基于超声波非线性的汉语语音防窃听方法

高 铭[1],陈奕可[2],陈佳彤[2],肖 甫[1*],韩劲松[2]
(1. 南京邮电大学计算机学院/软件学院/网络空间安全学院,江苏南京 210023;
2. 浙江大学计算机科学与技术学院,浙江杭州 310007)

摘 要: 语音隐私安全对于国家和个人信息安全至关重要. 为了确保用户的语音隐私免受窃听,超声波录音干扰技术被广泛采用. 此技术利用电子录音设备中的超声波非线性特征,在不影响正常交流的前提下,实现了高效且低成本的窃听录音干扰. 然而,现有的语音防窃听技术仍存在安全隐患. 由于以往技术仅采用简单的噪声掩盖技术,窃听者能够通过先进的去噪技术恢复语音信息,威胁语音隐私安全. 特别地,这些方法的设计主要针对英语语音,对汉语语音的适用性有限. 因此,针对汉语语音的隐私保护需求更为迫切. 为提高超声波录音干扰的安全性和适用性,本文针对汉语语音隐私保护,设计了一种安全稳健的防窃听方法. 本文分析汉语语音独特特征,以此为基础设计一种耦合噪声生成算法,该算法所生成的超声干扰噪声与汉语用户的语音信号紧密耦合、高度相关,因此难以分离,能够有效抵御各种去噪手段. 本文充分考虑了窃听者的能力,实现不可恢复的录音干扰,在不影响用户听力及正常交流的情况下,构建了安全的防窃听方案,全面保护用户的语音隐私安全. 为验证该方法的有效性,本文设计了超声波录音干扰原型系统. 实验结果表明,在 6 m 的范围内,本文方法能够确保 90% 以上的用户语音内容无法被窃听者识读,为汉语语音隐私保护提供了强有力的技术支持.

关键词: 语音安全与隐私;物联网安全;移动感知与安全;汉语语音;超声波非线性
基金项目: 国家杰出青年科学基金(No.62125203);国家自然科学基金(No.62372400)
中图分类号: TP393 **文献标识码:** A **文章编号:** 0372-2112(2025)03-0986-14
电子学报 URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20240237

Microphone Jamming Against Eavesdropping on Chinese Based on Ultrasonic Non-Linearity

GAO Ming[1], CHEN Yi-ke[2], CHEN Jia-tong[2], XIAO Fu[1*], HAN Jin-song[2]
(1. *School of Computer Science*, *Nanjing University of Posts and Telecommunications*, *Nanjing*, *Jiangsu* 210023, *China*;
2. *School of Computer Science and Technology*, *Zhejiang University*, *Hangzhou*, *Zhejiang* 310007, *China*)

Abstract: The privacy and security of speech are fundamental to both national and personal information security. To protect users' speeches from being eavesdropped on, ultrasonic microphone jammers are widely utilized. These jammers utilize the nonlinear characteristics of ultrasound in digital recording devices to inject noise into microphones efficiently and cost-effectively, without disrupting normal communication or human hearing. However, existing microphone jammers are vulnerable. They merely introduce simple noise to mask speeches. As a result, eavesdroppers can employ advanced denoising techniques to recover speech information, posing a significant threat to speech privacy and security. Moreover, existing jammers have primarily been designed for English speech, limiting their applicability to Chinese speech. Therefore, there is an urgent need for privacy protection for Chinese speech. To enhance the security and adaptability of ultrasonic microphone jammers, this paper introduces a robust jammer for Chinese speech privacy protection. Based on the unique characteristics of Chinese phonetics, we design a coherent noise generation algorithm, which produces real-time ultrasound noise intimately coupled with the protected speech signal. This noise is designed to be difficult for adversaries to separate from the speech, ensuring that any attempts at eavesdropping will be frustrated. Comprehensively considering the capabilities of the potential adversaries adversary, our proposed jammer realizes the robust protection against eavesdropping. The generated noise cannot be removed by adversaries using state-of-the-art denoising techniques and is imperceptible to human hearing. Thereby, we

收稿日期:2024-03-14;修回日期:2024-09-16;责任编辑:李勇锋
*通讯作者:肖甫

comprehensively safeguard speech privacy and security. We develop a prototype of the proposed ultrasonic microphone jammer to validate its effectiveness. Experimental results demonstrate that over 90% of protected speeches remain unrecognizable to adversaries within a range of 6 meters under the protection of the proposed jammer, even if the adversary adopts state-of-the-art denoising techniques. Therefore, we provide robust technical support to protect Chinese speech privacy.

Key words: speech security and privacy; internet of things security; mobile security; Chinese phonetics; ultrasonic non-linearity

Foundation Item(s): National Science Fund for Distinguished Young Scholars of China (No.62125203); National Natural Science Foundation of China (No.62372400)

1 引言

窃听行为作为间谍活动的主要手段之一,对国家、社会和个人的信息与隐私安全构成了严重威胁. 随着科技迅猛发展,窃听设备逐渐微型化,使得窃听行为更加难以被察觉. 在智能设备普及的背景下,利用智能手机、智能手表、智能扬声器系统(智能音箱)等设备开展非法录音的行为愈发猖獗. 例如,苹果手机语音助手Siri曾被报道出在未经许可的情况下采集周边用户语音信息[1]. 此外,部分恶意软件试图窃取智能手机、智能手表等移动设备中传声器(麦克风)权限,进行非法监听[2]. 同时,智能扬声器系统存在安全漏洞[3],可能引发远程窃听等隐私隐患. 据统计,2022年全球智能扬声器系统销售总量高达1.466亿台,中国出货量为2 631万台[4],这些数量繁多的"窃听器"已然成为隐私安全的重大威胁. 数据安全与用户隐私问题在2023年全国两会中被多次提及,防止语音窃听与非法录音成为保障国家和个人信息安全的重要议题.

随着对语音隐私保护需求的不断增长,各种防窃听技术应运而生. 其中,超声波录音干扰技术凭借其独特的优势,例如:人耳无法感知、适用设备广泛、覆盖范围大以及部署成本低等,受到了研究者和消费者市场的广泛关注[5]. 超声波防窃听技术利用传声器(麦克风)的非线性特性[6],使用人耳听不见的超声波信号在电子录音设备中产生低频噪声,从而有效抵御窃听行为,保护语音隐私. 由于电子录音设备中的传声器普遍存在非线性特性,高频超声波信号超出其频率响应范围时会发生失真,进而畸变为可听声频段上的低频噪声. 通过精心设计与调制超声波信号,能够高效干扰电子录音设备,防范窃听行为的发生和危害[7].

在实际应用中,超声波防窃听技术面临着一大挑战:在真实的窃听环境中,窃听者并不会轻易放弃窃听行为,他们会采取各种技术手段从受干扰的录音中恢复语音信息. 然而,大多数防窃听方案都忽略了这一问题,很少考虑到窃听者具备去噪能力的情况,因此缺乏有效的应对措施[7~10]. 这些方案虽然能够运用不同类型的超声调制手段远程、无声地向窃听设备注入大量噪声,例如:高斯白噪声(White Gaussian Noise)、啁啾(Chirp)噪声等,但窃听者仍可以使用先进的去噪技术,例如:滤波降噪技术、盲源信号分离技术(Blind Source Separation, BSS)[11],来降低录音中的噪声干扰,从而恢复原始语音信息,实现语音隐私窃取目的[5]. 研究表明,先进超声波干扰设备(如:MicShield[12])在去噪技术的作用下,仅能保护约25%的语音片段不被识读[5,12],这远不能满足防范窃听的实际需求.

针对这一问题,两篇最新的研究工作[13,14]尝试提供解决方案,但它们在实际应用中存在一定的限制,无法直接适用于汉语对话的日常环境. Gao等人[13]设计的MicFrozen系统能够有效抵御降噪技术,但其需要实时地采集用户语音以生成所需噪声,这种方法主要适用于如会议、演讲等用户面前放置传声器的场景,对于日常生活中的对话情境,其应用受到限制. InfoMasker[14]则主要面向于以英语为代表的印欧语系,未考虑以汉语为代表的汉藏语系的独特特征,难以直接应用于汉语语音保护. 一方面,相较于印欧语系,汉藏语系具有更为丰富的辅音种类,并且在日常发音中辅音的出现频率更高[15],有些字读音甚至完全由辅音构成,例如:"嗯/ŋ/". 这使得在汉藏语系中,基于元音音素片段噪声的干扰方法可能无法达到理想的语音掩蔽效果. 另一方面,汉语中的音调变化,例如:普通话中的阴平、阳平、入声和去声[15,16],是汉语语音中非常重要的特征. 这些音调变化在频域上的体现为频率的起伏,而这些特征很难被不具备类似特征的元音音素片段噪声有效掩盖,这使得窃听者可以利用先进的盲源分离技术[11]高效地去除噪声并恢复汉语语音内容. 因此,针对汉藏语系的语音隐私保护需要开发更为针对性的技术和方法,以提高保护效果.

针对上述存在的问题,本文提出了一种专为汉语语音设计的超声波录音干扰方法. 该方法利用传声器中普遍存在的超声非线性特征,旨在为汉语用户提供一种更有效、更安全的语音隐私保护. 考虑到窃听者可能采用多种去噪手段,本文深入分析了汉语语音的特性,并设计了一种耦合噪声生成算法. 无需实时采集用户语音内容,该算法所生成的超声干扰噪声与待保护语音信号高度相关且难以分离. 其核心原理在于利用并破坏现有去噪算法的一个基本前提:要有效地分离

和提取信号与噪声,两者在时间、频率或空间关系上必须有显著差异. 例如,常见的去噪技术如高通、低通、带通滤波器主要利用信号与噪声在频率上的差异进行分离;而盲源信号分离方法则依赖于待分离信号之间的独立性. 然而,本文设计的耦合噪声与原始语音信号在时域和频域特征上高度相似或重叠,这使得窃听者难以提取有用的语音信息,从而全面保护了用户的语音隐私安全. 此外,为了提升本文方法的实用性和用户体验,本文设计超声载波调制策略,这有效避免了振铃效应等硬件不完善性导致的可听频段噪声产生与泄露,确保用户听觉在使用过程中不会受到任何影响,并兼顾了对不同类型窃听设备的高效噪声注入. 本文开发了一种超声波录音干扰原型系统,实现了大覆盖范围的语音隐私保护. 实验结果表明,在 6 m 的范围内,本文方法能够保护 90% 以上的用户语音内容免受窃听威胁.

2 基础知识与相关工作

在电子录音设备中,传声器(麦克风传感器)系核心环节. 传声器由一个可移动薄膜、一个固定电极或驻极体的空气隙电容及后续模拟电路构成. 当声波抵达并作用于可移动薄膜时,薄膜会因声压变化而发生弯曲. 这种弯曲运动改变了薄膜与固定电极或驻极体之间的距离,使得电容随之变化. 这个过程中,由于电容上的电荷保持恒定,电容变化将导致两端电压变动,进而将声波信号转换为电信号. 经过一系列处理,包括放大、滤波和模数转换,这个模拟电信号最终转化为数字电信号.

在理想的传声器中,其放大器展现出极佳的线性特征,即输出信号与输入信号保持正比关系. 这意味着传声器数字输出信号 $y(t)$ 的相位和频率将与输入信号(原始语音)$x(t)$ 相同,且幅值成比例变化,波形无任何失真,即

$$y(t)=k_1\cdot x(t) \tag{1}$$

其中,k_1 为传声器的线性放大系数. 在商用传声器中,这种线性特征在可听声频段(低于 16 kHz)内得到较好的实现,确保了语音拾取的质量.

然而,传声器在处理超声频段信号时,其内部放大器环节呈现出非线性的特点. 具体来说,输出信号不再是输入信号的简单线性变换,而是包含了输入信号的更高次幂项. 由于高阶(三阶及以上)系数通常极小,因此可忽略不计. 因此,在超声频段上通常呈现为平方(二次)非线性,即

$$y(t)=k_1\cdot x(t)+k_2\cdot x(t)^2 \tag{2}$$

其中,k_i 表示第 i 阶非线性系数. 这种平方非线性的存在意味着,当传声器接收到两个或更多的超声波时,这些超声信号可能会互相作用,并在可听声频段中产生一个额外的信号.

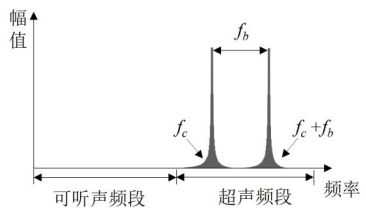
本节以两支单频超声信号作为输入信号为例. 假设这两个信号的频率分别为 f_c 和 f_c+f_b,均超过 20 kHz,那么输入信号可以表示为

$$x(t)=\cos\left(2\pi f_c t\right)+\cos\left[2\pi\left(f_c+f_b\right)t\right] \tag{3}$$

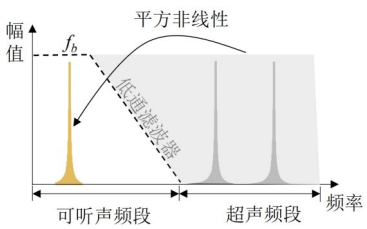
如图 1 所示,平方非线性放大后,高频超声分量会被滤波器消除,传声器的输出只保留了由非线性现象产生的低频可听声信号,即

$$y(t)=k_2\cos\left(2\pi f_b t\right) \tag{4}$$

利用这种超声非线性特性,研究者提出了一种超声波录音干扰方法. 该方法通过将噪声信号调制到超声载波上,能够在不影响人们正常听觉和交流的前提下,向电子录音设备中注入噪声,规避潜在窃听风险. Backdoor[7]测试了四种噪声信号(包括:固定单一频率噪声信号、频率调制噪声信号、幅值调制噪声信号和高斯白噪声),并通过实验验证噪声分布带宽为 8 kHz 的高斯白噪声具有性能最佳的干扰效果. Chen 等人[8]选用[0, 4] kHz 频带内、以 0.45 ms 为跳频周期的跳频(Frequency-hopping)信号作为干扰噪声,并将系统设计为手镯等可穿戴形态. 此外,为了满足仅允许授权设备录音、禁止其他录音设备的现实需求,研究人员提出了选择性干扰机制. MicShield[12]确保智能语音助手正常收听用户合法语音命令的同时,利用超声非线性引入噪声以掩盖非交互式语音,例如:对话、电话通话,防止智能语音助手在用户不知情时窃听隐私语音. 该系统的超声干扰噪声为 4 kHz 带宽的高斯白噪声. Li 等人[9]设计的 Patronus 系统的超声干扰噪声频率在[85, 255] Hz 范围内随机变化. 为了避免“振铃效应”产生可听噪声,该系



(a) 输入:两支超声信号



(b) 输出:非线性解调得到的可听声信号

图 1 超声波非线性特性示意图

统采用扫频啁啾信号来实现频率的平滑变化. Patronus系统通过蓝牙等其他信道向授权设备发送噪声分布频率与特征,允许授权的录音设备分离噪声并记录用于语音信息,同时干扰未经授权的录音设备,有效防止隐私泄露.

然而,上述超声波录音干扰方案常采用与原始语音信号无关的简单噪声信号来掩盖语音信息. 它们往往低估了现实场景中窃听者的真实能力,窃听者可能会采用各种去噪技术来恢复噪声中的语音信息. 随着录音设备抗噪声能力的增强和去噪算法的不断发展,以及各类去噪算法的涌现,真实的窃听者能够有效降低干扰噪声的影响,恢复原始语音信息[5]. 为了更有效地保护语音信息,Gao等人[13]提出了一种实时生成反向信号的方法,该方法根据用户实时采集的语音信号生成反向信号,并调制于超声载波之上. 当携带反向信号的超声干扰信号到达窃听设备时,信号在传声器(麦克风)中发生解调,与用户原始语音信号相互抵消,从而降低窃听设备所能够采集到的语音信号能量,有效抵御窃听者可能采用的降噪技术. 这种方法适用于任何语系的语音保护,但是实时采集用户语音的传声器仅适用于会议等正式场景的部署,在私人交谈等场景中应用受限. InfoMasker[14]面向以英语为代表的印欧语系,利用声音掩蔽效应(Masking Effects),无需实时采集用户语音内容使用户语音难以被辨识. 这种方法利用用户预先录制的语音信息,在语料库中选择与用户音色相近的随机元音音素片段作为干扰噪声,从而增加干扰噪声与原始语音信号之间的相关性. 这种方法可以避免干扰噪声被具有抗噪声功能的录音设备降低噪声强度,提高语音信息的保护效果. 然而,这种方法并不适用于以汉语为代表的汉藏语系. 不同于印欧语系中元音音素对语音语义的表达占到主导地位,汉藏语系中辅音在日常用语中比重较高,对语音语义也有不可忽视的作用. 这在信号层面上的体现为:英语语音的能量大多聚集在[85,255]Hz频段上(元音音素片段主要能量也集中于这个频段),而汉语语音中辅音能量更高,因此在频谱上能量分布更宽[16]. 因此,仅用元音音素片段无法对汉语语音进行足够掩盖. 另一方面,由于汉语中存在语调变化[15],相同音素不同语调具有不同的频率特征[16],所以元音音素片段与汉语语音的相关性较弱,这种噪声模型易被窃听者使用盲源信号分离等技术去除. 由于汉语中截然不同的元音辅音体系与独特的音调变化[15,16],使用元音音素片段作为干扰噪声的方法对于汉语等汉藏语系的语音保护效果有限. 在汉语等汉藏语系的语音保护方面,仍需探索更加针对性的防窃听技术,以应对窃听者的不断发展和变化的技术手段.

3　威胁模型

窃听者的目标是在未经许可的情况下偷录用户语音等声音信息. 他们可能暗中设置各种窃听设备,甚至通过不正当手段获取智能手机等智能设备的传声器访问权限,以实现清晰的窃听.

在现实环境中,窃听者通常具备噪声去除与信息恢复能力. 他们不会因干扰而轻易放弃窃听,反而会利用机器学习算法和先进的噪声处理技术,从充满干扰的录音中提取有价值的语音信息. 此外,窃听者会采取一系列预防措施以降低干扰噪声的影响,如部署多个窃听设备或使用高效降噪技术的窃听设备,提高窃听录音的清晰度. 为应对这一挑战,本节分析窃听者的手段和策略,从而更加全面地认知语音窃听的风险,并在此基础上设计出更加安全可靠的超声波录音干扰方案.

首先,窃听者可以利用信号与噪声在时间分布上差异,在时域上进行信号分离. 其中,盲源信号分离技术[12]是一例典型的时域去噪技术. 即使在源信号信道参数、统计分布等信息均未知的情况下,仅利用源信号之间相互独立这一微弱已知条件,盲源信号分离技术也能够通过接收到的观测信号恢复出源信号. 盲源信号分离技术通常要求使用多个传声器采集多路信号,并利用各路信号之间的独立性(或相关性)作为优化目标,通过不断迭代优化,从混合信号中分离出独立源信号,提取语音信息.

其次,窃听者可以利用信号与噪声在频域特性上差异来降低干扰噪声的影响. 他们利用诸如短时傅里叶变换和离散小波变换等频谱分析技术,深入挖掘并理解噪声的频率特性. 在此基础上,他们会精心设计低通、高通或带通滤波器等频域滤波器,以便更精确地区分语音信号和噪声. 一个常见的策略是选用宽带带阻滤波器或者陷波滤波器,在特定频段或频率点上强度显著的干扰噪声,从而将混合信号中的干扰噪声去除,分离语音信息.

最后,窃听者还可以综合利用时域和频域的多个特征实现信号分离. 例如,窃听者结合频谱嗅探技术与时域自适应滤波技术以实现高效去噪. 首先,在超声频段检测超声干扰信号,并根据非线性特征计算出在可听声频段上的干扰噪声. 随后,窃听者利用时域自适应滤波技术,根据干扰噪声的统计特性自动调整滤波器参数,从而消除干扰噪声.

在现实环境中,窃听者可能会采取多种去噪手段来恢复待保护的语音隐私信息. 由于用户无法预知窃听者可能使用的去噪手段,因此窃听干扰方法必须具备一体化防护能力,综合运用时域和频域噪声生成技术,有效地抵御多种去噪方法.

4 本文方法

本节提出了一种稳健的超声波录音干扰方法,该方法能够生成与待保护语音信号紧密耦合且难以从录音中分离的干扰噪声. 这一设计旨在全方位地抵御窃听者的各种噪声分离手段. 首先,提出了一种基于可微非双射映射的耦合噪声生成算法,加强干扰噪声与语音信号之间的相关性,避免语音信息的泄露与恢复. 然后,使用语音生成技术,通过少量用户注册语音生成大量语料库,并基于汉语语音特征进行预处理,以便无需实时采集用户语音即可生成耦合噪声. 最后,设计超声载波,使之能够有效地向不同类型窃听设备中注入干扰噪声,并避免这个过程中可能会存在的可听噪声泄露现象.

4.1 耦合噪声设计

本节提出了一种基于非线性映射的耦合噪声生成算法. 该算法通过生成具有复杂特性的强噪声,增强了干扰噪声与语音信号在时域和频域等特征上的相关性,避免窃听者从录音中分离语音信号.

首先,利用可微非双射映射混叠技术[17],生成与原始语音信号具有相似的时域特征的噪声信号:

$$n_{\text{time}}(t)=\boldsymbol{M}_{1\times 2}(t)\cdot L\begin{bmatrix}x(t)\\ n_{r1}(t)\end{bmatrix} \tag{5}$$

其中,$\boldsymbol{M}_{1\times 2}(t)$是一个1×2阶不可逆随机矩阵;$x(t)$是语音信号;$n_{r1}(t)$是一个随机噪声;$L[\cdot]$是不可积非线性混叠函数:

$$L[x]=\frac{\left[\sin x-\ln\left(1+\mathrm{e}^{x}\right)\right]}{x} \tag{6}$$

由于可微非双射映射混叠具有多对多映射关系,其逆变换并不唯一. 这一特性使得生成的耦合噪声具有高度的复杂性和稳健性,增加了从语音信号中分离噪声的难度.

其次,本节结合信号频域特征与随机噪声,从频域角度上加强所生成的噪声与语音信号的相关性. 具体地,本节在时域与频域中分别通过卷积运算引入一项随机噪声. 所得耦合噪声为

$$n_{\text{fre}}(t)=n_{r3}(t)*\boldsymbol{M}_{1\times 2}(t)\cdot L\begin{bmatrix}n_{r2}(t)\cdot x(t)\\ n_{r1}(t)\end{bmatrix} \tag{7}$$

其中,$x(t)$为语音信号;$n_{ri}(i=1,2,3)$为互相独立且分布不同的随机噪声;$*$为卷积运算. 需要注意的是,式(7)中$n_{r2}(t)\cdot x(t)$在时域上的点乘运算对应了在频域中的卷积运算. 通过在时域和频率中的卷积操作,所得的噪声与语音信号紧密耦合. 这种噪声与语音信号在频域上完全重叠且互相不独立,因此频域滤波器无法在不破坏语音信号的同时去除噪声信号,而依赖于信号之间的独立性进行分离的独立分量分析等其他频域分离方法也无法奏效.

最后,为了抵御利用时频域特征的超声嗅探等去噪手段,本节利用跳频扩谱技术[18]拓宽耦合噪声的频谱带宽. 所得耦合噪声为

$$n_{\text{mix}}(t)=\cos\left[\omega_n t+\varphi(d,t,\Delta\omega)\right]\cdot n_{\text{fre}}(t) \tag{8}$$

其中,$\cos\left[\omega_n t+\varphi(d,t,\Delta\omega)\right]$为跳频信号;$\omega_n$为跳频频表中第$n(n\in N)$个频点;$n$满足$nT_h\leqslant t(n+1)T_h$;$t$为时间;$T_h$为跳频周期;$\varphi(d,t,\Delta\omega)$为相位调制函数;$d$为差错编码后的信码;$\Delta\omega$为跳频前后的频率差分量. 通过这种方法,耦合噪声$n_{\text{mix}}(t)$的带宽被大幅扩展,经验值设置可使之超过12 kHz. 由于本文不需要对耦合噪声进行解码,因此跳频信号中所有参数均可随机设置,以加大噪声复杂度. 这种情况下,窃听者难以获取干扰噪声的全部特征,故而无法恢复语音信号. 尽管可能存在强大窃听者使用宽通带超声波传声器来探测耦合噪声,但这类设备不仅价格高昂(往往超过十万元人民币),而且体积较大(一般大于50 cm×40 cm×25 cm),易被发现.

加入耦合噪声$n_{\text{mix}}(t)$后,窃听录音设备收到:

$$y(t)=x(t)+n_{\text{mix}}(t) \tag{9}$$

为了恢复语音信号$x(t)$,窃听者需要找到逆函数$G(\cdot)$,使之满足:

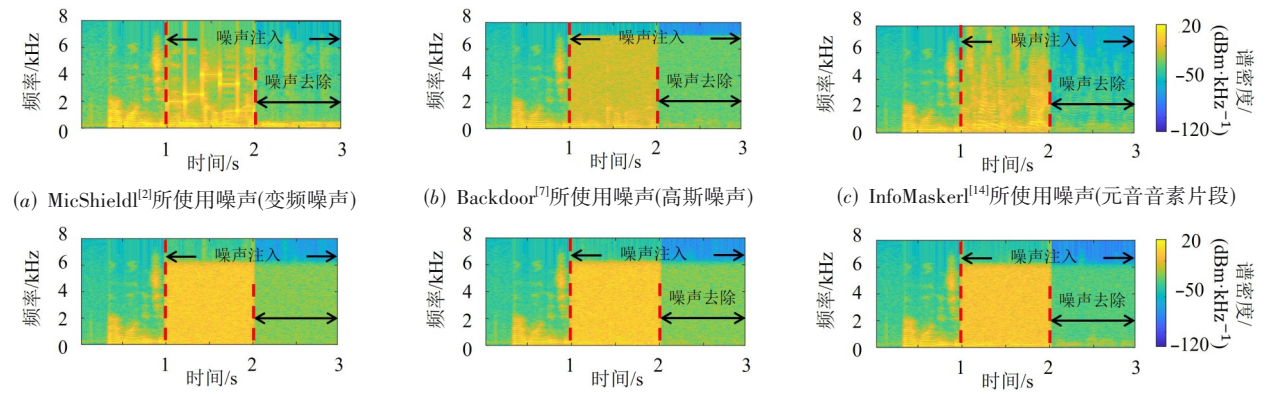
$$G\left(r(t)\right)=\left[x(t)\ n_{\text{mix}}(t)\right]^{\mathrm{T}} \tag{10}$$

得益于耦合噪声$n_{\text{mix}}(t)$,$y(t)$与$x(t)$依然保持可微非双射映射关系. 这导致存在无数个$G(\cdot)$函数满足式(12),即该式有无数解. 由于非线性混叠产生的噪声中存在无穷多的逆函数$G(\cdot)$和无数对$[x(t)\ \ n_{\text{mix}}(t)]^{\mathrm{T}}$可以满足统计独立性的要求[19],因此基于信号独立假设的盲源信号分离技术只能获得局部最优解[11],并且存在无数组局部最优解. 这意味着窃听者无法正确分辨出哪一组解是原始语音信号. 即使使用互信息等技术优化$G(\cdot)$[20],得到的也只是无意义的随机噪声,无法恢复出可辨析的语音信号. 研究表明,通过分析信号的统计学特征并增加时间相关性[17]、正则化[21]或混合模型结构[19]等约束条件,可以一定程度上减少局部最优解的数量. 针对这一现象,本节进一步增加随机噪声$n_{ri}(t)(i=1,2,3)$的复杂度,破坏上述约束条件. 具体来说,本算法每间隔一段时间就会随机改变$n_{ri}(t)$的特性,例如:分布特征和统计规律等. 这种措施将显著增加局部最优解的数量. 在这种情况下,窃听者使用任何去噪技术手段也只能得到严重失真的信号,其中不包含可辨析的语音信息.

为了展示本文设计的耦合噪声在不可分离领域的优异性能,本节与现有文献[7~9,12,14]所使用的噪声进行对比. 本节对现有文献所使用的噪声种类进行分类,并选取三个典型噪声作为对比:(a)变频噪声,噪声能量集中在特定频率或窄频段上,且噪声信号的中心

频率随时间变化,例如参考文献[8]和Patronus[9];(b)高斯噪声,宽带宽的高斯白噪声,例如Backdoor[7]、MicShield[12]和参考文献[10];(c)元音噪声,使用元音音素片段作为干扰噪声,例如InfoMasker[14]. 根据参考文献[5],所使用的干扰噪声的带宽越大,其抗去噪性能更加优异,因此本文选择(a)MicShield[12](使用4 kHz带宽的跳频噪声)、(b)Backdoor[7](使用8 kHz带宽的高斯白噪声)和(c)InfoMasker[14](元音音素片段)作为比较对象,图2(a)~图2(c)展示了原始语音音频信号、加对应噪声后的混合音频信号和使用盲源信号分离技术去噪后的音频信号的频谱图. 本实验中,语音音频内容为汉语普通话"白雪覆盖了一株小草",取自汉语语音数据集ST-CMDS[22],由EDIFIER M230扬声器播放,平均声强为63.1 dB,实验环境背景噪声为49.3 dB. 超声波干扰原型系统(系统详细设计见4.4节)放置于声源正前方5 cm处,发射调制了干扰噪声的超声信号(载波频率为39 kHz,选择依据见4.3节),平均声强为97.7 dB. 智能手机Samsung S8作为窃听录音设备,放置于原型机正前方2 m处. 相同实验条件下,本文设计的耦合噪声性能见图2(d). 图2(d)中耦合噪声系使用带保护语音信号作为$s(t)$并根据式(10)生成. 图2(a)~图2(c)中,所使用的干扰噪声虽然均能够很好地掩盖待保护语音信号,但是它们不能抵御去噪技术. 由此可见,现有的超声波录音干扰技术并不适用于汉语语音防窃听. 相比之下,图2(d)中耦合噪声能够有效防止窃听者恢复语音信息,体现出优异的抗去噪性能.

本段进一步探讨使用何种信号生成的耦合噪声能够具有足够强的抗去噪能力. 最佳情况下,超声波录音干扰方法能够实时获取用户语音内容并生成耦合噪声. 但是这一条件要求在用户身边部署传声器,这一要求显然不适用于私人对话等场景. 为此,本节在上段实验条件下测试:利用同一用户不同语音和不同用户语音所生成的耦合噪声的抗干扰性能,实验结果依次如图2(e)和图2(f)所示. 实验结果表明,根据同一用户语音所生成的耦合噪声能够对该用户语音做到较好的保护. 因此,用户可以在正式使用前注册音色,无需实时获取当前用户语音内容,即可在各类现实场景中抵御窃听者.



(a) MicShieldl[2]所使用噪声(变频噪声)　(b) Backdoor[7]所使用噪声(高斯噪声)　(c) InfoMaskerl[14]所使用噪声(元音音素片段)

(d) 耦合噪声(根据带保护语音信号生成)　(e) 耦合噪声(根据同一用户其他语音生成)　(f) 耦合噪声(根据其他用户语音生成)

图2　干扰噪声(耦合噪声$n_{\mathrm{mix}}(t)$)及其抗干扰性能对比图

4.2　用户语料生成

本节设计一种面向汉语语音用户的语料生成方法,无需实时采集用户语音内容即可实现高度相关的耦合噪声生成. 首先,本节使用语音生成技术扩充用户语料数量,降低用户注册所需时长. 然后,本节分析汉语语音的元辅音系统和语调特征,提高耦合噪声与未知内容的待保护语音之间的相关度,提高其抗去噪性能. 最后,本节设计基于随机音素生成耦合噪声的策略,增加耦合噪声的复杂度,进一步降低耦合噪声被窃听者破解的风险.

为获取用户音色特征的同时兼顾用户友好度,本文方法仅使用短时长的注册数据,基于深度神经网络技术生成并扩充用户语料数量. 本文采用语音生成网络SV2TTS[23],以包含520 000余条汉语普通话语音的语音数据库CN-Celeb2[24]为训练集,生成与注册用户具有相同音色的任意内容的大量语料内容. SV2TTS[23]首先使用GE2E loss函数对用户语音进行聚类,生成代表音色的向量;然后使用带有注意力(Attention)机制的编码解码(Encoder-deconder)网络将编码后的文本数字信息与语音信息建立映射,实现文本转语音(Text-To-Speech,TTS)的任务,输出文本内容对应的具有用户音色的语音信号的梅尔频谱图(Mel-spectrogram);最后使用WaveNet网络将其转换为时序语音信号.

如果具有足够多的用户注册数据,SV2TTS[23]所生成的语音能够与原始用户语音具有极其相似的音色. 然而,在实际使用中,考虑到用户使用的友好度,用户语音音色

注册时间不宜过长. 为了探究注册时长与语音生成效果之间的关系,从而选择合适的用户注册时长,本节测试了不同注册时长下所生成语音与用户原始语音的相似度. 本实验中,SV2TTS[23]生成一条指定内容的用户语音,该语音内容在用户注册过程中并未出现(即不包含于训练集中),并要求用户录制相同内容的语音,与所生成的信号进行比较. 由于用户语音与生成信号直接可能存在语速上的差异,这种语速差异并不能反映音色生成的性能,使用欧氏距离计算相似度有失偏颇,因此本节采用动态时间规整(Dynamic Time Warping,DTW)[25]计算两者之间的相似度. 注册过程中,用户平均语速约为每分钟170个字. 实验结果如图3所示,生成语音与用户语音之间的相似度随着用户注册语音时长的增加而提升. 当用户注册语音时长达到55 s后,DTW相似度高达95%. 当注册时长超过55 s后,随着注册时长继续增加,相似度的增速明显变缓,性能提升有限. 因此,本文方法推荐持续1 min的注册时长,兼顾用户体验的同时保障良好的语音生成性能. 利用语音生成网络SV2TTS[23],本节生成了大量与用户具有相同音色的语料数据,并根据式(10)获得大量耦合噪声.

为了进一步提高基于生成语音信号的耦合噪声与待保护内容之间的相关性,本节着手于汉语语音的元辅音系统和语调特征,设计更有针对性的用户语料. 本节按照音素将所生成的语料内容进行分割,并按照元辅音与音调进行分类,并获得待保护用户各个音素的平均语料. 首先,本节使用最大类间方差法(又称大津法,OTSU)[26]计算语音信号强度阈值,并以此为基础切割音素. 然后,本节使用DTW相似度将各个音素按照10个单元音,20个双元音,22个辅音[15]进行聚类. 考虑到汉语语音中的音调对语音频谱特征影响极大[15,16],本节将相同音素的不同音调视为不同的类别,因此共生成208(=(10+20+22)×4). 此处,类别数量可以根据语言类型进行调整,例如:部分方言中存在双辅音音素,吴语等方言中存在“阴上”“阳上”等音调. 最后,本节对相同音调音素内的语料信号在使用DTW对齐后进行平均,得到各个音素的平均信号.

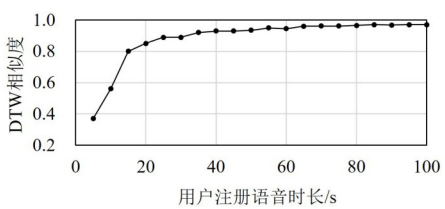


图3 不同注册时长下生成语音性能对比图

为了进一步提高耦合噪声的安全性,本节设计随机选择音素生成耦合噪声的策略. 直接使用基于各音素的耦合噪声存在被窃听者嗅探并去除的可能. 窃听者可能在其他场景中录制用户语言或者使用用户公开的语言信息,采用语言生成网络生成语料库并计算平均音素,以此为参考信号嗅探并使用自适应滤波技术滤除超声噪声. 为了避免这种情况的发生,本节首先对每次用于生成耦合噪声的音素进行随机抽取,每间隔0.5 s随机抽取任意数目的音素生成耦合噪声,并将这些噪声进行卷积操作,作为最后的干扰噪声. 这种策略使得干扰噪声的情况上升至$2^{208}-1$(为10^{62}数量级),极大地增加了噪声的复杂度,使得窃听者无法通过嗅探对比寻找到所对应的音素,因此难以滤除. 此外,本文使用的随机噪声n_{ri}会随着时间变换生成方式、频率分布和统计规律等特征,使得窃听者完全无法从混合信号中分离语言信息.

4.3 超声载波调制

本节设计超声载波调制方案,选择合适的载波频率以实现对任意窃听设备的噪声注入,并调整载波避免在超声发射过程中由于超声波扬声器等发声设备的硬件瑕疵而产生可听噪声.

4.3.1 超声载波频率选择

在现实场景中,用户往往无法了解窃听者所使用的录音设备的相关信息. 因此,本节选择超声载波频率以充分利用超声非线性特征,以对不同录音设备均实现高效的干扰效果.

已有的超声载波设计方案中仅选择载波频率为单一经验值的单频超声波,例如:Backdoor[7]、MicShield[12]和Patronus[9]均选用40 kHz的超声载波;参考文献[8]则设置超声载波频率为25 kHz. 然而,由于不同设备的非线性特征上存在差异性,这种基于经验的超声载波选择方法并不能总是对任意窃听设备保持较佳的干扰效果.

超声非线性现象在电子录音设备中是普遍存在,但它们在不同的频率上的非线性响应强度并不相同,且存在某些特定频率的超声信号使得非线性响应最为显著. 这一现象表明,录音设备的非线性系数会受到载波频率的影响,即式(6)中的第二阶非线性系数k_2取决于载波频率f_c. 在理想情况下,为了最有效地向窃听设备中注入干扰噪声,超声干扰器应发出具有最佳载波频率的超声波. 这一频率应能充分利用录音设备的非线性特性,以最大限度地增强干扰噪声的效果. 然而,由于硬件上的差异,不同录音设备的非线性系数并不相同,导致引发最大非线性响应的超声载波频率也存在差异. 在实际应用中,用户无法提前了解窃听设备的相关参数. 因此,需要设计合适的超声载波频率,以高效地向参数未知的各种录音设备注入耦合噪声.

本节通过大量实验发现,不同传声器之间的非线性系数差异并不明显. 基于这一观察,本实验测量了众多传声器的非线性系数,并计算其平均值作为超声载波频率. 实验涵盖了10个来自Panasonic、Hosiden、

Harman和Bosung等知名品牌的商用传声器模组,以及10种商用录音设备(如智能手机、平板电脑和智能扬声器系统). 实验结果表明,相同采样率设置下(如:分别为48 kHz和96 kHz)的传声器非线性系数非常接近. 综合实验结果,本节选择了两个超声载波频率,以针对不同采样率的窃听设备进行有效干扰:39 kHz用于干扰采样率低于48 kHz的设备,而80 kHz则针对采样率为96 kHz的设备. 本文方法可以灵活采用更高频率的超声载波以应对更高采样率的窃听设备,实现对各类窃听设备的高效干扰. 这种策略简单易行且效果显著,使该方法能够在无需了解窃听设备具体参数的情况下,实现高效的耦合噪声注入和窃听干扰.

载波频率选择也将影响到超声波传播距离,从而决定了录音干扰器有效覆盖范围. 在空气介质中,超声波声学强度随着传播距离d的增加而递减:

$$P(d)=P_0\cdot \mathrm{e}^{-\alpha(f)d} \tag{11}$$

其中,P_0为超声信号在声源处的发射声强;f为超声信号的频率;$\alpha(f)$为衰减系数,其取值取决于超声频率f,具体为

$$\alpha(f)=\frac{2\pi f\rho c}{Q(f)} \tag{12}$$

其中,ρ为介质密度;c为波速;$Q(f)$为取决于超声频率f的品质因数. 经验地,载波频率为25 kHz的超声信号的衰减系数最低. 然而,该载波频率在大多数录音设备中均不能较好地引发超声非线性现象. 相较之下,载波频率为39 kHz的超声信号虽然衰减系数略逊一筹,但能够兼顾较优的有效作用距离,并在更多设备上普适地引发超声非线性现象.

此外,增加超声扬声器的发射功率(即提高P_0)能够有效提高作用范围. 然而,这种方法不仅消耗大量能量,还可能对用户的生理健康产生负面影响. 根据国际非电离辐射委员会建议,超声声压的限值为110 dB[27]. 如果用户长时间暴露在超过110 dB的超声环境中,可能会出现生理不适,甚至引发疾病. 因此,本文设置超声发射功率为100 dB,这一设置兼顾用户生理健康并保障了足够强的超声能量.

4.3.2　可听噪声泄露及解决方案

超声波防窃听技术具备独特的优势在于利用人耳不可听的超声频段上的干扰信号. 然而,在实际应用中,由于发声设备的硬件特性,长时间播放超声信号可能会导致失真,进而产生可听噪声,这限制了超声波录音干扰技术的现实应用. 然而,已有超声载波设计方法中鲜有涉及可听噪声泄露的分析及其解决方案. 本节分析可听噪声的成因,并设计超声载波调制方案以避免可听噪声的产生.

(1)振铃效应

“振铃效应(Ringing Effect)”[7]是导致可听噪声产生的一个重要因素. 当声学信号的频率发生急剧变化时,由于发声设备硬件的不完善性,会在时域中产生一个脉冲波信号. 这个脉冲波在可听声频段具有较高的能量,其声音波形上的振荡类似于铃声,因此被称为“振铃效应”. Patronu[9]采用跳频(Frequency-hopping)信号作为干扰噪声,在频率跳变间人为地引入调频连续信号以平滑突变. 然而,在超声扬声器发声过程中,由于硬件不完美性,可能出现随机的频率突变导致的振铃效应,这个频率突变无法预测或控制,因而无法预先或及时插入调频连续信号以避免振铃效应.

在超声扬声器的发声过程中,带通滤波器被用来截断信号频率,以保障无偏差地产生特定频率的超声信号. 然而,由于带通滤波器的硬件不完美性,在这个过程中可能会发生频率的突变,从而触发了振铃效应的出现. 具体地,理想带通滤波器的频谱是一个矩形窗口函数,其频谱分布特性为:当输入频率ω小于截止频率ω_0时,$X(\omega)=1$,即保留该频段内信号;反之,当输入频率ω大于等于截止频率ω_0时,$X(\omega)=0$,即滤除该频段内信号. 傅里叶逆变换可得该带通滤波器的时域表达为辛格函数:

$$x(t)=\frac{1}{2\pi}\int_{-\infty}^{\infty}X(\mathrm{j}\omega)\mathrm{e}^{-\mathrm{j}\omega t}=\frac{\sin(\omega_0 t)}{\omega_0 t} \tag{13}$$

其中,t表示时间. 该带通滤波器在时域上的响应表现为一个显著的主脉冲,其幅值较高,随后伴随着无数逐渐衰减的周期性脉冲. 这种现象中,幅值最高的主脉冲被称为“振铃”脉冲波信号. 由于实际硬件存在瑕疵,这个显著的“振铃”脉冲波信号有可能会泄露到低频的可听声频段内.

为了减小振铃脉冲的影响,本节采用频域滤波器在频域上对超声波信号进行平滑滤波. 具体地,本节将频域中的辛格函数应用于带通滤波器,以替代矩形窗函数. 辛格函数的表达式为

$$X(\omega)=\frac{2\sin\left(\frac{\omega}{2}\right)}{\omega} \tag{14}$$

根据傅里叶变换的可逆性,辛格函数带通滤波器在时域中的表达形式与矩形窗函数相似. 这种滤波器设计能有效平滑信号频率的突变过程,从而消除“振铃”脉冲现象. 虽然这种处理会产生一些带外频率分量,但是这些带通频段外的频率分量能量较低,且仍位于超声频段范围内. 因此这种调制方式能够不显著降低带通频段内超声能量的同时,避免可听噪声的产生.

(2)超声扬声器的非线性特征

超声扬声器的内部放大器会产生非线性效应,导

致发出的超声信号被解调成低频声音. 假设待发射的超声信号的中心频率为f_c,带宽为2 BW. 由于超声扬声器的非线性特性,它将产生一个中心频率为0、带宽为BW的可听声信号. 这种非线性效应会让超声信号在空气中传播时变成可听噪声.

为了解决这一问题,本节将超声信号分解成多个较窄带宽的信号,并使之频率低于20 Hz,即转化为人耳无法感知的次声波,从而避免可听噪声的出现. 以带宽为2 BW的超声信号为例,从频域上将它$k(k\in N)$等分,分解为$k(k\in N)$个带宽为$\frac{B}{k}$的信号. 通过选择较大的k,使得$\frac{B}{k}<20$ Hz,即可使自解调所产生的低频信号无法被人耳察觉.

总的来说,本文的超声频率设计方法的创新性和特点体现在以下两个方面. 在载波频率选择方面,本文方法在兼顾普适性的同时提高了有效干扰距离. 基于不同传声器之间的非线性系数差异并不明显这一观察,本文通过大量实验选定20款传声器的非线性系数平均值作为超声载波频率,从而使得系统能够兼顾多种类型传声器的干扰. 与此同时,本文的超声载波频率设计还考虑到了超声波的传播距离这一因素. 在可听噪声泄露及解决措施方面,本文分析了普遍存在的振铃效应与扬声器非线性特征. 针对这两个引发可听噪声泄露的问题,本文使用频域辛格函数平滑频率突变以避免振铃效应,并将超声信号分解成多个较窄带宽的信号以降低扬声器非线性特征对人耳的影响.

4.4 原型系统实现

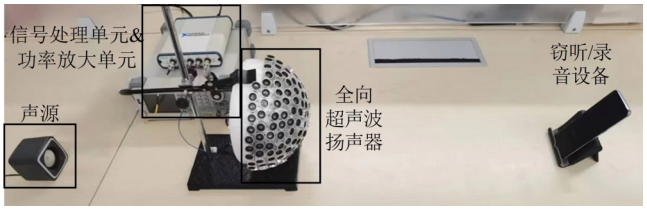
本节使用商用模组设计实现超声波录音干扰原型系统. 该原型系统由三部分组成:信号处理单元、功率放大单元与全向超声扬声器.

信号处理单元中,NI USB-4431信号处理器生成耦合噪声,并将其调制于由信号发生器(SIGLENT SDG1020)产生的频率分别为39 kHz和80 kHz的超声波载波信号之上. 功率放大单元中,数模转换器(CJC4344)将超声干扰信号转换为模拟信号,并由功率放大器以100 dB的发射声强驱动全向超声扬声器.

为了实现大范围的录音干扰,本节设计一款全向超声扬声器. 由于超声扬声器往往具有较强的方向指向性,即在空气介质中传播的超声信号能量集中辐射于特定方向,无法向各个方向同时发射超声干扰信号. 为了提高覆盖角度,该原型系统的全向超声扬声器由110个超声波换能器(NU40A14TR-1)组成,并沿半球面均匀分布,如图4(*a*)所示. NU40A14TR-1超声波换能器能够发射35~85 kHz的超声信号,其部分频率响应曲线如图5所示. NU40A14TR-1超声波换能器在37~41 kHz频段内展现出约4 kHz带宽的平坦特性,且在该带宽内最高能够输出超过100 dB以上的超声能量. 通过上述选型和设计,该原型系统能够实现大范围的超声干扰噪声发射.

该原型系统参数总结于表1.

(*a*) 全向超声波扬声器



(*b*) 原型系统与实验设置

图4 原型系统与实验设置

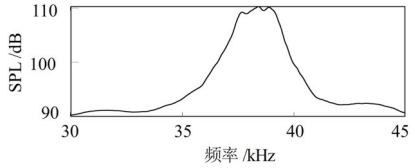


图5 NU40A14TR-1超声波换能器频率响应曲线

表1 超声波录音干扰原型系统参数表

载波频率	超声声强	超声波换能器及数量
39 kHz, 80 kHz	100 dB	110个NU40A14TR-1

5 实验结果与分析

5.1 实验设置

实验布局 超声波录音干扰原型系统位于声源与窃听设备之间,并放置于一个安静的实验室(环境噪声为48.1 dB)中,如图4所示. 三者之间的默认布局:声源处于原型机正后方0.5 m处,窃听设备处于原型机正前方2 m处. 三者之间距离的影响将由后续实验进行定量分析.

声源与待保护信号 本文使用EDIFIER M230扬声器作为声源,并播放预先录制的音频文件. 该扬声器

以65 dB(约为人类正常交谈时声音的强度水平)播放音频信号. 音频信号来源于三个数据集,包括两个开源汉语普通话数据集(ST-CMDS[22]和CN-Celeb1[24])和一个自制汉语普通话数据集. 自制数据集中,共12名志愿者(包括6位男性和6位女性,年龄分布为22~55岁),平均语速分别约为每分钟130字(慢语速)、170字(正常语速)和200字(快语速),朗读《傅雷家书》. 各个语速下各位志愿者分别阅读10 min,共计时长为6 h的语音数据集.

窃听/录音设备　本实验共选用40款传声器或录音设备作为窃听设备,包括:20款商用传声器模块和20款具有录音功能的商用电子产品. 本实验预先随机选取其中一半设备测量非线性系数,并按照4.3.1节中策略计算非线性系数的平均值用于超声载波频率的选择. 另一半设备的非线性系数未经过测量. 由于后续实验结果表明,本文的超声波录音干扰方法在有无测量非线性系数的录音设备上均取得较优的干扰性能,因此后续不进行区分,以这40款录音设备上的平均干扰性能作为最终结果.

评价指标　本文选用协同单词错误识别率(Cooperative Word Error Rate, CWER)[13]和信噪比(Signal-to-Noise Ratio, SNR)作为评价指标. CWER的定义:由人类听众和自动语音识别(Automatic Speech Recognition, ASR)系统协同工作,所能够正确识别出的字数比例:

$$\mathrm{CWER}=1-\frac{\mathrm{card}\left(\bigcup_{i=1}^{n}\mathrm{R}_{A_i}\bigcup_{j=1}^{m}\mathrm{R}_{H_j}\right)}{\mathrm{card}(S)} \tag{15}$$

其中,card(·)为集合的基数,即该集合中元素的数量;由m个ASR系统和n个人类听众共同协作,识别语音数据集S中的字/内容;第i个ASR系统所正确识别出来的字集合为$\mathrm{R}_{\mathrm{ASR}_i}$;第$j$个人类所正确识别出来的字集合$\mathrm{R}_{H_j}$. 该指标能够合理地反映窃听者使用各种手段对语音内容进行识读,涵盖了人耳听力对词句和识别与理解能力的同时,结合多种使用了不同深度学习网络和不同训练数据的ASR系统,最大程度地反映了窃听者对语言内容的识读能力. 本实验选用了科大讯飞[28]、ASRT[29]和腾讯微信语音转文字服务这三款汉语/中文ASR系统,并招募7名无听力障碍的汉语母语志愿者(年龄分布在22~40岁之间).

窃听者去噪技术　根据第3节中威胁模型,本文从利用时域、频域与时频域特征的去噪技术中各选择一种代表性方法.

(1)BSS技术:窃听者利用多个传声器(本实验中采用4个)录制多声道音频,并通过基于快速独立成分分析算法[30]的盲源分离技术,利用语音与噪声在时域特征上的独立性进行信号分离.

(2)频率滤波器:窃听者采用短时傅里叶变换分析带噪录音的频谱,并挖掘噪声的频率特征. 随后,利用带阻滤波器来去除高功率的噪声成分.

(3)嗅探器辅助的自适应噪声滤波器:窃听者运用嗅探器这一辅助工具发送超声波,以捕捉干扰噪声[31]. 随后,通过基于归一化最小均方算法的自适应滤波器,实现对干扰噪声的有效去除.

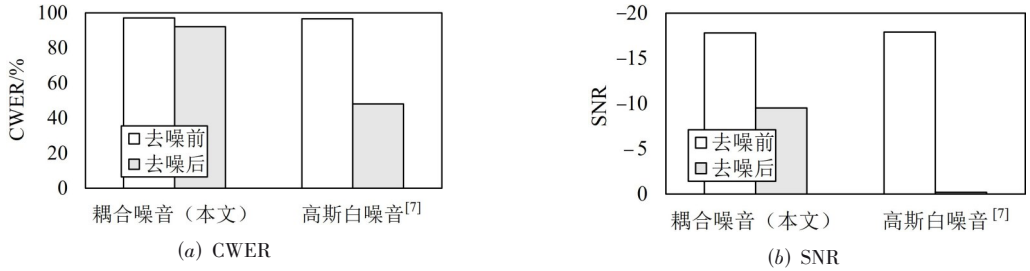
为了评估超声波录音干扰方法的抗去噪能力上限,本文选择在各种去噪技术后所得到的CWER和SNR最低值作为本文方法安全性评价依据. 由于不同条件下不同去噪手段的影响并不相同,选择指标的最低值能够正确反映窃听者采用这种手段对语音的恢复的情况下超声波录音干扰方法的真实性能. 后续实验将证明该方法具有极佳的录音干扰性能且在不同条件下变化不大,本文实验重点将侧重于抗去噪性能的评估. 因此,除5.2节外,本文实验结果均为去噪手段作用后的录音干扰方法抗去噪性能.

5.2　整体性能评估

本实验测评使用耦合噪声的超声波录音干扰方法的干扰与去噪抵御能力,并与Backdoor[7]对比. 根据文献[5]与图2中结果可得,Backdoor[7]系已有超声波录音干扰技术中抵御去噪能力最优的,因此被选作比较对象. Backdoor[7]使用8 kHz带宽高斯白噪声,并部署于5.1节的实验条件下,实验结果对比如图6所示. 使用高斯白噪声与耦合噪声均能达到96%以上CWER的干扰效果,耦合噪声的性能以不足1个百分点CWER的优势微高于高斯白噪声. 然而,当面临去噪手段后,高斯白噪声的性能大幅度下跌,CWER由去噪前的96.6%大幅度下降至48.1%,SNR由−18.9 dB提升至−0.2 dB. 相比之下,耦合噪声有效地抵御了去噪手段,CWER仅下降了5.1个百分点,保持在了92.0%以上;SNR虽有上升,但幅度远低于高斯白噪声,仅上升至−9.5 dB,保存了大量噪声. 上述结果说明,本文超声波录音干扰方法在抗去噪性能上优于现有干扰技术,能够避免窃听者恢复语音.

5.3　干扰覆盖范围

本节评估超声波干扰方法的有效作用范围. 本实验测试范围为以超声扬声器所在位置为原点,半径6 m范围内的半圆形范围,录音设备依次放不同的位置下进行录制语音. 具体地,在该半圆形范围内绘制正交网格图,且网格内最小矩形边长为5 cm,在每个网格点上依次放置录音设备进行音频录制(即以5 cm的间隔步长遍历整个半圆形区域). 这种以固定距离正交网格式遍历的方式能够保证所采集的数据较为均匀,避免遍历角度进行测量的方式出现的中心区域测量点密集,外围区域测量点稀疏的情况,有利于真实反映覆盖范



(a) CWER (b) SNR

图6 整体性能

围内的性能表现，防止稀疏测量部分出现过大或过小的异常值而导致整体结果偏离实际情况. 对于所得到的音频信号，分别采用第5.1节中所介绍的三种去噪技术进行噪声消除，降噪后的音频文件分别交由7名志愿者和3款汉语ASR系统进行汉语语音内容识别，并根据所能够正确识别出的字数比例计算CWER. 为了直观展示数据结果，本节使用三次样条插值(Cubic Spline Interpolation)将CEWR插值到1 cm步长密度，并用MATLAB绘制结果云图，如图7所示.

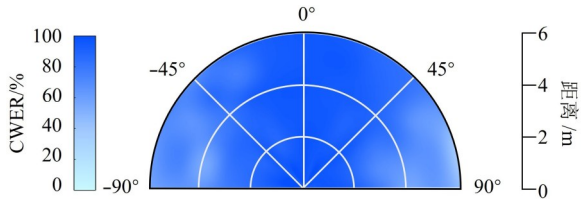


图7 干扰覆盖范围性能

原型系统能够抵御各种去噪技术，实现有效干扰，CWER均高于50%. 在半径6 m、角度为±30°的扇形区域内，CWER仍高于80%. 当范围进一步扩大至±45°的扇形区域，原型系统保持76.4%以上的CWER；角度继续扩大至±60°，CWER均高于60%.

在超声扬声器正对方向上录音干扰方法的抗去噪性能极佳，CWER始终保持在90%以上；在距离4 m、角度在±60°范围内，该方法的抗去噪表现依旧优异，CWER超过75%. 上述结果不仅适用于二维平面，在三维空间中仍具有类似性能表现. 通过调整干扰系统的数量、部署位置和覆盖角度(如同时使用两个全向超声扬声器即可实现360°的广角覆盖)，用户可以轻松实现全面立体的干扰效果.

5.4 窃听设备的影响

本实验测试了对40款录音设备的干扰效果，包括20款传声器模组(7款Panasonic传声器、5款Bosung传声器、4款PRIMO传声器、3款HOSIDEN传声器和1款KNOWLES传声器)和20款消费电子设备(包括10款智能手机、4款智能扬声器系统、3款录音笔、2款笔记本电脑和1款平板电脑). 这40款录音设备依次放置于超声波录音干扰原型系统正前方2 m处. 即使窃听者采用各项去噪手段，该原型系统对各个录音设备均有较优的干扰效果，平均CWER高达92.0%，并且对各个录音设备上测试所得CWER均大于75%. 值得注意的是，测试设备中仅有20款提前测量非线性系数以用于设计超声载波频率，其余设备的所有参数均未经任何测量. 对于这些未知参数的录音设备，该原型系统仍然能够保持良好的录音干扰与去噪抵抗性能. 这说明，本文方法能够在实际应用中有效地防范各类型的录音设备.

5.5 声源的影响

5.5.1 数据声源距离的影响

在现实场景中，用户并不会始终与超声波录音干扰设备保持较近的位置. 本节评估声源(扬声器)与该系统之间的距离对本文干扰方法性能的影响. 实验过程中声源方向与超声扬声器方向保持一致. 实验结果如图8所示，其中距离的正方向为超声扬声器轴向(即正对方向)，窃听设备与超声扬声器间距离为2 m. 对于不同距离的声源，本文方法实现了均值为92.34%、最小值为91.8%的干扰性能. 这使得实际使用中用户无需刻意停留在特定位置，即可实现灵活地窃听干扰.

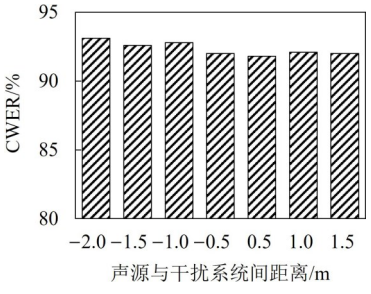


图8 声源距离的影响

5.5.2 多声源/用户场景的干扰性能

在现实场景中，多人交谈的情况很常见，因此超声波录音干扰技术需要能够有效保护多用户的语音隐私. 7位用户分别进行时长为1 min的注册后，聚集在距离该超声波录音干扰原型系统2 m范围(位置随机)，依次递增谈话中用户数量，实验结果如图9所示. 当用户数量小于3位时，系统性能几乎没有影响. 当用户数量

大于4位时,系统性能随着用户数量的增加呈现略微下降的趋势,但仍保持高于90%的语音内容无法被恢复识别. 这种下降趋势的原因可能在于用户的增加使得超声传播的多径效应变得显著,增加了超声波的衰减,略微削弱了能够抵达到窃听设备处的干扰噪声强度. 总的来说,本文干扰方法能够支持多用户语音隐私保护,在任意数量声源的情况下保护了所有用户超过90%的语音内容.

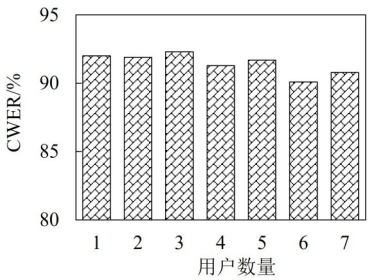


图9　多用户场景下的干扰性能

5.5.3　语速的影响

用户的语速常常千差万别,本实验将用户语速纳入考量,测试不同语速下该超声波录音干扰方法的性能. 12名志愿者分别被要求以平均语速分别约为每分钟130字(慢语速)、170字(正常语速)和200字(快语速)分别注册,实验结果如表2所示. 当待保护语音语速与注册语速相近时,该原型系统抗去噪性能分别为92.1%(慢语速)、92.0%(正常语速)和92.8%(快语速);当待保护语速与注册语速不同时,原型系统性能依次为92.2%、91.8%和92.9%. 实验结果说明,用户语速对本文方法几乎没有影响.

表2　不同语速下系统性能表　　单位:%

注册\测试	慢语速	正常语速	快语速
慢语速	92.1	91.6	92.8
正常语速	92.3	92.0	92.9
快语速	92.0	91.9	92.8

5.6　原型系统功率消耗

本实验使用数字功率计PF9800测量超声波录音干扰原型系统的功率消耗. 如表3所示,在持续工作3 h内,该原型系统的功率总消耗仅为1.063 W,平均功耗为0.354 W/h或者5.9 mW/min. 低功耗的特性使得该系统能够在各类场景中保证较长的续航时间,满足用户的语音保护需求.

表3　超声波录音干扰原型系统功耗结果表

三小时总功耗	平均功耗
1.063 W	0.354 W/h或者5.9 mW/min

5.7　用户感知度评估

本节从超声暴露声强和用户主观性实验两个维度评估该方法对用户听力的影响.

为了避免对用户生理健康产生不良影响,本文干扰方法及原型系统的最高声强严格限制在100 dB以下. 本实验使用NI USB-4431声音测量仪测量声场强度. 距离超声扬声器0.5 m,测量所得的超声信号强度为95 dB,远低于国际非电离辐射委员会建议的超声声强限值110 dB[27].

用户主观性实验中,22名志愿者被随机分为两组. 两组人员依次进入部署超声波录音干扰原型系统的会议室中自由活动与交谈30 min,其中一组人员活动期间超声波录音干扰原型系统持续工作,另一组则保持关闭. 在这之后,志愿者立即被要求根据自己在活动过程中的实际感受,对是否察觉到噪声或耳部不适进行1~5分的评分. 其中,1分表示完全没有异常感觉,5分则表示能明显感觉到噪声或不适. 统计结果如表4所示,原型系统开启组的平均评分为1.97,而关闭组则为1.91. 这两个评分非常接近,进一步证实了本文干扰方法所发射的超声波在用户日常活动中几乎不会被人耳所察觉.

表4　用户主观感知度结果表

开启系统用户感知度评分	关闭系统用户感知度评分
1.97	1.91

综上所述,本文提出的超声波录音干扰方法不仅具备卓越的窃听干扰能力,而且严格遵循国际安全标准,确保用户生理健康,实现了既安全又舒适的语音隐私保护体验.

6　结论

本文提出了一种专为汉语语音设计的安全且稳健的超声波录音干扰方法. 鉴于对手可能采用复杂的去噪技术所带来的潜在威胁,本文深入剖析了汉语语音的独特特征,并基于此设计了一种耦合噪声生成算法. 该算法生成的超声波干扰噪声与待保护语音信号紧密耦合,使得两者在时域和频域上展现出高度相似的特性. 因此,即便窃听者采用各种去噪技术,也难以从中提取出有价值的语音信息. 本文噪声生成算法无需实时采集用户语音,这一特点极大地拓宽了干扰器的应用场景. 本文设计超声波录音干扰原型系统,实现了6 m范围的有效保护,为用户的语音隐私提供了坚实保障.

参考文献

[1] 盛玉雷. 语音入口得加把“隐私锁”[N/OL].(2019-08-13)[2024-03-14]. http://opinion. people. com. cn/n1/2019/0813/

c1003-31290817.html.

[2] THOMAS G. How to protect yourself from camera and microphone hacking[EB/OL]. [2024-03-14]. https://www.consumerreports.org/electronics-computers/privacy/how-to-protect-yourself-from-camera-and-microphone-hacking-a1010757171.

[3] WU H, QIAN W. Breaking Smart Speakers: We are Listening to You[R]. Las Vegas: DEF CON Hacking Conference, 2018.

[4] SLOTTA D. Smart Speaker Market in China - Statistics & Facts[R]. Germany: Statista, 2023.

[5] CHEN Y K, GAO M, LI Y M, et al. Big brother is listening: An evaluation framework on ultrasonic microphone jammers[C]// Proceedings of the International Conference on Computer Communications. Piscataway: IEEE, 2022: 1119-1128.

[6] ZHANG G M, YAN C, JI X Y, et al. DolphinAttack: Inaudible voice commands[C]//Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security. New York: ACM, 2017: 103-117.

[7] ROY N, HASSANIEH H, ROY CHOUDHURY R. BackDoor: Making microphones hear inaudible sounds[C]//Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services. New York: ACM, 2017: 2-14.

[8] CHEN Y X, LI H Y, et al. Wearable microphone jamming[C]//Proceedings of the Conference on Human Factors in Computing Systems. New York: ACM, 2020: 1-12.

[9] LI L K, LIU M N, YAO Y G, et al. Patronus: Preventing unauthorized speech recordings with support for selective unscrambling[C]//Proceedings of the 18th Conference on Embedded Networked Sensor Systems. New York: ACM, 2020: 245-257.

[10] CHEN Y K, GAO M, LIU Y J, et al. Implement of a secure selective ultrasonic microphone jammer[J]. CCF Transactions on Pervasive Computing and Interaction, 2021, 3(4): 367-377.

[11] MAKINO S, LEE T W, SAWADA H. Blind Speech Separation[M]. Berlin: Springer, 2007.

[12] SUN K, CHEN C, ZHANG X Y. “Alexa, stop spying on me!”: Speech privacy protection against voice assistan-ts[C]// Proceedings of the Conference on Embedded Networked Sensor Systems. New York: ACM, 2020: 298-311.

[13] GAO M, CHEN Y K, LIU Y J, et al. Cancelling speech signals for speech privacy protection against microphone eavesdropping[C]//Proceedings of the 29th Annual International Conference on Mobile Computing and Networking. New York: ACM, 2023: 1-16.

[14] HUANG P, WEI Y, CHENG P, et al. InfoMasker: Preventing eavesdropping using phoneme-based noise[C]// Proceedings of the Network and Distributed System Security Symposium. Piscataway: IEEE, 2023: 1-13.

[15] 王力. 汉语音韵,音韵学初步[M]. 北京: 中华书局, 2014: 29.

[16] WANG X H, XU L. Speech perception in noise: Masking andunmasking[J]. JournalofOtology,2021,16(2):109-119.

[17] ZIEHE A, KAWANABE M, HARMELING S, et al. Blind separation of post-nonlinear mixtures using linearizing transformations and temporal decorrelation[J]. Journal of Machine Learning Research, 2003, 4: 1319-1338.

[18] FREITAG L, STOJANOVIC M, SINGH S, et al. Analysis of channel effects on direct-sequence and frequency-hopped spread-spectrum acoustic communication[J]. IEEE Journal of Oceanic Engineering, 2002, 26(4): 586-593.

[19] JUTTEN C, BABAIE-ZADEH M, HOSSEINI S. Three easy ways for separating nonlinear mixtures? [J]. Signal Processing, 2004, 84(2): 217-229.

[20] ALMEIDA L B. Linear and nonlinear ICA based on mutual information[C]//Proceedings of the Adaptive Systems for Signal Processing, Communications, and Control Symposium. Piscataway: IEEE, 2002: 117-122.

[21] TALEB A. A generic framework for blind source separation in structured nonlinear models[J]. IEEE Transactions on Signal Processing, 2002, 50(8): 1819-1830.

[22] OpenSLR.ST-CMDS-20170001_1, Free ST Chinese MandarinCorpus[EB/OL]. (2022-08-29) [2024-03-24]. https://openslr.org/38/.

[23] JIA Y, ZHANG Y, WEISS R, et al. Transfer learning from speaker verification to multispeaker text-to-speech synthesis[J]. Advances in Neural Information Processing Systems, 2018, 31: 4485-4495.

[24] FAN Y, KANG J W, LI L T, et al. CN-celeb: A challenging Chinese speaker recognition dataset[C]//IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE, 2020: 7604-7608.

[25] SAKOE H, CHIBA S. Dynamic programming algorithm optimization for spoken word recognition[J]. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1978, 26(1): 43-49.

[26] OTSU N. A threshold selection method from gray-level histograms [J]. IEEE Transactions on Systems, Man, and Cybernetics, 1979, 9(1): 62-66.

[27] DUCK F A. Medical and non-medical protection stan-

dards for ultrasound and infrasound[J]. Progress in Biophysics and Molecular Biology, 2007, 93(1-3): 176-191.

[28] An artificial intelligence platform focusing on intelligent speech interaction which provides solutions for developers[EB/OL]. [2024-03-14]. https://www.xfyun.cn.

[29] Nl8590687. ASRT: a DL-based Chinese ASR system[EB/OL]. (2022-08-29)[2024-03-14]. https://www.xfyun.cn.

[30] OJA E, YUAN Z J. The fastICA algorithm revisited: Convergence analysis[J]. IEEE Transactions on Neural Networks, 2006, 17(6): 1370-1381.

[31] HE Y T, BIAN J Y, TONG X Y, et al. Canceling inaudible voice commands against voice control systems[C]// The 25th Annual International Conference on Mobile Computing and Networking. New York: ACM, 2019: 1-15.

作者简介

高　铭　男,1996年2月出生于江苏省连云港市. 现为南京邮电大学计算机学院、软件学院、网络空间安全学院校长专聘教授. 主要研究方向为智能感知、物联网安全与隐私保护. 中国电子学会会员编号:E190159838M.
E-mail: gaomingppm@njupt.edu.cn

陈奕可　男,1998年1月出生于江苏省泰州市. 现为浙江大学计算机科学与技术学院博士研究生. 主要研究方向为物联网安全与无线感知.
E-mail: chenyike@zju.edu.cn

陈佳彤　男,2002年5月出生于重庆市. 现为浙江大学计算机科学与技术学院硕士研究生. 主要研究方向为物联网安全.
E-mail: 3200105258@zju.edu.cn

肖　甫　男,1980年10月出生于湖南省邵阳市. 现为南京邮电大学计算机学院、软件学院、网络空间安全学院教授、博士生导师. 主要研究方向为物联网感知计算、物联网安全技术与数据中心网络.
E-mail: xiaof@njupt.edu.cn

韩劲松　男,1975年12月出生于山东省淄博市. 现为浙江大学计算机科学与技术学院教授、博士生导师. 主要研究方向为物联网安全、智能感知与隐私保护.
E-mail: hanjinsong@zju.edu.cn