VoiceSecure: A Microphone-Module to preserve Speech Privacy in Voice Communication

ABSTRACT

This paper addresses a critical concern regarding user privacy in audio calls. Voice communication has become an integral part of modern life. However, such audio calls frequently convey sensitive information. If intercepted by malicious actors, the audio content becomes vulnerable to speaker identification and the extraction of confidential information. To mitigate this risk, we introduce *VoiceSecure*, a first microphone module that protects speech privacy by thwarting automatic speaker identification and speech recognition while keeping the sound perceptually similar to the human users. *VoiceSecure* employs a set of carefully designed speech modifications tuned to optimally secure speech privacy without degrading its usability. The microphone module seamlessly integrates with existing devices via an audio jack or Bluetooth connection. By anonymizing speech before transmitting it to device software, *VoiceSecure* mitigates the possibility of software-based spoofing attacks. We develop a prototype using a commercial microphone and microcontroller. Our experiment shows that *VoiceSecure* outperforms in concealing sensitive information from human speech without perceptibly altering it for users.

1 INTRODUCTION

In today's interconnected world, Voice over Internet Protocol (VoIP) stands as a cornerstone of modern communication, seamlessly facilitating voice calls across the globe. Speech privacy over voice calls is an important aspect of communication, particularly in situations where sensitive or confidential information is being exchanged. It is crucial to ensure that the conversation remains private and cannot be intercepted by unauthorized individuals while making voice calls. In an ideal scenario, end-to-end encryption should be used for voice communication. This encryption technique ensures that the sender transmits encrypted audio that can only be decrypted by the intended receiver, without any possibility of decryption by cell phone carriers or app servers. However, Voice over Internet Protocol (VoIP) is currently the most commonly used technology for voice communication [4]. Although it uses digital encryption between your phone and the cellular telephone base station, this encryption method can only protect data from eavesdroppers. Cell phone companies can still access and listen to private speech, which allows them to run mass surveillance systems on speech conversations, which poses a significant risk to public privacy. The National Security Agency has built a surveillance system capable of recording a foreign country's telephone calls [2, 15]. The law, Section 702 of the Foreign Intelligence Surveillance Act, allows warrantless telephone and internet monitoring of non-US citizens abroad Despite the fact that millions of these communications are obtained without warrants, they have been used in ordinary criminal investigations undermining a core purpose of the Fourth Amendment of the US Constitution to protect the people against searches without probable cause [3].

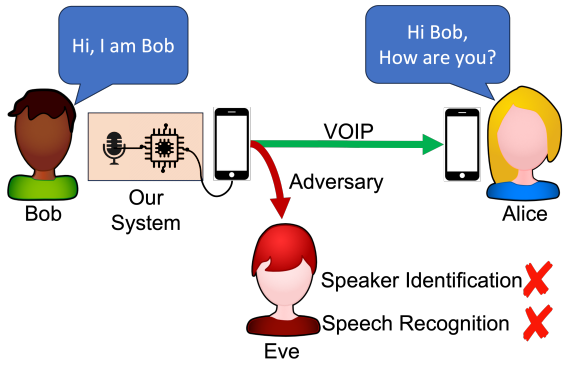

Figure 1: *VoiceSecure*, a microphone module to protect speech privacy in voice communication.

The COVID-19 pandemic has led to an increased usage of teleconferencing applications for conducting business meetings and staying in touch with family and friends. This increased usage highlights the importance of analyzing the security and privacy concerns associated with these communication channels. Two of the most commonly used teleconferencing applications, Zoom and Skype, do not ensure end-to-end encryption. Zoom uses transport encryption which means that the content is encrypted when shared between your device and Zoom servers, but it is not accessible to Zoom [21]. This enables Zoom to spy on private meetings and hand over recordings of meetings to governments or law enforcement in response to legal requests. Skype uses 256-bit AES symmetric encryption where a unique key is generated for each session [1]. If you make a call from Skype to mobile or landline phones, the part of your call that takes place over the Public Switched Telephone Network (PSTN) is not encrypted and hence vulnerable to even basic eavesdropping attacks [16].

In addition to the privacy threat posed by service providers, researchers have also investigated attacks on encrypted speech passing over the network. A team from BlackBerry discovered a vulnerability that allows attackers to intercept and decrypt GSM calls. The vulnerability is present in the GSM encryption key exchange process that establishes a secure connection between a phone and a nearby cell tower during a call. This problem is present in all GSM implementations up to 5G [27]. Another work ReVoLTE [32], demonstrates an attack on Voice over LTE (VoLTE), which uses a packet-based telephony service integrated into the Long Term Evolution (LTE) standard. This attack exploits an implementation flaw in LTE which allows an attacker to recover the contents of an encrypted VoLTE call. In this paper, we ask a fundamental question: *Is it possible to shield users from mass surveillance by thwarting automatic systems from speaker identification and safeguarding private speech from automated transcription?*

There are existing methods aimed at protecting user privacy by shielding speakers from automatic identification [8, 11, 13, 40] and preventing speech from automatic recognition [9, 10, 42]. However, existing anonymization methods encounter two primary challenges in these scenarios: firstly, they are not suitable for low-latency and computationally efficient frame-by-frame transmissions. This is because they often rely on deep neural networks, which are computationally expensive and/or require processing entire audio pieces rather than individual frames. Secondly, the anonymization process is rendered ineffective if an attacker gains access to the original audio prior to its anonymization at the software level [18, 19]. For instance, an attacker might compromise the victim's device and gain control over the microphone or directly access local audio file. In conclusion, there is currently no system capable of protecting both speaker identity and speech content in real-time before being accessed by the device software.

To address these concerns, we introduce *VoiceSecure*, an innovative microphone module designed to provide real-time speech anonymization that protects user privacy from automatic speaker identification and speech recognition while keeping speech intelligibly the same for the human listener. This ensures users can enjoy the same audio call experience without concerns of privacy breaches over voice calls. *VoiceSecure* functions as an external microphone, capturing speech audio, applying anonymization in real-time, and transmitting it to the user's device via either an audio jack or Bluetooth connection.

There are several challenges in realizing this system, primarily due to the constrained capabilities of the hardware module. Existing computationally expensive speech anonymization techniques [11, 40, 42] are not suitable for the hardware module, so, we need to design a new computationally efficient anonymization technique. This new technique must be designed to jointly optimize to evade automatic speaker identification and speech recognition. Moreover, the system must operate seamlessly in real-time without disturbing the natural call experience, ensuring that the anonymization process remains imperceptible during normal communication. Maintaining indistinguishability to users is paramount for preserving the natural flow and integrity of communication.

Our solution leverages the inherent differences between how the human brain perceives speech sounds and how speaker identification and speech recognition models process speech signals [24, 45]. Specifically, *VoiceSecure* adopts a multi-step approach. Initially, it segregates voiced and unvoiced segments of speech, leveraging the human tendency to focus on the high-energy voiced regions to understand the content, while unvoiced regions aid in smooth sound transitions. Subsequently, *VoiceSecure* eliminates inaudible frequency components from the speech, as they are not important for human perception. Additionally, *VoiceSecure* introduces random temporal distortions by flipping small time-domain speech windows, exploiting a phenomenon elucidated in the psychoacoustic literature. Furthermore, *VoiceSecure* manipulates the pitch and formants of the speech signal, which are crucial attributes for automatic speaker identification. However, random alterations to these features risk altering the sound to human listeners, potentially leading them to believe they are conversing with a different individual. To overcome this challenge, we observe that minor variations in pitch or formants within small speech windows (less than 50 milliseconds) are imperceptible to humans. Hence, *VoiceSecure* strategically selects small time windows with intervals in between and applies these modifications, effectively fooling automatic speaker identification while preserving speech perceptually similar for human listeners.

Summary of contributions:
In developing the proposed system, we have made the following three contributions:

- We propose the first microphone module to prevent automatic speaker verification and speech recognition while preserving speech intelligibility for humans.
- We develop a signal processing-based algorithm for speech manipulation and tune its parameters using particle swarm optimization.
- We implement *VoiceSecure* on an off-the-shelf microcontroller, and evaluate its performance under various scenarios.

2 OVERVIEW

Before delving into the details of our system, we first discuss the basics of human speech, and how state-of-the-art automatic speaker verification and speech recognition systems work.

2.1 Fundamentals of Human Speech

Human speech is a complex and intricate process involving various physiological and linguistic components. At its core, speech production relies on the coordinated movement of the vocal tract, including the lungs, vocal cords, pharynx, and articulators such as the tongue and lips. These organs work in harmony to produce sounds that convey meaning and intention. Human speech is made up of various sound components known as phonemes. These components are combined to form any speech. The set of possible phonemes is fixed and the English language is made up of 44 unique phonemes. Phonemes can be divided into vowels, consonants, fricatives, nasals, and stops based on the type of corresponding sound they produce. Since the possible number of phonemes is fixed for a particular language, Automatic Speech Recognition systems are commonly trained to identify the sequence of phonemes from the speech signal and then combine them to decode the actual content of the speech. In addition to the phonemes that are used to articulate words, speech sounds exhibit distinct pitches, harmonic structures, formants, rhythms, and intonations, which are intricately linked with the speaker's biological traits. These distinctive features serve as pivotal signatures for speaker identification, forming the base of automatic speaker verification systems.

2.2 Automatic Speaker Verification (ASV)

Automatic Speaker Verification (ASV) systems are innovative systems designed to identify speakers based on their unique vocal characteristics. These models utilize advanced algorithms to analyze speech patterns, pitch, formants, intonation, and other vocal attributes to verify a speaker's identity with a high level of accuracy. There are signal-processing-based methods that can extract these attributes from the speech signal and then compare them for speaker verification. Although these methods have shown good performance and are favorable in terms of computation, they are not robust against various environmental factors. Recently, machine learning-based techniques have also been used for speaker identification [12, 34]. Deep neural networks are trained to compute hierarchical features from the speech signal and learn robust speaker representations that are agnostic to the speech content, recording channel, and ambient noise.

2.3 Automatic Speech Recognition (ASR)

Speech Recognition systems are widely used in applications like voice assistants. These systems are designed to transcribe spoken language into text, enabling seamless human-computer interaction. Initially, these models extract features from the speech signal, employing techniques such as Discrete Fourier Transform (DFT), Mel Frequency Cepstral Coefficients (MFCC), Linear Predictive Coding, and the Perceptual Linear Prediction Method. Subsequently, various statistical models like Hidden Markov Models, Gaussian Mixture Models, and Deep Neural Networks are utilized for decoding actual content from the extracted features. Recently, there has been a surge in end-to-end learning-based methods, combining whole feature extraction and decoding phases into a unified model, thus enhancing efficiency and accuracy in Automatic Speech Recognition.

3 ADVERSARY MODEL

Audio calls over the internet have become an important aspect of modern life, serving various purposes from confidential business meetings to personal conversations with family and friends. However, these calls often transmit highly sensitive information. If intercepted by adversaries, the audio content can be exploited in various ways. For instance, adversaries could extract voice prints that enable them to conduct replay or voice mimicry attacks on speaker verification systems which are commonly utilized for authentication.

Furthermore, adversaries may exploit intercepted audio to infer sensitive details discussed during conversations. For example, during a call with a bank, account numbers, and financial details could be inferred, or health issues could be inferred during a conversation with a medical professional. Such information can then be linked to specific speakers, leading to more targeted attacks.

Our adversary model supposes the ability of adversaries to eavesdrop on Voice over Internet Protocol (VoIP) channels. Existing literature [27, 32] has exposed the vulnerability in current GSM and LTE channels that allow attackers to access data over voice communication. We also assume a scenario where service providers can easily access data of audio calls [2, 15]. We also envision a scenario where an adversary can tamper with the device software to get access to the raw audio samples.

In addressing these concerns, our proposed solution must fulfill three crucial criteria: Speech Anonymity, Usability, and Security.

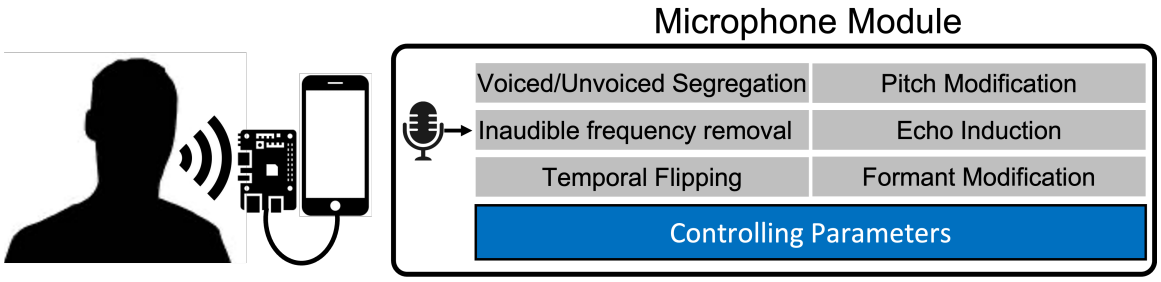

Figure 2: A systematic overview of *VoiceSecure*.

Speech Anonymity: The solution must prevent the adversary from inferring speaker identities, and sensitive content from the speech audio.
Usability: It is important that the resultant audio should sound perceptually similar to human listeners to avoid any disruption to the user's call experience.
Security: The solution should apply speech anonymization before the audio is accessed by device software to mitigate the possibility of any software-based spoofing attacks.

VoiceSecure is specifically designed to safeguard speech privacy over audio calls. Our approach introduces a speech anonymization technique that effectively conceals speaker identities and sensitive content from adversaries while preserving the naturalness and intelligibility of the speech audio for human listeners. *VoiceSecure* develop a microphone module that anonymizes speech audio in real-time before transmitting it to the device software. This module can seamlessly integrate with existing devices, either through traditional audio jack connections or via Bluetooth technology.

It is important to note that our solution is not limited to audio calls but can also be extended to scenarios where adversaries gain access to audio content uploaded by users on social media platforms. While existing solutions may address such scenarios, we emphasize that *VoiceSecure* offers a versatile solution applicable across various applications.

4 SPEECH SECURE

4.1 Design Overview

The primary objective of *VoiceSecure* is to develop a system that can protect user privacy by protecting human speech over audio calls. *VoiceSecure* prevents the automated system from identifying speaker identity, and extracting sensitive content from the speech signal. While modifying speech signals *VoiceSecure* aims to keep sound perceptually similar to human listeners in order to not interfere with user experience. *VoiceSecure* targets to apply speech anonymization before getting accessed by the device software mitigating the possibility of any software-based spoofing attacks.

The development of *VoiceSecure* is divided into three key phases:
Designing Speech Modifications: This phase focuses on designing a set of speech alterations capable of deceiving automatic speaker identification and speech recognition systems.
Optimization of Parameters: The objective of this phase is to identify optimal parameters that effectively protect speech privacy while preserving natural sound qualities for human perception.
Microphone Module Development: This phase involves the implementation of the designed modifications on microcontrollers to create a functional microphone module. The performance of this module will be demonstrated in real-world scenarios.
In the subsequent sections, we will delve into each of these three components in detail.

4.2 Designing Speech Modifications

The design of speech modifications is a pivotal aspect of *VoiceSecure*, as it aims to conceal information from automated speaker identification and speech recognition systems without perceptibly altering speech signals for human listeners. Simply adding noise or randomly altering speech signals would not only hinder automatic systems from extracting crucial information but also impede effective voice communication for users. To devise efficient modifications, *VoiceSecure* capitalizes on the inherent disparities between how the human brain perceives speech sounds and how speaker identification and speech recognition models process speech signals.

Humans perceive sounds differently from the physical sounds that may be present in reality which in turn differs from what might be recorded by the microphone [24, 45]. This variance arises due to the intricate non-linearities inherent in the auditory process and the

unique physical attributes of the ear, nervous system, and brain. Psychoacoustics, a dedicated field of study, delves into understanding these nuances of human perception. Moreover, research suggests that the human brain processes speech in a markedly distinct manner compared to other non-speech sounds [26]. Extensive research in existing literature has conducted numerous experiments and formulated comprehensive psychoacoustic models to understand human perception.

We will highlight some key insights from the literature on psychoacoustics that *VoiceSecure* leverages in the development of our speech modifications. Firstly, the closure principle elucidates how the brain fills auditory gaps, ensuring continuity and coherence in sound perception. This phenomenon, also known as auditory induction, allows for seamless integration of fragmented auditory information. Secondly, studies reveal that even when small speech signal segments are reversed, they remain intelligible. Because it has been divided into roughly phoneme-sized blocks, and, while each phoneme is replayed backward, the overall sequence of phonemes remains intact. Additionally, the Haas effect dictates that the brain prioritizes the first arriving auditory signal when multiple versions reach the ear with slight delays, influencing our perception of spatial and temporal aspects of sound [38]. We also found out that two complex musical tones are perceived as separate when they have different fundamental frequencies, however, the hearing process is capable of dealing with slight mistuning, so almost equal frequencies can be perceived as similar [24]. These insights underscore the complex nature of human auditory perception and provide invaluable insights into the development of *VoiceSecure*.

Considering these observations, *VoiceSecure* adopts a multi-step approach. Initially, it segregates voiced and unvoiced segments of speech, leveraging the human tendency to focus on the high-energy voiced regions to understand the content, while unvoiced regions aid in smooth sound transitions. The closure principle helps to fill the missing unvoiced regions. Subsequently, *VoiceSecure* eliminates inaudible frequency components from the speech, as they do not play any part in human perception. Additionally, *VoiceSecure* introduces random temporal distortions by flipping small time-domain speech windows, exploiting a phenomenon elucidated in the psychoacoustic literature. Utilizing the Haas effect, *VoiceSecure* induces echoes in the speech signal, serving as noise for automatic identification systems while remaining imperceptible to humans. Furthermore, *VoiceSecure* manipulates the pitch and formants of the speech signal, which are crucial attributes for automatic speaker identification. However, random alterations to these features risk altering the sound to human listeners, potentially leading them to believe they are conversing with a different individual. Based on the closure principle, we observe that minor variations in pitch or formants within small speech windows (less than 50 milliseconds) are imperceptible to humans. Hence, *VoiceSecure* strategically selects time windows with intervals in between to apply these modifications. This ensures that minor variations in pitch or formants within these windows remain imperceptible to humans, effectively deceiving automatic speaker identification systems while maintaining perceptual similarity for human listeners.

4.3 Optimization of Parameters

VoiceSecure relies on a range of speech modifications discussed in the previous section, each governed by specific parameters. These parameters control various aspects of speech modification such as the threshold for voiced/unvoiced removal, minimum energy for audible frequencies, amount of pitch and formant variation, and delay for subsequent echoes. Moreover, we incorporate parameters to control the frequency of particular modifications. For instance, to prevent perceptibility to humans, pitch modification is applied to only a certain percentage of time windows.

Achieving optimal performance for *VoiceSecure* relies on finding the ideal set of parameters that effectively safeguard user speech from speaker identification and speech recognition while maintaining sound quality for human listeners.

To find the optimal parameters, we can run a grid search over all possible combinations, which can be computationally intensive and time-consuming. To deal with this challenge, *VoiceSecure* employs Particle Swarm Optimization (PSO). PSO is inspired by the social behavior of bird flocks. This approach randomly selects 'N' sets of parameters from our parameter space, applies speech modifications accordingly, and evaluates their performance in terms of speaker identification, speech transcription, and speech intelligibility. Based on these evaluations, the optimizer refines its selection for the next step, ultimately returning a set of parameters that yield optimal performance in thwarting speaker identifiers and speech recognizers while preserving high speech intelligibility.

4.4 Microphone Module Development

Finally, we develop a microphone module that can be integrated with existing devices either via audio jack or Bluetooth. For that we use Raspberry Pi 3 b+ having a 1.4GHz quad-core processor, and 1GB of RAM. We have

attached a commercial microphone with the Raspberry pi 3. Raspberry has both audio jack and Bluetooth which enable seamless integration with any device. We implement speech modifications on Python using Libros library, a Python package for music and audio analysis. This module listens to the speech audio, applies frame-by-frame modifications in real-time, and transmits it using an audio port to any connected device. This ensures the effective deployment of *VoiceSecure*'s speech anonymization techniques across diverse communication platforms.

5 EXPERIMENTAL EVALUATION

5.1 Metrics

To evaluate the performance of *VoiceSecure* in preventing speaker verification and speech anonymization while keeping the speech intelligibly similar to the humans, we employ the following three metrics commonly used in the field:

- **Equal Error Rate (EER):** EER represents the point at which the false acceptance rate (FAR) equals the false rejection rate (FRR). It provides a balanced measure of performance, indicating the threshold at which the model's discriminative ability is optimal.
- **Word Error Rate (WER):** WER measures the accuracy of automatic speech recognition systems by comparing the number of substitution, deletion, and insertion errors in the recognized transcription against the true speech.
- **Speech Intelligibility (STOI):** STOI quantifies the intelligibility of speech signals by assessing the similarity between the original and test speech signals.

5.2 DataSets

To train and evaluate the performance of *VoiceSecure*, we employ two widely recognized datasets in the domain of speech processing.

LibriSpeech: LibriSpeech is a widely used dataset for speaker verification, consisting of audio recordings of read English speech from public domain audiobooks. The "test-clean" subset is a clean subset of the test set.

VoxCeleb1: VoxCeleb1 is a dataset collected for speaker verification and identification, containing short clips of celebrities speaking from YouTube videos. The "test" subset is commonly used for evaluation purposes.

5.3 State-of-the-art Models

In developing *VoiceSecure* and evaluating its efficacy, we employed the following two state-of-the-art models for speaker verification and speech recognition:

X-Vector is a deep neural network architecture designed for extracting speaker embeddings. It effectively captures speaker-specific characteristics from speech signals, enabling accurate speaker verification across various conditions and scenarios.

DeepSpeech is a deep neural network-based architecture for end-to-end speech-to-text recognition. The core of the engine is a recurrent neural network (RNN) trained to ingest speech spectrograms and generate English text transcriptions.

5.4 Comparison with Baseline

In this section, we compare our results with MicPro [39]. MicPro proposes a microphone module aimed at preserving speaker identity without compromising speech usability for humans. It represents the most closely related system to *VoiceSecure*. However, MicPro does not address the protection of sensitive speech content from speech recognizers. For comparison, we use LibriSpeech DataSet, and for evaluation, we use X-vector for speaker identification and DeepSpeech for Speech Recognition. Figure 4 shows that *VoiceSecure* achieves similar performance in safeguarding speaker identity and maintaining speech intelligibility. However, *VoiceSecure* outperforms in protecting speech content for automatic transcription.

5.5 Comprision with Benign Noises

In this section, we conduct a comparative analysis of *VoiceSecure*'s performance against common environmental noises, including Additive White Gaussian Noise, Babble noise, crowd talking, city sidewalks, and restaurants. As depicted in Figure 3, *VoiceSecure* consistently achieves a higher equal error rate and word error rate, indicating its superior efficacy in concealing sensitive information from automated systems. Despite a slight degradation in speech intelligibility, *VoiceSecure* maintains a sound quality nearly indistinguishable to human listeners.

6 RELATED WORK

This section discusses about the existing methods of voice anonymization and adversarial attacks on automatic speech recognition (ASR).

6.1 Voice Anonymization

Existing voice anonymization methods are primarily based on signal processing (SP), voice conversion (VC), and voice synthesis (VS). Signal processing methods [28, 36] directly apply signal processing techniques to modify speaker-related features such as formant positions, pitch, tempo, and pauses in speech audios to obscure voice prints. However, these methods usually induce large quality degradation as they fail

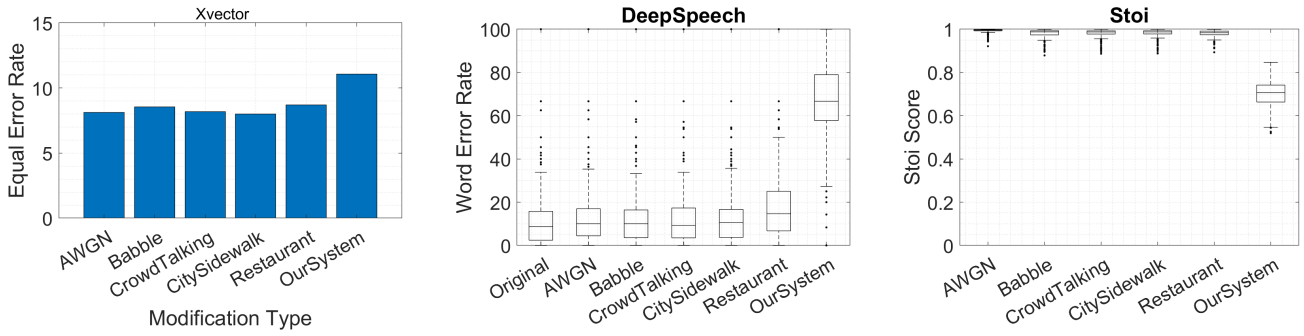

Figure 3: Comparing performance of *VoiceSecure* with benign noises.

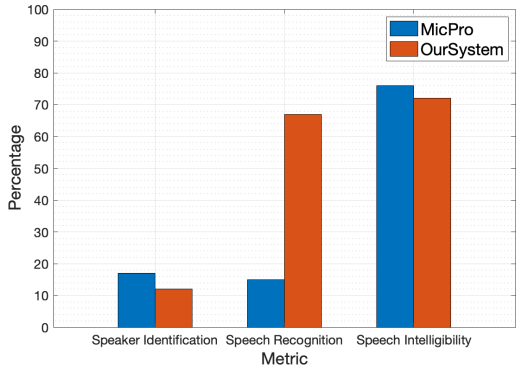

Figure 4: Comparison with the baseline

to consider the intelligibility and naturalness of the speech. Voice conversion [29, 30, 35, 41] and voice synthesis [14, 17, 20, 25] methods convert the original audio into another audio that sounds completely different from the original speaker. Although voice conversion and voice synthesis may achieve anonymity, they are not suitable for scenarios where the user wants to hide their identity from ASVs but hopes to preserve their natural voice to human audiences.

Apart from traditional signal processing, voice conversion, and synthesis methods, researchers have explored adversarial examples that can trick ASV into misidentifying speakers by adding imperceptible noise to the speech samples. However, existing ASV adversarial examples [8, 13, 22, 23, 44] are constructed via iterative updates, which cannot be used to achieve real-time voice anonymization. Recently, FAPG [40] and VCloak [11] developed a real-time voice anonymization system which achieves anonymity while preserving the intelligibility, naturalness, and timbre of the audio. Although these recent works have achieved voice anonymization, no system jointly protects speaker identity from ASV and the content of private speech from ASR.

6.2 Adversarial attacks on ASR

Adversarial attacks on automatic speech recognition involve creating examples that are transcribed differently by machines while appearing natural to humans. Previous research has explored various adversarial attack methods on speech recognition systems [37]. For instance, a study proposed the hidden voice command attack by designing obfuscated audio fragments to mislead GMM-based recognition systems while remaining unintelligible to humans [6]. Commandersong [43] investigated attacks on DNN-based speech recognition systems by adding adversarial perturbation on songs to deliver hidden commands. Adversarial audio examples were also optimized using gradient descent through the use of Connectionist Temporal Classification (CTC) loss [7]. Researchers have also looked at ways to improve the practicality and stealthiness of adversarial audio examples. Metamorph [9] studied mechanisms to enhance the survival of the adversarial audio examples in over-the-air transmission, while Schonherr et al. [33] adopted a psychoacoustic model lowering the signal guided by human hearing thresholds to increase the stealthiness of adversarial audio examples. Similarly, imperceptible and robust adversarial examples were generated by researchers to successfully attack the Lingvo ASR system in real-world scenarios [31]. Despite their success, they rely on white-box knowledge of the targets. To overcome this limitation, existing work has explored black-box attacks that exploit signal processing algorithms prior to the DNN-based classification stage [5]. Recently, Devil's Whisper [10] proposes a general approach for physical adversarial attacks by enhancing the simple local model approximating the target black-box model. Although the above attacks showed excellent results, none of these attacks generate adversarial perturbations in

real-time to prevent both automatic speaker verification and speech recognition.

Recently, SMACK [42] proposed a semantic adversarial perturbation attack aimed at deceiving automatic speech recognition (ASR) and automatic speaker verification (ASV) systems. Their approach involves modifying speech prosodies, such as tone, pitch, and phoneme durations, and is also not capable of generating perturbations in real-time, which limits its effectiveness in live call scenarios. In this paper, we propose a real-time system that can jointly deceive both speaker verification and speech recognition systems.

7 CONCLUSION

In conclusion, *VoiceSecure* proposes a microphone module that can protect user privacy by thwarting automatic systems from speaker identification and speech recognition while keeping the sound perceptually similar to humans. *VoiceSecure* can seamlessly integrate with existing devices either via audio jack or Bluetooth.

REFERENCES

[1] Does skype use encryption? https://support.skype.com/en/faq/FA31/does-skype-use-encryption. Last accessed 28 March 2023.

[2] Upstream vs. prism. https://www.eff.org/pages/upstream-prism, Oct 2017. Last accessed 28 March 2023.

[3] Secret evidence and the threat of more warrantless surveillance. https://www.hrw.org/news/2018/01/11/secret-evidence-and-threat-more-warrantless-surveillance, Oct 2020. Last accessed 28 March 2023.

[4] What is end-to-end call encryption? https://www.uctoday.com/unified-communications/what-is-end-to-end-call-encryption/#:~:text=VoIP%20phones%20use%20digital%2C%20encrypted,gold%20standard%20for%20protecting%20communication., Jun 2022. Last accessed 28 March 2023.

[5] Hadi Abdullah, Washington Garcia, Christian Peeters, Patrick Traynor, Kevin RB Butler, and Joseph Wilson. Practical hidden voice attacks against speech and speaker recognition systems. *arXiv preprint arXiv:1904.05734*, 2019.

[6] Nicholas Carlini, Pratyush Mishra, Tavish Vaidya, Yuankai Zhang, Micah Sherr, Clay Shields, David A Wagner, and Wenchao Zhou. Hidden voice commands. In *Usenix security symposium*, pages 513–530, 2016.

[7] Nicholas Carlini and David Wagner. Audio adversarial examples: Targeted attacks on speech-to-text. In *2018 IEEE security and privacy workshops (SPW)*, pages 1–7. IEEE, 2018.

[8] Guangke Chen, Sen Chenb, Lingling Fan, Xiaoning Du, Zhe Zhao, Fu Song, and Yang Liu. Who is real bob? adversarial attacks on speaker recognition systems. In *2021 IEEE Symposium on Security and Privacy (SP)*, pages 694–711. IEEE, 2021.

[9] Tao Chen, Longfei Shangguan, Zhenjiang Li, and Kyle Jamieson. Metamorph: Injecting inaudible commands into over-the-air voice controlled systems. In *Network and Distributed Systems Security (NDSS) Symposium*, 2020.

[10] Yuxuan Chen, Xuejing Yuan, Jiangshan Zhang, Yue Zhao, Shengzhi Zhang, Kai Chen, and XiaoFeng Wang. Devil's whisper: A general approach for physical adversarial attacks against commercial black-box speech recognition devices. In *USENIX Security Symposium*, pages 2667–2684, 2020.

[11] Jiangyi Deng, Fei Teng, Yanjiao Chen, Xiaofu Chen, Zhaohui Wang, and Wenyuan Xu. V-cloak: Intelligibility-, naturalness-& timbre-preserving real-time voice anonymization. *arXiv preprint arXiv:2210.15140*, 2022.

[12] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. *arXiv preprint arXiv:2005.07143*, 2020.

[13] Tianyu Du, Shouling Ji, Jinfeng Li, Qinchen Gu, Ting Wang, and Raheem Beyah. Sirenattack: Generating adversarial audio for end-to-end acoustic systems. In *Proceedings of the 15th ACM Asia Conference on Computer and Communications Security*, pages 357–369, 2020.

[14] Fuming Fang, Xin Wang, Junichi Yamagishi, Isao Echizen, Massimiliano Todisco, Nicholas Evans, and Jean-Francois Bonastre. Speaker anonymization using x-vector and neural waveform models. *arXiv preprint arXiv:1905.13561*, 2019.

[15] Barton Gellman and Ashkan Soltani. Nsa surveillance program reaches 'into the past' to retrieve, replay phone calls. https://www.washingtonpost.com/world/national-security/nsa-surveillance-program-reaches-into-the-past-to-retrieve-replay-phone-calls/2014/03/18/226d2646-ade9-11e3-a49e-76adc9210f19_story.html, Oct 2014. Last accessed 28 March 2023.

[16] DAVID GILBERT. Is skype safe and secure? what are the alternatives? https://www.comparitech.com/blog/information-security/is-skype-safe-and-secure-what-are-the-alternatives/, Apr 2020. Last accessed 28 March 2023.

[17] Yaowei Han, Sheng Li, Yang Cao, Qiang Ma, and Masatoshi Yoshikawa. Voice-indistinguishability: Protecting voiceprint in privacy-preserving speech data release. In *2020 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2020.

[18] Wenbin Huang, Wenjuan Tang, Hanyuan Chen, Hongbo Jiang, and Yaoxue Zhang. Unauthorized microphone access restraint based on user behavior perception in mobile devices. *IEEE Transactions on Mobile Computing*, 2022.

[19] Wenbin Huang, Wenjuan Tang, Kuan Zhang, Haojin Zhu, and Yaoxue Zhang. Thwarting unauthorized voice eavesdropping via touch sensing in mobile systems. In *IEEE INFOCOM 2022-IEEE Conference on Computer Communications*, pages 31–40. IEEE, 2022.

[20] Tadej Justin, Vitomir Štruc, Simon Dobrišek, Boštjan Vesnicer, Ivo Ipšić, and France Mihelič. Speaker de-identification using diphone recognition and speech synthesis. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, volume 4, pages 1–7. IEEE, 2015.

[21] Micah Lee and Yael Grauer. Zoom meetings aren't end-to-end encrypted, despite misleading marketing. https://theintercept.com/2020/03/31/zoom-meeting-encryption/, Mar 2020. Last accessed 28 March 2023.

[22] Zhuohang Li, Cong Shi, Yi Xie, Jian Liu, Bo Yuan, and Yingying Chen. Practical adversarial attacks against speaker recognition systems. In *Proceedings of the 21st international workshop on mobile computing systems and applications*, pages 9–14, 2020.

[23] Zhuohang Li, Yi Wu, Jian Liu, Yingying Chen, and Bo Yuan. Advpulse: Universal, synchronization-free, and targeted audio adversarial attacks via subsecond perturbations. In *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*, pages 1121–1134, 2020.

[24] Ian McLoughlin. *Speech and Audio Processing: a MATLAB-based approach*. Cambridge University Press, 2016.

[25] Xiaoxiao Miao, Xin Wang, Erica Cooper, Junichi Yamagishi, and Natalia Tomashenko. Language-independent speaker anonymization approach using self-supervised pre-trained models. *arXiv preprint arXiv:2202.13097*, 2022.

[26] Brian CJ Moore. An introduction to the psychology of hearing. academic press. *San Diego*, 1997.

[27] Lily Hay Newman. Hackers could decrypt your gsm phone calls. https://www.wired.com/story/gsm-decrypt-calls/, Aug 2019. Last accessed 28 March 2023.

[28] Jose Patino, Natalia Tomashenko, Massimiliano Todisco, Andreas Nautsch, and Nicholas Evans. Speaker anonymisation using the mcadams coefficient. *arXiv preprint arXiv:2011.01130*, 2020.

[29] Jianwei Qian, Haohua Du, Jiahui Hou, Linlin Chen, Taeho Jung, and Xiang-Yang Li. Hidebehind: Enjoy voice input with voiceprint unclonability and anonymity. In *Proceedings of the 16th ACM Conference on Embedded Networked Sensor Systems*, pages 82–94, 2018.

[30] Jianwei Qian, Haohua Du, Jiahui Hou, Linlin Chen, Taeho Jung, and Xiang-Yang Li. Speech sanitizer: Speech content desensitization and voice anonymization. *IEEE Transactions on Dependable and Secure Computing*, 18(6):2631–2642, 2019.

[31] Yao Qin, Nicholas Carlini, Garrison Cottrell, Ian Goodfellow, and Colin Raffel. Imperceptible, robust, and targeted adversarial examples for automatic speech recognition. In *International conference on machine learning*, pages 5231–5240. PMLR, 2019.

[32] David Rupprecht, Katharina Kohls, Thorsten Holz, and Christina Pöpper. Call me maybe: Eavesdropping encrypted lte calls with revolte. In *USENIX Security Symposium*, pages 73–88, 2020.

[33] Lea Schönherr, Katharina Kohls, Steffen Zeiler, Thorsten Holz, and Dorothea Kolossa. Adversarial attacks against automatic speech recognition systems via psychoacoustic hiding. *arXiv preprint arXiv:1808.05665*, 2018.

[34] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. X-vectors: Robust dnn embeddings for speaker recognition. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5329–5333. IEEE, 2018.

[35] Brij Mohan Lal Srivastava, Nathalie Vauquier, Md Sahidullah, Aurélien Bellet, Marc Tommasi, and Emmanuel Vincent. Evaluating voice conversion-based privacy protection against informed attackers. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2802–2806. IEEE, 2020.

[36] Tavish Vaidya and Micah Sherr. You talk too much: Limiting privacy exposure via voice input. In *2019 IEEE Security and Privacy Workshops (SPW)*, pages 84–91. IEEE, 2019.

[37] Donghua Wang, Rangding Wang, Li Dong, Diqun Yan, Xueyuan Zhang, and Yongkang Gong. Adversarial examples attack and countermeasure for speech recognition system: A survey. In *Security and Privacy in Digital Economy: First International Conference, SPDE 2020, Quzhou, China, October 30–November 1, 2020, Proceedings*, pages 443–468. Springer, 2020.

[38] Frederick Andrew White et al. *Our acoustic environment.* 1975.

[39] Shilin Xiao, Xiaoyu Ji, Chen Yan, Zhicong Zheng, and Wenyuan Xu. Micpro: Microphone-based voice privacy protection. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, pages 1302–1316, 2023.

[40] Yi Xie, Zhuohang Li, Cong Shi, Jian Liu, Yingying Chen, and Bo Yuan. Enabling fast and universal audio adversarial attack using generative model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14129–14137, 2021.

[41] In-Chul Yoo, Keonnyeong Lee, Seonggyun Leem, Hyunwoo Oh, Bonggu Ko, and Dongsuk Yook. Speaker anonymization for personal information protection using voice conversion techniques. *IEEE Access*, 8:198637–198645, 2020.

[42] Zhiyuan Yu, Yuanhaur Chang, Ning Zhang, and Chaowei Xiao. Smack: Semantically meaningful adversarial audio attack. In *USENIX Security Symposium*, 2023.

[43] Xuejing Yuan, Yuxuan Chen, Yue Zhao, Yunhui Long, Xiaokang Liu, Kai Chen, Shengzhi Zhang, Heqing Huang, Xiaofeng Wang, and Carl A Gunter. Commandersong: A systematic approach for practical adversarial voice recognition. In *27th {USENIX} Security Symposium ({USENIX} Security 18)*, pages 49–64, 2018.

[44] Baolin Zheng, Peipei Jiang, Qian Wang, Qi Li, Chao Shen, Cong Wang, Yunjie Ge, Qingyang Teng, and Shenyi Zhang. Black-box adversarial attacks on commercial speech platforms with minimal information. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, pages 86–107, 2021.

[45] Eberhard Zwicker and Hugo Fastl. *Psychoacoustics: Facts and models*, volume 22. Springer Science & Business Media, 2013.