on.

Recording Device(s): The recording device is the collection of all devices equipped with microphone(s) in the environment, such as mobile phones and smart home devices. Due to the black box nature of most electronic products, they are not fully controlled by the users and may secretly record users' conversations. Considering the ease of use, users are unlikely to put these devices in hidden areas which may cause non-line-of-sight (NLOS), so we do not consider NLOS scenario in this paper.

InfoMasker: InfoMasker is a jamming device equipped with ultrasound transmitters. It can generate special noise signals according to the users' registration information and inject these signals into surrounding recording devices to make the recordings unrecognizable to both humans and machines. Please note that we do not cover the voice call scenario since all microphones in the environment will be jammed.

B. Threat Model

We consider an adversary who can control recording devices in the environment (e.g., the manufacturer of the smart home devices). To eavesdrop on the content of the conversation, the adversary would record the speech signals, improve the recordings' intelligibility using speech enhancement methods, and then extract semantic information. In this work, we assume the adversary has the following capabilities in each step of eavesdropping:

Audio Recording. We assume the adversary can control one or more recording devices in the environment to record single-channel or multi-channel audio signals.

Speech Enhancement. To improve the intelligibility of recordings, the adversary can enhance the quality of speech signals with different speech enhancement methods including noise reduction algorithms and *BSS* (Blind Signal Separation) algorithms if multi-channel recordings are acquired.

Semantic Information Extraction. We consider the adversary can jointly use different ASRs in conjunction with human listening to extract semantic information from the enhanced recordings, where the human can recognize recordings accurately and ASRs can interpret speech contents efficiently.

We also consider a powerful adversary who knows our noise generation methods and can train a specific ASR system to extract information from the recordings jammed by our noise.

C. Design Goals

We envision the following design goals of InfoMasker:

Effectiveness: The noise transmitted by InfoMasker should be able to prevent the target speech from being recognized by ASR systems and the human auditory system, and the target speech signal can be produced by a single people or by multiple different people.

Robustness: The noise signal injected by InfoMasker should be robust against noise reduction and speech enhancement methods.

Low-interference: The noise signal transmitted by the InfoMasker should be inaudible to human beings.

IV. Phoneme-Based Informational Masking

A. Key Insight

The main goal of this paper is jamming microphones in the environment by transmitting noise to make the disturbed recordings unrecognizable to both human and ASR systems. The success of jamming ASR systems mainly relies on the noise's robustness against noise reduction methods, while the success of jamming the human auditory system mainly relies on the two masking effect: informational masking and energetic masking. The effectiveness of energetic masking mainly depends on the relative energy of the noise to the target signal, while for informational masking, it mainly depends on the content of the noise [7], [27], [14]. Existing works [40], [28], [10] focus on utilizing energetic masking to jam speech signals, which leads to a high demand for energy and is not suitable in the ultrasonic transmission scenario.

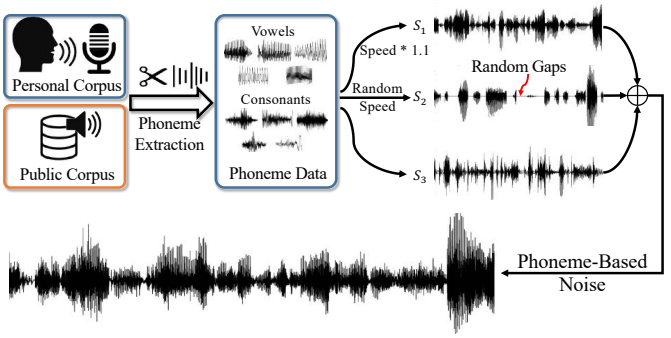


Fig. 4: Generation of phoneme-based noise.

In this paper, our key insight is exploiting the informational masking effect to form noise. Previous studies have shown that when presented with noises, the differences in phonetical properties such as fundamental frequency (F0) between the noise and the target speech signal will assist listeners to filter out the noises and understand the target speech signal[7]. Therefore, we first construct noise that has similar F0 properties to the target speech signal. Additionally, through experiments we find that the difference in speech rate between the noise and the speech also affects the effectiveness of informational masking, which inspires us to treat speech rate as another factor in our phoneme-based noise design.

B. Noise Design

As shown in Figure 4, our phoneme-based noise consists of three phoneme sequences: S_1, S_2, and S_3. S_1 is a sequence of random vowels without gaps in between, and it is accelerated by a preset parameter to include more vowels per unit time for better jamming effectiveness, and the parameter is set to 1.1 according to our experiments. The reason we choose vowels is that compared to consonants, vowels make up most of the energy in speech signals. Because the difference in phonetical properties such as fundamental frequency (F0) between the speech signal and the noise will assist listeners in separating the noise, we extract the phoneme data from the target people's speech materials to minimize such difference. Simple concatenation of the phoneme data will cause discontinuity at phoneme boundaries. To smooth the connected phoneme data, we apply a hamming window with a length of 25ms at the juncture between phonemes. Without special mention, this smoothing method is also applied to the following sequences that make up our noises.

Then we consider narrowing down the speech rate gap between the noise and the target signal. Since the target signal's speech rate is unknown, we add another vowel sequence S_2 with random speech rate. Different from S_1, there are gaps

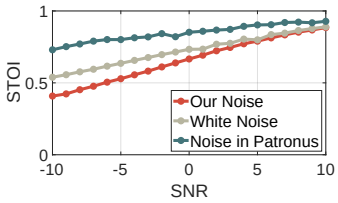


Fig. 5: Comparison of different types of noise via STOI

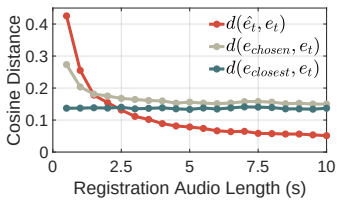
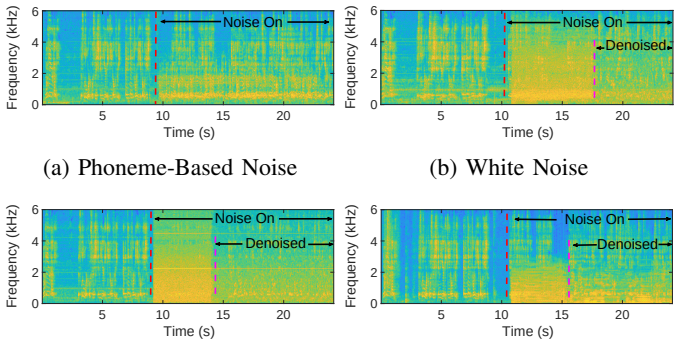


Fig. 6: Comparison of different registration speech length



(a) Phoneme-Based Noise

(b) White Noise

(c) Noise in [10]

(d) Noise Generated by [3]

Fig. 7: Comparison of the robustness of different noises against the built-in noise reduction in a Vivo Nex smartphone.

of random length between the vowels in S_2. Each vowel is accelerated or decelerated with a randomly chosen speed factor that follows a uniform distribution $U(0.3, 1.8)$ which is chosen according to our experimental results. In addition to making the rate more similar to the target signal, S_2 also makes the noise more varied and shows less constant patterns, which helps with the robustness against noise reduction methods.

Apart from vowels, we add a consonant sequence S_3 to increase the diversity of noise. As the difference in consonants between different people is much less than that in vowels, we extract the consonant data from a public speech corpus (LibriSpeech [35]). Then we connect these consonants without gaps to form S_3.

Our noise is the superposition of the above three phoneme sequences. We adopt the metric STOI [47], a function of Time-Frequency(TF)-dependent intelligibility measure, to test the intelligibility of the speech signal jammed by different noises. STOI ranges in [0, 1] and a higher value indicates better intelligibility. The result in Figure 5 shows that our noise outperforms other tested noises. We further test the robustness of our noise. As shown in Figure 7, our noise is robust against the noise reduction process in the Vivo Nex smartphone, while the other types of noise are suppressed significantly.

V. System Design

A. System Overview

The system overview of InfoMasker is shown in Figure 8. Its workings involve three phases: registration, data augmentation, and jamming. In the registration phase, InfoMasker will first collect speech materials from the users and then extract

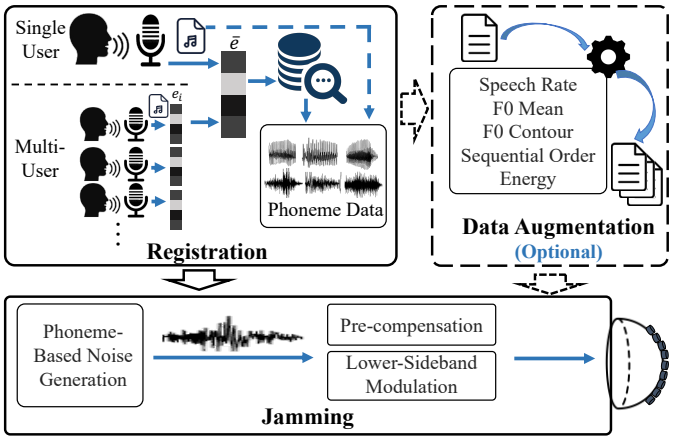


Fig. 8: Overview of InfoMasker.

their voice features from these materials. Then depending on the amount of speech materials, InfoMasker will segment the speech materials into phoneme data or search its database to find a shadow speaker whose voice features are the most similar to the extracted features of the target speaker. Data augmentation is then applied optionally according to the users' security requirements to expand the data samples. Finally, during the jamming process, the system will continuously generate phoneme-based noise based on the augmented data and transmits the noise to jam microphones in the environment.

B. Registration Phase

As stated in Section IV, the phoneme-based noise is generated from the target people's speech data. However, collecting a large number of speech samples of the target people is time-consuming thus not practical all the time. Our solution is to pick speech data similar to the target people's voices from a dataset. The similarity between the target people's voices and the dataset audio clips is represented by the cosine similarity of their feature vectors, extracted from the speech data with the method proposed in [50]. Denote two feature vectors as $\boldsymbol{e_1}$ and $\boldsymbol{e_2}$, then the cosine similarity of them can be represented as $1 - d(\boldsymbol{e_1}, \boldsymbol{e_2})$, where $d(\boldsymbol{e_1}, \boldsymbol{e_2}) = \frac{\boldsymbol{e_1} \cdot \boldsymbol{e_2}}{||\boldsymbol{e_1}|| ||\boldsymbol{e_2}||}$.

To evaluate the effectiveness of this solution, we test four types of noise showing different similarity to the target people's speech: the target people's voice, the speech with the highest similarity in the dataset, the dataset speech with the same gender to the target, and the dataset speech with different gender. We also use Gaussian white noise as a baseline. The dataset we used here is train-clean-360 from LibriSpeech [35]. As shown in Figure 9, the jamming performance of the noises decreases as the voice similarity drops, and there is about 10% gap in Word Error Rate (WER) between the best and the worst case [1]. Based on the results, we consider two types of registration scenarios according to the number of users.

Single-User Registration. In this scenario, our priority is protecting a specific user's voice privacy. Meanwhile the voice privacy of others is also protected, merely with a 10% less effectiveness as shown in Figure 9.

[1] In this paper we use Tencent ASR [48] for speech recognition if there is no special description

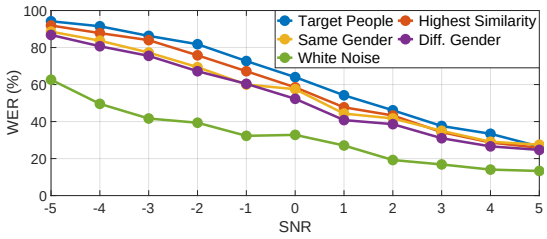


Fig. 9: Effectiveness of different types of nosie.

There are two registration ways depending on whether the user has enough time or not. Given enough time, the user can record adequate speech data with Harvard Sentences dataset [1], which has a set of phonetically balanced sentences. With these recordings, we can extract sufficient and balanced phoneme data for noise generation. In contrast, if the registration is time-restricted, the user just need to record a few sentences for voice feature extraction. Then the data showing the highest similarity in the dataset will be chosen for noise generation. To get the appropriate recording length for this time-restricted registration, we test the impact of the length on the cosine similarity between the feature vector extracted from the recording of the limited length $\hat{\boldsymbol{e_t}}$ and the ideal feature vector $\boldsymbol{e_t}$ (i.e., ground truth) (see Figure 6). We find that when the length is smaller than 5 seconds, the distance first drops rapidly with increasing speech length, and then drops much slower. This shows that when the length reaches certain level, its benefit to the similarity turns moderate.

We further calculate the similarity value between the people in the dataset who match the target people best under varying length of the target people's recording, which is shown as $d(\boldsymbol{e_{chosen}}, \boldsymbol{e_t})$ in Figure 6. Meanwhile, we use the target people's long enough recording to extract a feature vector, find the people showing the highest similarity to it, and calculate their similarity which is presented as a relatively flat curve denoted as $d(\boldsymbol{e_{closest}}, \boldsymbol{e_t})$.

The result in Figure 6 shows that with the increase of the length of the registration speech, $d(\boldsymbol{e_{chosen}}, \boldsymbol{e_t})$ becomes close to $d(\boldsymbol{e_{closest}}, \boldsymbol{e_t})$, and the distance between them is neglectable when the length is greater than 5 seconds, indicating 5-second speech is appropriate for registration.

Multi-User Registration. In this scenario, we protect the voice privacy of all the people present in the environment. Please note that the online scenario (such as a video call) is not covered since all microphones in the environment will be jammed. Similar to the second way of registration in the single-user scenario, the users need to read the sentences selected from Harvard dataset and their voice features are extracted from the recordings. However, different from the single-user scenario, each user may be matched to a different speaker in the dataset. To improve the jamming performance for each user, we use the average of all user's voice vectors as the representative feature for this group of people, then finds the best-matching data in the dataset and for noise generation.

Please note that our system is totally offline during the usage, so the registration phase will not cause privacy leakage.

C. Data Augmentation

After the registration, we are now able to generate phoneme-based noise. However, since the amount of phoneme data is limited by the dataset, the probability of occurring recurring fragments in the noise will increase as more noise is generated, especially when the dataset is small. When an attacker locates these fragments, it is very possible for him to recover part of the semantic information from the recordings with the help of language models, similar to the key reused scenario in running key cipher. To prevent this, a data augmentation process that fine-tunes the phoneme data is applied to increase the total amount of data. To retain a high similarity between the augmented data and the target people's audio, the fine-tune should be restricted to a people's inner differences. Here we adopt the following properties used in emotion recognition [12], speech rate, F0 mean, F0 contour, and energy, which have inherent inner differences because of the people's different emotions as shown in Table I In addition, we randomly reverse each single phoneme data, which has little effect on the data's phonetical properties but disrupts its semantic information.

Phonetical Properties	Modification Range	Emotional Impact	
		↑	↓
Speech Rate	0.3-1.8	Fear or Disgust	Sadness
F0 Mean	0.9-1.1	Anger or Happiness	Disgust or Sadness
F0 Contour	0.7-1.3	Anger or Happiness	Sadness
Energy	0.5-2.0	-	-
Sequential Order	-	-	-

TABLE I: Phonetical properties for data augmentation

Speech rate. In emotion recognition, a high speech rate usually implies disgust or fear, while a low speech rate indicates sadness [12]. We change the original speech rate with a factor α according to [13] and α is uniformly sampled from $[0.3, 1.8]$, which is obtained experimentally to guarantee an acceptable impact on phonetical properties. Please note that the change of speech rate here is independent of the acceleration of S_1 in the noise design, which aims to increase the number of vowels per unit time of our noise.

F0 mean. A typical adult male's F0 mean always falls in $[85, 155]$ Hz, and that of an adult female is $[165, 255]$ Hz [6]. A high F0 mean usually implies anger or happiness, and vice versa sadness [15]. Similar to before, we vary the F0 mean by a multiplicative factor α which is uniformly sampled from $[0.9, 1.1]$. This range is also determined experimentally to limit the impact on phonetical properties.

F0 contour shows the audio's pitch change along with time. An exaggerated contour usually implies anger or happiness, and vice versa means sadness. We exaggerate or flatten the F0 contour according to Equation 1 using World vocoder [30] with a factor α uniformly sampled from $[0.7, 1.3]$.

$$F0^{'} = \alpha(F0 - \overline{F0}) + \overline{F0} \tag{1}$$

Energy. We modify the average amplitude of the data in the time domain with a factor randomly sampled from $[0.5, 2]$.

Sequential order. For human speech data, the reverse in the time domain has little effect on phonetical properties

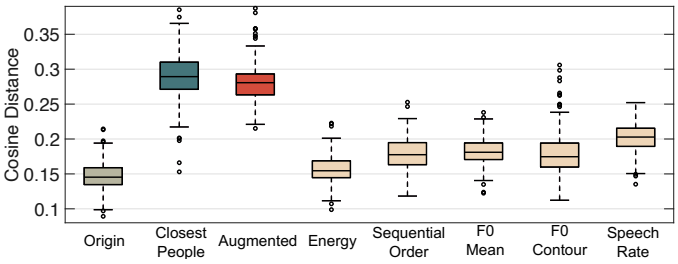


Fig. 10: Impact of data augmentation on the audio's phonetical features. The boxes illustrate the distribution of $d(\hat{e}, \boldsymbol{e_t})$ where $\boldsymbol{e_t}$ is the target people's voice feature and $\hat{e}$ is the feature extracted from different types of audio.

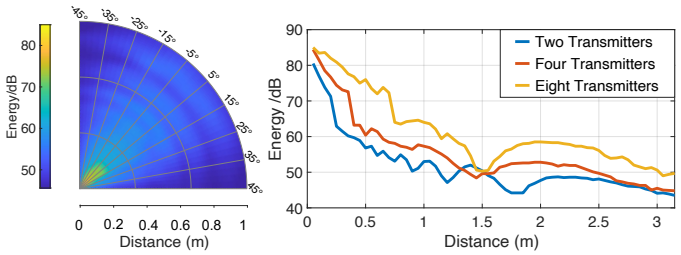


Fig. 11: Characteristics of the transmitter. Left: Energy distribution of two transmitters. Right: Energy attenuation with distance when the angle is $0°$.

but will disrupt the semantic information. In this paper, we randomly reverse each phoneme data in the time domain.

To visualize the impact of data augmentation, we encode each audio into a feature vector [50] thus obtaining vectors from different types of audio, then calculate the cosine distance between these vectors and the average feature vector of the corresponding people. We compare four types of data: the data from the same people, from the people who have the closest voice feature to the target people among the dataset (LibriSpeech train-clean-100 [35]), the data processed by all five augmentation methods, and processed by every single method. Results in Figure 10 show that augmenting the data with a single method has limited impacts on its phonetical features. Even when augmented by all the five methods, the data is more similar to the original data than that of the closest people in the dataset.

VI. Noise Transmission

Considering the implementation of InfoMasker, we need to address the limitations involved with acoustic sensors and the conflict between jamming range and inaudibility. We detail the noise transmission method that responds to these challenges.

A. Characteristics of Transmitters

We first study the characteristics of a widely-used off-the-shelf transmitter [4] to prepare for our array design. To simulate the noise injection scenario, we play 39kHz and 41kHz single tones through two transmitters separately and measure the energy of the demodulated signal around transmitters with a sound level meter. The result in the left of Figure 11 shows that the energy around transmitters attenuates rapidly with angle, which drops nearly 10 dB within $25°$ and so cannot meet the jamming requirements in the real-world. There are two reasons account for this. Firstly, compared to human audible sound, ultrasound propagates straight. Secondly, the success of noise injection requires both the carrier and the modulated noise signal arrive at the recorder. We further analyze the energy attenuation with varying distance when different number of transmitters are deployed. As shown in the right of Figure 11, although the energy attenuates rapidly as distance increases, we also witness the increase of ultrasound energy via utilizing more transmitters.

There are two insights from these results: 1. To increase the effective jamming distance, we can simply increase the number of transmitters. 2. To increase the effective jamming angle, it is necessary to further design the distributions of the transmitters for carrier signal and modulated noise signal.

B. Transmitter Array Design

To extend the coverage of the transmitter, we design a transmitter as shown in Figure 12 (left). We install transmitters in a hexagonal shape on the spherical foam and separate them into two groups: one group sends the carrier signal and the other sends the modulated noise signal. The curvature of the spherical foam and the distribution of the two groups of transmitter enable the large effective coverage of the transmitter array. The energy distribution around the transmitter array is shown in Figure 12 (right). It is obvious that the transmitter array can cover a large span of angle up to about $90°$ and a much longer distance compared to using a single transmitter.

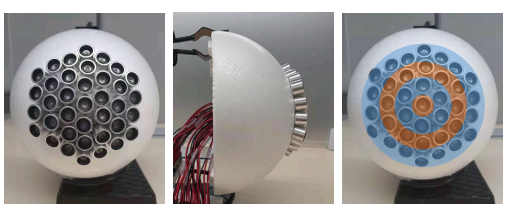
C. Pre-compensation

Most acoustic sensors exhibit non-flat frequency response due to the imperfection of manufacturing, which could cause distortion of the transmitted noise thus decreasing the jamming effectiveness. To address this, we analyze the frequency response $H_2(f)$ of the recorder, and the equivalent frequency response $H_1(f)$ between the transmitter and the recorder. Then an inverse filter $h_1^{-1}(t) \circledast h_2(t)$ [1] is applied on the noise signal before modulation to compensate the unwanted distortion. We analyze $H_1(f)$ and $H_2(f)$ using multiple recorders and then use the average of them respectively. Figure 14 shows that this process suppress the distortion evidently. Please note that only the frequency blow 4kHz is considered for compensation because the majority of energy in human speech falls in this range. The drop below 1kHz in the ideal spectrum is caused by the common recorders' imperfection.

D. Lower-Sideband Noise Modulation

Modulating the noise signal onto high-frequency carriers makes the jamming signal inaudible. However, the two widely used modulation methods, double sideband amplitude modulation (DSB-AM) [55], [28] and frequency modulation (FM) [40], are not suitable here. The wide frequency range of our noise results in a large bandwidth of the modulated signal if DSB-AM is applied, which increases the audibility of the jamming noise because of the self-demodulation in the

[1] $h(t) \overset{\mathcal{F}}{\longleftrightarrow} H(f)$ means Fourier transform and $\circledast$ means convolution.



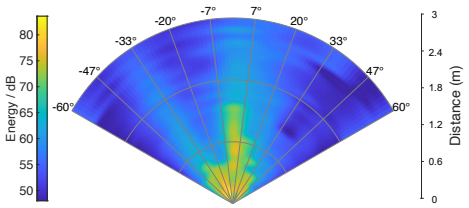


Fig. 12: The transmitter array. Left: The transmitters in orange transmit the carrier signal and the others in blue transmit the modulated noise signal. Right: Energy distribution of the array

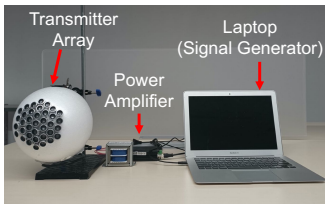


Fig. 13: Hardware implementation.

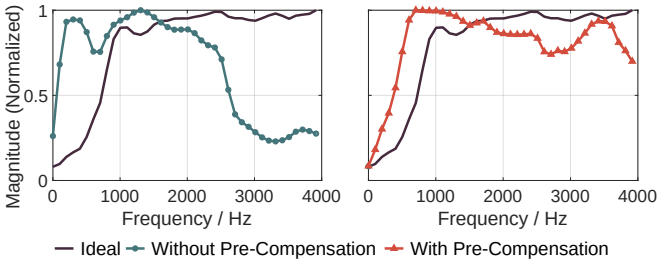


Fig. 14: Spectrum of recorded sweep signals.

transmitter [40], [41]. As for FM, our phoneme-based noise cannot be recovered by self-demodulation at the receiver end.

In this paper, we use single-sideband modulation (SSB-AM) to reduce the noise audibility. Take the lower-sideband amplitude modulation (LSB-AM) as an example, assume the carrier signal is $cos(2\pi f_c t)$ and the noise signal is $n(t)$, the LSB-AM can be represented by $n_L(t) = n(t)cos(2\pi f_c t) + \hat{n}(t)sin(2\pi f_c t)$, where $\hat{n}(t)$ is the Hilbert transform of $n(t)$. Theoretically, SSB-AM can reduce the power of the audible sound generated at the transmitter to 0.5 times the original (see proof in Appendix-A). Additionally, we choose lower-sideband instead of higher-sideband because of its lower attenuation in the air. The f_c we used is 40 kHz, which has the maximum resonance response in most microphones [40].

We conduct an experiment with 3 male and 3 female aging from 22 to 31 [1] to test the impact of modulation methods on audibility. We gradually increase the transmission power until volunteers perceive the noises. The normalized maximum transmission energy is shown in Table II.

Noise	Normalized Energy		
	DSB-AM	LSB-AM	USB-AM
White Noise	1.00	1.49	1.29
Phoneme-Based Noise	2.77	**4.14**	3.61

TABLE II: Normalized max transmission energy.

VII. Evaluation

A. Experimental Setup

Evaluation Methodology. In this part we systematically evaluate the performance and the robustness of our system under various variables. At last, a case study in a common office is conducted to validate the practicality of our system.

Dataset. We use three datasets here. TIMIT [16] is used in Section VII-D because it small size which cam make machine learning models converge faster. Harvard Sentences [1] is used for human perception test. The sentences included in the dataset are short, which can reduce the difficulty for human to recognize. LibriSpeech [35] is used in the remaining parts.

Hardware Implementation. The hardware implementation is shown in Figure 13, which includes a transmitter array, two power amplifiers (one for the carrier and the other for the modulated noise), and a signal generator (a laptop with a soundcard sampling rate >80kHz). Without considering the laptop, the hardware implementation costs about 70 dollars.

Please note that the inconsistency in the accuracy of Tencent ASR across experiments is caused by differences in test sets and test dates. But for each experiment, the test set is consistent and is done within a short time period.

B. Baseline

Overall Performance. We first evaluate our noise in the single-user scenario. We test the jamming performance of our noise under different ASRs and use a [0, 8] kHz band-limited gaussian white noise for comparison. A test set containing 27000 words is generated from LibriSpeech. Since built-in noise reduction mechanism of recording devices may affect the reception of white noise (proven in Fig 7b), we directly mix noises with speech signal in digital domain and feed the mixed signals to ASRs for a fair comparison.

We test four commercial ASR systems (Amazon Transcribe [5], Tencent ASR [48], Xunfei ASR [22] and Google Speech-to-Text [17]) and two commonly used open source ASR systems (DeepSpeech [19] and WeNet [53]). The results in Figure 15 show that our noise performs significantly better than white noise when $SNR \leq 4$, and the gap between them gradually increases as the SNR decreases. Besides, the advantage of our noise is more obvious on commercial ASRs than on open-source ones. We suppose this is because commercial systems have been enhanced for the interference of white noise.

Multi-User Scenario. We evaluate our system with different number of randomly chosen user. The results in Figure 16 show that when the number of users is smaller than 10, the jamming effectiveness is lower than that of the single user scenario where the noise is generated based on the target person's speech, but significantly higher than the noise generated from the speech data of the other person with the

[1] All procedures performed in studies involving human in this paper are validated through an institutional review (IRB)

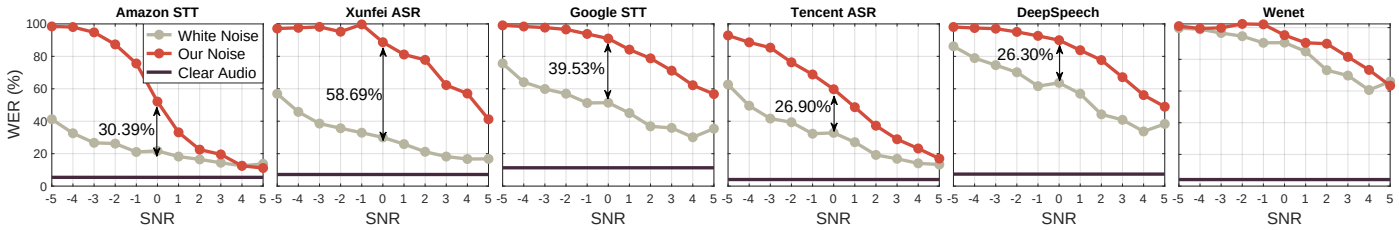


Fig. 15: Recognition results of different ASR systems. We use the clear audio as a baseline.

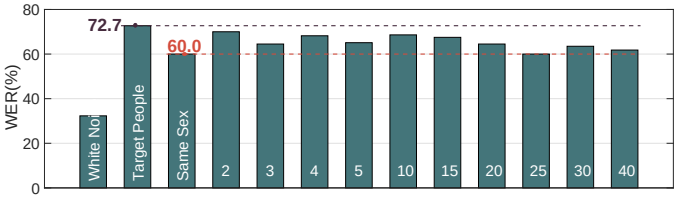


Fig. 16: Recognition results for multi-user scenario. The numbers in the bar represent the number of user.

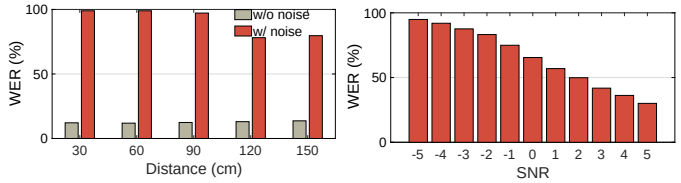


Fig. 17: Results of the end-to-end scenario. Left: results in different distances. Right: results in different SNR.

same sex. When the number of users is greater than 10, the noise's jamming effectiveness gradually close to the noise generated from the people with same sex.

Simulations of Real-World Scenario. Here we simulate the real-world scenarios under different conditions. As there are two main factors that affect the quality of recordings, namely room impulse resonances (RIRs) and environmental noises, we adopt the RIRs Noises dataset [25] which contains various room impulse responses and point-source noises for the simulation. Besides, we also use a tool to produce RIR parameters corresponding to different room sizes. We simulate the scenarios where speech undergoes various reverberation conditions during the propagation: real-world RIRs, small room RIRs, medium room RIRs, and large room RIRs respectively, and random point-source noises are included in each case of them. We also consider the scenario where only the point-source noise exists. The impact of reverberation on recognition is acquired by feeding these audios to an ASR. For comparison, We mix our noise and audio recordings incorporated with RIRs and noises following a typical setting (i.e., SNR = 0) and test what the recognition results will be. The results in Figure 21 show that RIRs and point-source noises have limited impact on the WER of clear audios ($< 12\%$). But when the audio is jammed by our noise, a significant increase on WER is obtained. Especially, compared with the origin situation (i.e., no RIR), the jamming sees an increase by $10\% \sim 35\%$ while RIRs and noises are considered. These results indicate that RIRs and point-source noises in real-world scenarios could benefit our noise for jamming eavesdroppers.

Real-World Scenario. To evaluate our system in real-world scenario, we place several smartphones around the transmitter array and adjust the transmitter's energy to get different SNRs. We use another smartphone to play speech signals around these smartphones. As it is hard to get a stable and precise SNR in real-world scenario, we test several times in each SNR interval and then calculate the average WER and the minimum WER. We collect more than 70 hours data in totally. The results in Table III show that when $SNR < 0$, the WER in real-world scenario is slightly lower than that in digital domain, but significantly higher when $SNR > 0$.

SNR	<-4	[-4,-2]	[-2, 0]	[0,2]	[2,4]	>4	Clear
Avg WER(%)	85.8	81.6	77.6	70.2	56.4	42.3	11.5
Min WER(%)	68.6	77.0	62.4	62.2	45.3	30.3	-
Digital WER(%)	88.6	85.4	68.8	48.67	28.9	17.0	4.1

TABLE III: Recognition result in real-world scenario.

Real-World End-to-End Scenario. We further evaluate our system in a real-world end-to-end scenario with two volunteers (one male and one female). Either of them reads one sentence (about 5 seconds) randomly selected from [1] for registration. We extract their voice embeddings and then match the corresponding closest people in LibriSpeech for noise generation. To simulate a realistic scenario, we place the transmitter array on an office table and then place 5 smartphones acting as eavesdroppers with different distances to the array. We have the volunteer sit at the table and read 50 sentences selected from [1]. As it is hard to control the SNR precisely in a real-world scenario, we record the noise and speech separately and then mix them under various SNRs. The recognition results of the mixed audios (i.e., jammed speech) are shown in Figure 17. The results show that our system performs well in the real-world end-to-end scenario.

Human Perception. We recruit 15 testees including 5 females and 10 males aged from 22 to 31 to test the intelligibility of jammed audios. To make testees put their effort into the study, we adopt an accuracy-related reward to incent them. Tested audios are either generated in digital domain or recorded over-the-air. In the digital domain, in addition to white noise and our noise generated with the target's speech, we also test our noise generated from audios of same & different sex people (SNR=-1). In the over the air setting, we test clean audios and jammed audios recorded in VII-J. In addition to recognizing the audio, we also ask testees to

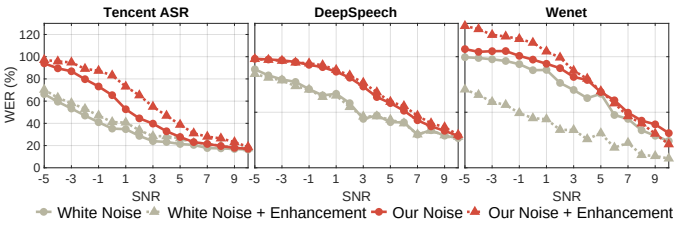


Fig. 18: Recognition results after speech enhancement.

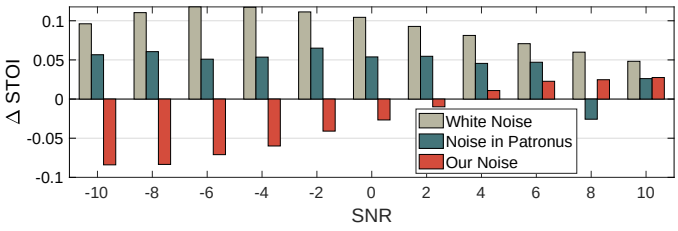


Fig. 19: Change of intelligibility after speech enhancement.

score the intelligibility of each audio from 0 to 5, where 5 indicates easily understandable and 0 is the opposite. In total, each testee needs to recognize 37 sentences selected from Harvard Sentences. The results in Table IV show that our noise performs better than white noise. Besides, audios jammed by our noise also exhibit the worst intelligibility.

	Clear	White Noise	Same People	Same Sex	Diff Sex	Clear	Same People
WER (%)	26.69	64.28	75.89	67.09	60.72	26.55	99.9
Score (0-5)	4.88	2.46	1.54	1.91	2.3	4.62	0.18

TABLE IV: Human perception result. Audios of the right two types are recorded in the over-the-air scenario.

C. Robustness against Speech Enhancement Method

To test the robustness of our noise against speech enhancement methods, we use a SOTA speech enhancement algorithm [20] to enhance the speech signal obscured by different types of noise before feeding them to ASRs. We test three ASR systems: Tencent ASR, DeepSpeech, and Wenet. The results in Figure 18 show that the Tencent recognition accuracy decreases significantly after enhancement (for both our noise and white noise cases). We speculate this is caused by the conflict between Tencent inherent speech enhancement and the SOTA algorithm we used. For DeepSpeech, the enhancement has limited impact on the result. However, for Wenet, the accuracy improves significantly for the white noise jamming case, which even surpasses the accuracy of DeepSpeech. While for our noise, the accuracy decreases after enhancement when $SNR < 5$. We speculate this is due to the difference of training data for DeepSpeech and WeNet. Pretrained models of DeepSpeech are trained with more versatile data (e.g., Common Voice Corpus), which grants DeepSpeech certain robustness against speech enhancement so there is almost no difference before and after the processing. Overall, our noise show robustness against the SOTA denoising method on both commercial and advanced ASRs.

STOI Test. We also compare the intelligibility of the disturbed audios before and after speech enhancement via STOI. We compare our noise with white noise and the noise proposed in [28]. As shown in Figure 19, for the other two types of noise, the enhancement process improves their intelligibility significantly at each SNR level. While for our noise, the audio intelligibility only improves when SNR > 3. We think the noise enhancement process may enhance the phonemes in both the speech signal and our noise, which leads to a bigger difference between the disturbed audio and the original speech.

Robustness in Real-World Scenario. We further evaluate our noise's robustness in real-world scenario. The same enhancement method as before is used here to process the data collected in real-world scenario. The result is shown in Tab V. Similar to the results in digital domain, the enhancement process reduces the recognition accuracy at each SNR range.

SNR	<-4	[-4,-2]	[-2, 0]	[0,2]	[2,4]	>4
Avg WER(%) (with Enhancement)	87.1	87.6	82.9	79.1	65.3	53.7
Avg WER(%) (without Enhancement)	85.8	81.6	77.6	70.2	56.4	42.3

TABLE V: Result of the real-world scenario w/ enhancement.

D. Robustness against A Specialized ASR

We consider a powerful attacker who can train a specialized ASR to extract information from the jammed recordings. We choose TIMIT [16] as the training data and take CRDNN [39] as the network architecture for its SOTA performance in speech recognition on TIMIT. The metric used here is PER (Phoneme Error Rate). We compare the recognition accuracy of specialized ASRs trained to recognize our noise and white noise respectively. Specifically, we train multiple ASRs and each one takes jammed speech signals with a specific SNR as the training data. As last, we have 22 ASRs considering the combinations of two types of noise and 11 SNRs.

The results are shown in Figure 20. For white noise, the PER rises slowly with the decrease of SNR, which indicates the network's denoising ability for white noise. For our noise, the PER is slightly higher than white noise when $SNR \geq 0$, but increases rapidly when $SNR < 0$. These results means that even when the attacker can train a specialized ASR, it is hard to recognize the audio jammed by our noise when $SNR \leq 1$, which is a reasonable value in real-world scenarios. We also find that when $SNR < 0$, the training for the ASR targeting our noise can not converge properly for many reasons (e.g., exploding gradients), and requires careful parameter tuning.

E. Comparisons with the Speech Noise

In this part, we compare our noise with speech noises that contain 1, 2, and 3 speech series respectively. For a fair comparison, the speech series making up the noise is from the same people as the target. The results on the left of Figure 22 show that without considering noise reduction methods, our noise performs better when $SNR < -2$.

We then test the robustness of the noises against noise reduction methods. As all the tested noises are speech-like

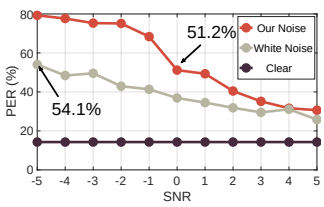


Fig. 20: PER of specialized ASR systems.

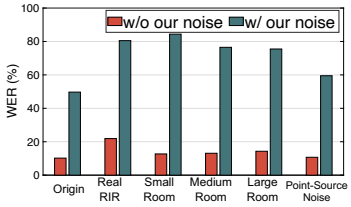


Fig. 21: Simulation of real-world scenarios.

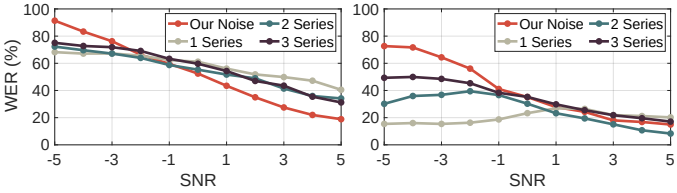


Fig. 22: Comparisons with speech-like noise. Left: WER before noise reduction. Right: WER after noise reduction.

noises, we choose a SOTA model structure of speech separation [45] and trained three models targeting noises containing 1, 2, and 3 series respectively. The detailed training strategy is presented in Appendix-B. During the experiment, we find that although the separation process improves the intelligibility of audio for human ear, the accuracy of the ASR recognition is even worse than before. We think this is caused by the noise residue. So we further process the separated results with a speech enhancement method [9]. For our noise, we try to separate it with all three models separately then apply the enhancement. The result with the lowest WER is chosen for comparison.

The result on the right of Figure 22 shows that when $SNR \geq 0$, the WERs of different tested noises are close; when $SNR < -1$, the WERs of speech noises with all SNRs are lower than 50% and our noise performs much better. The result also reveals a phenomenon that higher noise energy for the speech noise does not always mean higher WER, which is different from our noise. Instead, the jamming performance of the speech noise will decrease when its noise energy is above certain threshold. The corresponding SNR thresholds are about 1, -3, and -4.8 for the noise containing 1, 2, and 3 speech series respectively. This phenomenon indicates that compared to our noise, the speech noise is less applicable in the real-world scenario as it is hard to control the noise energy to stay within a specific range.

F. Impact of the Number of Phoneme Series

Here we investigate the impact of the number of phoneme series (S_1, S_2, and S_3) in our noise under different SNRs. We first test the effectiveness of different combinations of them and the results are shown on the left of Figure 23 (More results are presented in Appendix-C). The results show that the gap between different combinations is narrow. Generally, less series means better effectiveness (except for one series case). We think this is because the less series, the more energy can each phoneme series share.

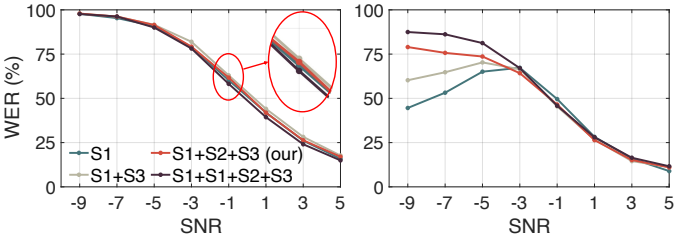


Fig. 23: Impact of the number of phoneme series. Left: WER before noise reduction. Right: WER after noise reduction.

We then further test the robustness of speech noise and our noise. We suppress the noises with speech separation and enhancement models as stated in VII-E and the results are presented on the right of Figure 23. We find that as opposed to effectiveness test, more series indicates higher robustness, which means there is a trade-off between effectiveness and robustness. Our noise ($S_1 + S_2 + S_3$) is relatively balanced between these two factors. Others can choose the combination arbitrarily according to application requirements.

G. Comparisons with Other Jamming Methods

Here we compare our noise with the counterparts in two other related works, namely Backdoor [40] and Patronus [28], and a commercial off-the-shelf device [3]. The noise in [28] consists of dynamic frequencies and chirps. The range of frequency is $[50, 40k]$ Hz and the duration of signal in each frequency is 0.2s. The noise in [40] is a band-limited (i.e., $[0, 12k]$) white noise modulated by a 40 kHz carrier. As for the commercial device [3], we do not have the knowledge of its internals, therefore we can only speculate that its noise is made of variable multi-frequency tones based on the spectrogram. We conduct this experiment in the real-world scenario. We play audios with a smartphone and play noises with our transmitter (except for [3]) with constant power in the meantime. Then we record the audio with one smartphone placed at different distances from the transmitter. The recordings are enhanced with different methods before being fed into an ASR for better recognition (the same process as VII-E for our noise and FullSubNet+ [9] for others). The results in Figure 24 show that our noise performs better than others significantly.

H. Impact of Data Augmentation Process

Here we evaluate the impact of the data augmentation process on the effectiveness of our noise. In the noise generation process, before concatenating a new selected phoneme data in to the noise, for each augmentation method, the phoneme data would be augmented with a probability p. Then we test the jamming effectiveness of the noise generated with different p. The results in Figure 25 show that as p increases, the jamming effectiveness of our noise gradually decreases. However, the impact is limited, with only a drop of 5% in WER when p increase form 0 to 0.5.

I. Impact of Recording Devices

To test the generalizability of InfoMasker among different recording devices, we use different appliances to record demodulated signals and them calculate their energy. We

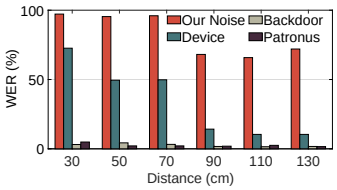


Fig. 24: Comparisons with other existing methods

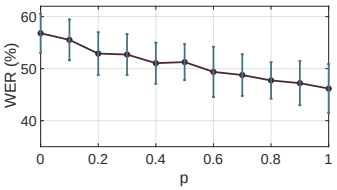


Fig. 25: Impact of data augmentation

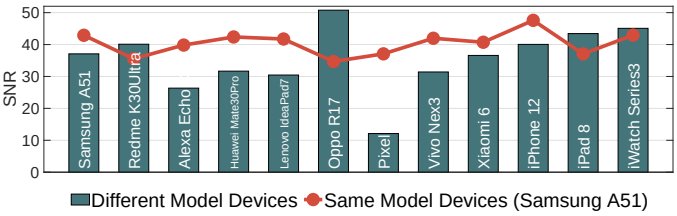
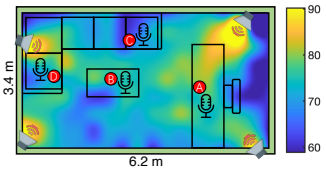


Fig. 26: Nonlinearity in different recording devices

(a) The office.



(b) Energy Distribution.

Fig. 27: Experiment Settings.

Types	WER(%)			
	Phone A	Phone B	Laptop	iPad
A	98.0	98.2	95.7	99.3
B	98.8	98.4	88.1	93.8
C	98.5	56.4	95.8	98.6
D	95.7	97.7	97.9	95.3
Amplifiers On	25.8	26.3	32.5	32.0
Clear	16.0	7.1	19.9	15.5

TABLE VI: Recognition results for the case study.

test a variety of recording devices, including nine models of smartphones, an iPad, a smart watch, a smart home device and a laptop. Besides, we also test the nonlinearity of twelve smartphones of the same model (Samsung A51). The results in Figure 26 shows the good generalizability of InfoMasker.

J. Case Study: A Common Office

Environment Setting. We further explore the possibility of deploying our system in a real-world environment. We deploy our system in a common office room, which is 6.2 meters long and 3.4 meters wide, as shown in Figure 27a. We place a transmitter array in each of the four corners of the room. We first measure the ultrasound energy distribution in the room, and the result is shown in Figure 27b.

Please note that except some areas within 10cm from the transmitter array, where the ultrasound energy reaches about 105 dB SPL, the energy in all other areas is lower than 95 dB SPL. This energy distribution meets the suggestion proposed by World Health Organization (WHO) that when humans are exposed to 40kHz ultrasound for more than 4 hours, the energy of the ultrasound should not exceed 110 dB SPL [2].

Recognition Results. We place four devices in the positions highlighted in red in Figure 27b. The noise energy in Point A and B is relatively high and the energy in Point C and D is relatively low. The devices placed include two smartphones (Samsung A51 and Huawei Mate30Pro), a laptop (Lenovo IdeaPad7), and an iPad (iPad8). For each device we record about 3 hours data and the result is shown in Table VI. We use three commercial ASR systems to recognize each data and choose the lowest WER as the result. For the reason that the power amplifiers will generate audible noise when turned on, we also test the scenario where the power amplifiers are turn on but no noise is transmitted.

Blind Signal Separation. Considering the recording device may have multiple microphones or there are multiple devices recording at the same time, the attacker could use BSS methods to denoise the recordings. In the case study, the recordings from the Huawei Mate30Pro have two channels, so we test the effectiveness of BSS on these recordings. We test five BSS algorithms: AuxIVA [43], ConsistentILRMA [24], FastMNMF [44], LaplaceFDICA [42], and t-ILRMA [32]. The results are shown in Table VII. It can be noticed that even at position C, where the noise energy is the lowest, BSS still cannot improve the recognition accuracy.

Position	WER (%)				
	AuxIVA	Consistent-t-ILRMA	FastMNMF	FDICA	ILRMA
A	98.6	98.6	98.5	98.8	98.5
B	98.4	98.4	98.8	98.8	98.4
C	67.8	67.4	72.0	69.2	67.4
D	97.9	97.9	99.6	97.9	97.9

TABLE VII: Recognition results after BSS.

VIII. Discussion

A. Design of Registration-Free System

The current system requires user registration before usage and the performance of multi-user scenario is about 7% worse than the single-user scenario. Besides, the registration for multi-user scenario is time-consuming. Therefore, We would like to optimize this process to relieve users from registration and to enable InfoMasker dynamically detect current speaker and transmit corresponding noises. However, the optimization process is challenging as when which speaker will start speaking is unpredictable. Additionally, our system has to extract speaker feature from a severely obscured speech signal as the jamming noise is "always on".

In our first step in addressing this challenge, we attempted to use an auto-encoder to recover the target speaker's voice feature from disturbed signals. Via feeding the spectrogram of the jammed speech recording and the ground-truth noise signal in digital domain as two input to the auto-encoder, we tried to obtain the spectrogram of the denoised signal that should be the clean speech signal of current speaker. Figure 28 illustrates

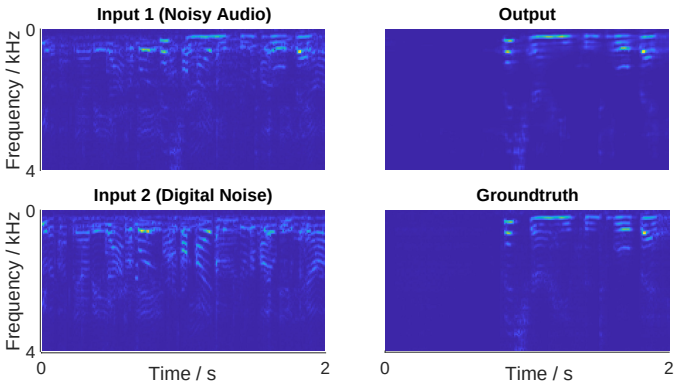


Fig. 28: Illustration of the Denoising Method.

the denoised spectrogram. Although the generated spectrogram is similar to the ground truth of the current speaker, the voice embedding extracted from the spectrogram output cannot be accurately matched to the embedding of that of the speaker.

We conjecture the inaccuracy of embedding calculation is caused by the voice feature extraction algorithm we use. When it is used in our case, the input fed to the algorithm inevitably contains residual. As a result, the performance of the algorithm may suffer because it has never seen such data in its training phase. We would like to verify this hypothesis and develop a robust feature extraction model which can accurately calculate the voice feature of target speakers under the interference from various kinds of noises including common ones and even our phoneme-based noise. By incorporating this component, InfoMasker can accomplish real-time user detection and achieves better jamming performance. Besides, according to the results in Figure 28, it is possible to remove our noise with the ground truth digital noise, which indicates the legitimate users launching the jamming could recover the speech from the jamming interference given the noise template. We leave this part to future work.

B. Non-Line-of-Sight Devices

Although we do not consider the NLOS scenario as stated in Section III, here we conduct an experiment to test the energy attenuation of audible sound and ultrasound when they travel through different types of material to simulate the NLOS scenario. We place the recorder in a soundproof box and replace one side of the box with different materials. The results in Table VIII show that except for clothes, where there is little attenuation for both audible sound and ultrasound, the attenuation for ultrasound is much greater than audible sound. This is a limitation for ultrasonic-based noise injection because ultrasound will attenuate significantly when traveling through different materials due to its high frequency. In contrast, human audible sound can reach hidden devices easily by reflection, diffraction, and even penetration.

C. Generality Across Devices

While our system is valid against all devices tested in Section VII-I, there could be some devices that can possibly bypass our system. For instance, the resonance frequency of some devices could be shifted away from 40 kHz [55], [40],

Material	Soundproof Box	Clothes	A4 paper	Plastic Bag	Plastic Box	Rubber Mat
Normal Audio	-76.9	-5.6	-28.7	0.89	-29.1	-92.6
Ultrasound	-183.2	1.27	-89.5	-25.3	-61.46	-165.2

TABLE VIII: Energy attenuation in NLOS scenario (dB)

which makes the energy of the injected noise relatively low. Besides, iPhone 6 Plus could resist ultrasound signals because of the poor nonlinearity in its microphone [55]. For the former, one possible solution is to adjust the carrier frequency of the ultrasonic noise to different resonance frequencies accordingly, but this would make the ultrasonic noise less effective across devices. In this paper, we choose 40 kHz as the carrier frequency because of its broad applicability [55], [40]. For the latter, although it is difficult to jam the phone model with ultrasonic noise. Some other methods, such as Electromagnetic Interference (EMI), may be effective.

D. Deployment of Transmitter Arrays

In a practical application scenario, the deployment of the transmitter array is relatively convenient as the probability of destructive interference happening is neglectable. Our noise in each location is a superposition of the ultrasounds from different transmitters with various phases, the short wavelength fact and over-complex phase of these noises result in a slim chance of neutralization. So we only need to choose an appropriate number of arrays and spread them in the target room to satisfy the coverage requirement.

E. Portable Implementation

Currently, the implementation of our system is relatively heavy-weight. However, it is possible for implementing a portable version of the system. The laptop can be replaced by a Raspberry Pi with a external soundcard; The power amplifier can be replaced by a smaller-size one (the current one is power excessive); the half-spherical foam can be replaced by a smaller hollow sphere containing all parts including a battery in it. This part is left for future work.

IX. Related Work

Preventing Eavesdropping with Microphone's Nonlinearity. Several existing works attempt to jam microphones using ultrasound. Roy et al. inject white noise to microphones using inaudible ultrasound [40]. Li et al. generate noise with variable frequency according to a preset key to prevent the unauthorized devices from recording audios, while enabling the authorized users to recover speech signals from the noisy recordings [28]. Sun et al. propose MicShield, which can prevent the always-on microphones in smart home devices from recording private speech while passing the preset voice commands [46]. Chen et al. integrate ultrasonic transmitters into a wearable bracelet to expand the effective jamming coverage [10]. However, the noises used in these works are either single tones with variable frequency or white noise, which are not robust when facing speech enhancement methods as demonstrated by the experiments in this paper.

Other Applications with Microphone's Nonlinearity. Many works explore the microphone nonlinearity for other purposes, such as inaudible voice commands injection [55], [41], [52], defense of inaudible commands [21], [54], [41], communication [40], and authentication [56]. Yan et al. transmit the inaudible voice commands through solid medium to improve the stealthiness of the attack [52]. Roy et al. expand the attack range by striping different frequency bands to different transmitters [41]. Zhang et all. exploit the difference of propagation attenuation between ultrasound and human voice to distinguish inaudible commands injection from normal voice commands. Zhou et al. validate that the parameters of the nonlinearity model in different microphones can be used as features for device authentication [56].

X. Conclusion

In this paper, we propose InfoMasker, a highly effective anti-jamming system. By exploring informational masking effect, we achieve phoneme-based jamming noise design for the first time. Our noise exhibits strong ability to interfere with both ASRs and human auditory system, and it shows robustness against noise removal methods. Moreover, our system optimizes the conventional ultrasonic transmission by using lower-sideband modulation and integrating pre-compensation. Digital-domain experiments show our noise can significantly obstruct the recognition accuracy of SOTA ASRs to below 50% with an SNR of 0, which is far better than the jamming performance of white noise. Our case study further validates the effectiveness of our noise in real-world scenario. When the noise energy is acceptable, the recognition accuracy of jammed speech signals are all below 50% in every tested cases, even less than 10% for certain situations.

Acknowledgment

This work is supported by National Key R&D Program of China (Grant No. 2020AAA0107700), National Natural Science Foundation of China (No. 62172359, 62032021, 61972348, 62102354, 62227805, 62072398, 61772236), Funding for Postdoctoral Scientific Research Projects in Zhejiang Province (ZJ2021139), Fundamental Research Funds for the Central Universities (No. 2021FZZX001-27), Research Institute of Cyberspace Governance in Zhejiang University, National Key Laboratory of Science and Technology on Information System Security (6142111210301), and State Key Laboratory of Mathematical Engineering and Advanced Computing.

References

[1] "IEEE recommended practice for speech quality measurements," *IEEE Transactions on Audio and Electroacoustics*, vol. 17, no. 3, pp. 225–246, 1969.

[2] "Environmental health criteria – ultrasound," https://apps.who.int/iris/bitstream/handle/10665/37263/9241540826-eng.pdf?sequence=1&isAllowed=y, 1982.

[3] "A typical commercial device," https://detail.tmall.com/item.htm?spm=a230r.1.14.1.323043c3LS6BEq&id=665020600411&ns=1&abbucket=7&skuId=4793894240789, 2022.

[4] "Ultrasonic transducer (nu40c16t-1), jinci technology," http://www.jinci.cn/index.php?c=article&id=133, 2022.

[5] Amazon, "Amazon Transcribe," https://aws.amazon.com/transcribe/, 2022.

[6] R. Baken and R. Orlikoff, *Clinical Measurement of Speech and Voice*, ser. Speech Science. Singular Thomson Learning.

[7] D. S. Brungart, "Informational and energetic masking effects in the perception of two simultaneous talkers," *The Journal of the Acoustical Society of America*, vol. 109, no. 3, pp. 1101–1109, Mar. 2001.

[8] G. K. C. Chen and J. J. Whalen, "Comparative rfi performance of bipolar operational amplifiers," in *1981 IEEE International Symposium on Electromagnetic Compatibility*, 1981, pp. 1–5.

[9] J. Chen, Z. Wang, D. Tuo, Z. Wu, S. Kang, and H. Meng, "Fullsubnet+: Channel attention fullsubnet with complex spectrograms for speech enhancement," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7857–7861.

[10] Y. Chen, H. Li, S.-Y. Teng, S. Nagels, Z. Li, P. Lopes, B. Y. Zhao, and H. Zheng, "Wearable microphone jamming," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, ser. CHI '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 1–12.

[11] J. Cosentino, M. Pariente, S. Cornell, A. Deleforge, and E. Vincent, "Librimix: An open-source dataset for generalizable speech separation," 2020.

[12] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal Processing Magazine*, vol. 18, no. 1, pp. 32–80, 2001.

[13] F. Developers, "ffmpeg tool [software]," http://ffmpeg.org, 2016.

[14] J. M. Festen and R. Plomp, "Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing," *The Journal of the Acoustical Society of America*, vol. 88, no. 4, pp. 1725–1736, Oct. 1990.

[15] I. Fónagy, "A new method of investigating the perception of prosodic features," *Language and Speech*, vol. 21, no. 1, pp. 34–49, 1978.

[16] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "Timit acoustic-phonetic continuous speech corpus ldc93s1," 1993.

[17] Google, "Google Speech-to-Text," https://cloud.google.com/speech-to-text/, 2022.

[18] T. Guardian, "Ukraine prime minister offers resignation after leaked recording," https://www.theguardian.com/world/2020/jan/17/ukraine-prime-minister-oleksiy-goncharuk-offers-resignation-after-leaked-recording, 17-Jan-2020.

[19] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates, and A. Y. Ng, "Deep speech: Scaling up end-to-end speech recognition," 2014.

[20] X. Hao, X. Su, R. Horaud, and X. Li, "Fullsubnet: A full-band and sub-band fusion model for real-time single-channel speech enhancement," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6633–6637.

[21] Y. He, J. Bian, X. Tong, Z. Qian, W. Zhu, X. Tian, and X. Wang, "Canceling inaudible voice commands against voice control systems," in *The 25th Annual International Conference on Mobile Computing and Networking*, ser. MobiCom '19. Association for Computing Machinery, pp. 1–15.

[22] IFLYTEK, "Xunfei ASR," https://www.xfyun.cn/services/lfasr, 2022.

[23] B. Karmann and T. Knudsen, "Project Alias," https://github.com/bjoernkarmann/project_alias, 30-Jun-2019.

[24] D. Kitamura and K. Yatabe, "Consistent independent low-rank matrix analysis for determined blind source separation," *Research Square*, Jul. 2020.

[25] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 5220–5224.

[26] Z.-Q. Lang and S. Billings, "Evaluation of output frequency responses of nonlinear systems under multiple inputs," *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing*, vol. 47, no. 1, pp. 28–38, Jan. 2000.

[27] M. R. Leek, M. E. Brown, and M. F. Dorman, "Informational masking

and auditory attention," *Perception & Psychophysics*, vol. 50, no. 3, pp. 205–214, May 1991.

[28] L. Li, M. Liu, Y. Yao, F. Dang, Z. Cao, and Y. Liu, "Patronus: Preventing unauthorized speech recordings with support for selective unscrambling," in *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*, ser. SenSys '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 245–257.

[29] S. L. Mattys, L. M. Carroll, C. K. W. Li, and S. L. Y. Chan, "Effects of energetic and informational masking on speech segmentation by native and non-native speakers," *Speech Communication*, vol. 52, no. 11, pp. 887–899, Nov. 2010.

[30] M. Moirise, F. Yokomori, and K. Ozawa, "World: A vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Transactions on Information and Systems*, vol. E99.D, no. 7, pp. 1877–1884, 2016.

[31] T. Moynihan, "Alexa and google home record what you say. but what happens to that data?" https://www.wired.com/2016/12/alexa-and-google-record-your-voice/, 2016.

[32] T. Nakashima, R. Scheibler, Y. Wakabayashi, and N. Ono, "Faster independent low-rank matrix analysis with pairwise updates of demixing vectors," in *2020 28th European Signal Processing Conference (EUSIPCO)*, 2021, pp. 301–305.

[33] B. News, "Gavin williamson interrupted by siri during commons statement," https://www.bbc.com/news/av/uk-politics-44701007, 03-Jul-2018.

[34] T. G. News, "Nsa monitored calls of 35 world leaders after us official handed over contacts," https://www.theguardian.com/world/2013/oct/24/nsa-surveillance-world-leaders-calls, 25-Oct-2013.

[35] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.

[36] Paranoid, "Paranoid Home Devices - HomeWave." https://paranoid.com/products, 2020.

[37] A. Police, "Google is permanently nerfing all home minis because mine spied on everything i said 24/7 [update x2]," https://www.androidpolice.com/2017/10/10/google-nerfing-home-minis-mine-spied-everything-said-247/, 2017.

[38] I. Pollack, "Auditory informational masking," *J. Acoust. Soc. Am.*, vol. 57, no. S1, pp. S5–S5, Apr. 1975.

[39] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong, J.-C. Chou, S.-L. Yeh, S.-W. Fu, C.-F. Liao, E. Rastorgueva, F. Grondin, W. Aris, H. Na, Y. Gao, R. D. Mori, and Y. Bengio, "SpeechBrain: A general-purpose speech toolkit," 2021.

[40] N. Roy, H. Hassanieh, and R. Roy Choudhury, "Backdoor: Making microphones hear inaudible sounds," in *Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services*, ser. MobiSys '17. New York, NY, USA: Association for Computing Machinery, 2017, p. 2–14.

[41] N. Roy, S. Shen, H. Hassanieh, and R. R. Choudhury, "Inaudible voice commands: The Long-Range attack and defense," in *15th USENIX Symposium on Networked Systems Design and Implementation (NSDI 18)*. Renton, WA: USENIX Association, Apr. 2018, pp. 547–560.

[42] H. Sawada, S. Araki, and S. Makino, "Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 3, pp. 516–527, 2011.

[43] R. Scheibler and N. Ono, "Fast and stable blind source separation with rank-1 updates," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 236–240.

[44] K. Sekiguchi, A. A. Nugraha, Y. Bando, and K. Yoshii, "Fast multichannel source separation based on jointly diagonalizable spatial covariance matrices," in *2019 27th European Signal Processing Conference (EUSIPCO)*. IEEE, Sep. 2019.

[45] C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi, and J. Zhong, "Attention is all you need in speech separation," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Jun. 2021.

[46] K. Sun, C. Chen, and X. Zhang, ""Alexa, stop spying on me!": speech privacy protection against voice assistants," in *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*. New York, NY, USA: Association for Computing Machinery, Nov. 2020, pp. 298–311.

[47] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2010, pp. 4214–4217.

[48] Tencent, "Tencent ASR," https://cloud.tencent.com/product/asr, 2022.

[49] P. Walker and N. Saxena, "Evaluating the effectiveness of protection jamming devices in mitigating smart speaker eavesdropping attacks using gaussian white noise," in *Annual Computer Security Applications Conference*, ser. ACSAC. New York, NY, USA: Association for Computing Machinery, 2021, p. 414–424.

[50] L. Wan, Q. Wang, A. Papir, and I. L. Moreno, "Generalized end-to-end loss for speaker verification," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 4879–4883.

[51] C. Wueest, "Everything you need to know about the security of voice-activated smart speakers. symantec." https://www.symantec.com/blogs/threat-intelligence/security-voice-activated-smart-speakers, 2017.

[52] Q. Yan, K. Liu, Q. Zhou, H. Guo, and N. Zhang, "SurfingAttack: Interactive hidden attack on voice assistants using ultrasonic guided waves," in *Proceedings 2020 Network and Distributed System Security Symposium (NDSS)*. Internet Society, pp. 1–18.

[53] Z. Yao, D. Wu, X. Wang, B. Zhang, F. Yu, C. Yang, Z. Peng, X. Chen, L. Xie, and X. Lei, "Wenet: Production oriented streaming and non-streaming end-to-end speech recognition toolkit," in *Proc. Interspeech*. Brno, Czech Republic: IEEE, 2021.

[54] G. Zhang, X. Ji, X. Li, G. Qu, and W. Xu, "Eararray: Defending against dolphinattack via acoustic attenuation," *Proceedings 2021 Network and Distributed System Security Symposium*, 2021.

[55] 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*, ser. CCS '17. New York, NY, USA: Association for Computing Machinery, 2017, p. 103–117.

[56] X. Zhou, X. Ji, C. Yan, J. Deng, and W. Xu, "NAuth: Secure face-to-face device authentication via nonlinearity," in *IEEE INFOCOM 2019 - IEEE Conference on Computer Communications*. IEEE, Apr. 2019.

APPENDIX

A. Proof of Single-Sideband Modulation

Assume the high-frequency carrier signal is $c(t) = cos(2\pi f_c t)$ and the noise signal is $n(t)$, the DSB-AM and LSB-AM can be represented by Equation 2 and 3:

$$n_D(t) = \sqrt{2}n(t)cos(2\pi f_c t) \tag{2}$$

$$n_L(t) = n(t)cos(2\pi f_c t) + \widehat{n}(t)sin(2\pi f_c t) \tag{3}$$

where f_c is the carrier frequency, and $\hat{n}(t)$ is the Hilbert transform of $n(t)$. The coefficient $\sqrt{2}$ in Equation 2 is to ensure the energy of the two modulated signals are same.

Similar to microphones, the nonlinearity in the ultrasonic transmitter also generates high order components. Because energy decays significantly as the order goes up, we only consider the quadratic term here. With DSB-AM and LSB-AM modulation, the quadratic terms of the modulated signal can be represented by Equation 4 and 5. As shown in these two equations, the human audible low-frequency components for DSB-AM is $n^2(t)$, while for LSB-AM it is $\frac{1}{2}(n^2(t) + \widehat{n}^2(t))$. Because the Hilbert transform imparts a phase shift of $\pm\frac{\pi}{2}$ to each frequency components, the amplitude of each frequency

components in the former is $\sqrt{2}$ times bigger than the latter. That is, the former has twice the energy of the latter.

$$\begin{aligned} n_D^2(t) =& n^2(t)(1+cos(4\pi f_c t)) \\ & \xrightarrow{Lowpass\ Filter} n^2(t) \end{aligned} \tag{4}$$

$$\begin{aligned} n_L^2(t) =& 0.5(n^2(t)+\widehat{n}^2(t)) + n(t)\widehat{n}(t)sin(4\pi f_c t) + \\ & 0.5(n^2(t)cos(4\pi f_c t) - \widehat{n}^2(t)sin(4\pi f_c t)) \\ & \xrightarrow{Lowpass\ Filter} 0.5(n^2(t)+\widehat{n}^2(t)) \end{aligned} \tag{5}$$

B. Training Strategies of Speech Separation Models

For each model, the training dataset is chosen from the combination of LibriSpeech train-clean-100 and train-clean 360, and the test dataset in all the experiments involving speech separation is chosen from LibriSpeech test-clean.

In this paper, we consider three speech separation models targeting at audios containing 2, 3, and 4 sources respectively. Use n to represent the number of sources, the model is first pretrained on the n-source LibriMix dataset [11]. The model is then fine-tuned on a modified n-source LibriMix dataset, in which the training data is a combination of n audios from the same speaker.

C. More Results of Different Number of Phoneme Series

The result of more types of phoneme series combination is in Table IX and Table X.

SNR	S1	S2	S1+S3	S2+S3	S1+S2+S3	S1+2S2+S3	2S1+S2+S3
-9	97.7400	95.8100	97.9000	96.8200	97.8800	97.3900	97.6500
-7	95.3400	91.2500	96.1300	93.7100	96.2200	95.7800	96.2400
-5	90.8100	82.8700	91.5700	87.5600	91.4300	90.1400	90.0000
-3	78.9300	68.0930	81.9700	75.4300	79.2100	77.7390	78.1500
-1	60.2400	52.3000	62.9800	57.9100	61.7100	57.2700	58.1720
1	41.9420	37.8900	44.0600	41.4300	41.9100	39.1800	39.3600
3	26.0800	24.7460	28.2200	28.5600	26.3120	24.1200	24.0500
5	15.9000	17.4000	17.7000	18.9400	17.0100	14.9400	14.9600

TABLE IX: WER (%) of the Recognition result of different number of series w/o noise reduction.

SNR	S1	S2	S1+S3	S2+S3	S1+S2+S3	S1+2S2+S3	2S1+S2+S3
-9	44.5600	17.6500	60.2400	39.8400	78.9800	81.7400	87.4700
-7	53.1900	21.2200	64.7500	38.8400	75.6900	79.9900	86.1800
-5	65.0700	31.2300	70.3000	45.0500	73.6400	76.8100	81.2300
-3	67.2800	42.7900	66.4500	47.8300	64.2200	64.6600	67.0580
-1	49.6000	41.3400	45.4400	43.7100	46.5000	45.9000	45.8100
1	28.2400	26.7900	26.7200	27.2500	26.4000	28.4500	28.0100
3	15.9700	16.2900	14.5400	15.9500	15.1700	16.5900	16.4500
5	8.9000	9.2700	11.2100	10.5200	11.0000	10.9100	11.6200

TABLE X: WER (%) of the Recognition result of different number of series w/ noise reduction.