USENIX

THE ADVANCED COMPUTING
SYSTEMS ASSOCIATION

For Human Ears Only: Preventing Automated Monitoring on Voice Data

Irtaza Shahid and Nirupam Roy, *University of Maryland, College Park*

https://www.usenix.org/conference/usenixsecurity25/presentation/shahid

This paper is included in the Proceedings of the
34th USENIX Security Symposium.

August 13–15, 2025 • Seattle, WA, USA

978-1-939133-52-6

Open access to the Proceedings of the
34th USENIX Security Symposium is sponsored by USENIX.

For Human Ears Only: Preventing Automated Monitoring on Voice Data

Irtaza Shahid, Nirupam Roy

University of Maryland, College Park
{irtaza, niruroy}@umd.edu

Abstract

As voice communication becomes an essential part of modern life, the exposure of sensitive information through audio calls presents significant privacy risks. Malicious actors can gain access to this data by compromising user devices, exploiting communication channels, or attacking data servers, making it vulnerable to automated monitoring systems that can identify speakers and extract speech content. To address these privacy concerns, we introduce *VoiceSecure*, the first microphone module designed to prevent automated monitoring of speech while preserving its natural sound for humans. By leveraging the principles of human auditory perception, *VoiceSecure* employs a set of speech modifications that are adaptively tuned in real-time to obscure speaker identity and speech content, without compromising the quality of the audio for human listeners. This hardware-based solution mitigates the risk of software-based attacks, integrating seamlessly with commercial devices via audio jack or Bluetooth. Comprehensive evaluation across state-of-the-art speaker verification and speech recognition models, and a variety of speech datasets, demonstrates that *VoiceSecure* outperforms traditional methods of protecting speech from automated monitoring while keeping it intelligible for humans.

1 Introduction

Online voice communication has assumed a fundamental role in modern life, bridging distances with immediacy and efficiency. Unlike text, which is stripped of tone and subtlety, voice conveys intent and emotion with clarity, fostering more genuine connections and direct exchanges. Besides traditional phone calls, teleconferencing platforms like Zoom, Teams, and Discord now serve as pillars of communication, quietly reshaping interactions within both personal and professional spheres. Being a rich source of personal and professional information, online transfer of voice data makes it vulnerable to various exploitations.

Today, over 95% of voice data is transferred digitally; with the widespread adoption of technologies like VoIP (Voice over Internet Protocol) [4], almost all phone calls, even traditional landline calls, are converted to digital signals for transmission over the internet. This makes the data readily available to computers at various stages of the transmission, starting from the user devices to internet middleboxes, application servers, and even eavesdropping on transmission channels. It has raised concerns about information security. Advancements in voice-to-text technologies and natural language processing algorithms make it possible to launch online monitoring and mass surveillance of digital voice data. Voice is not only a medium for information it carries, it also reveals users' emotions and nonverbal cues of personal feelings [83]. Moreover, voice is also a biometric identity of the person. Past works have shown that these nonverbal information in voice can be extracted computationally [11, 33, 71]. Unauthorized voice monitoring is a growing threat to user privacy as it can capture sensitive information, including secret government discussions, personal conversations, financial data, or health-related topics, exposing users to identity theft, information leaks, data profiling, or unwarranted surveillance. Moreover, without user consent, organizations or third parties may exploit voice data to train highly personalized profiles or avatars, potentially infringing on individual rights to privacy and autonomy.

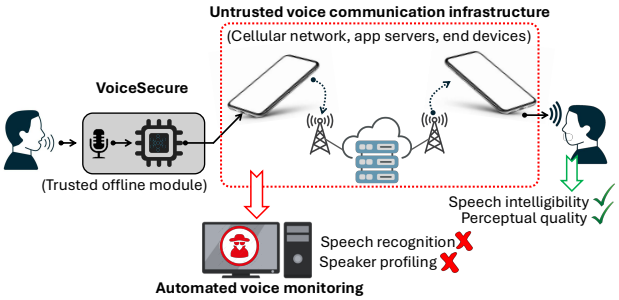

Figure 1: *VoiceSecure*, an offline hardware-software solution to protect speech privacy in real-time voice communication.

Recently surfaced events of digital espionage [44], monitoring [68], privacy breach [23], and mass-surveillance [34, 65]

Method	Speaker Identification	Speech Recognition	Naturalness	Real-time	Hardware
McAdams [47]	✓	✗	✓	✓	✓
VoiceMask [51]	✓	✓	✗	✗	✗
VCloak [19]	✓	✗	✓	✗	✗
SMACK [80]	✓	✓	✓	✗	✗
Stop Bugging Me [38]	✓	✗	✗	✓	✓
MicPro [76]	✓	✗	✓	✓	✓
***VoiceSecure* (Ours)**	✓	✓	✓	✓	✓

Table 1: Comparison of *VoiceSecure* with existing methods for protecting speech privacy. This table highlights the need for real-time speech protection methods while maintaining the natural flow of communication.

on voice data have triggered actions by service providers. The primary mode of this action is implementing data encryption. Voice communication services, such as WhatsApp and Facebook, offer end-to-end voice data encryption [1]. However, this does not fully address the overall vulnerability. For instance, Zoom fails to ensure end-to-end security when a resource-constrained device (e.g., a smartphone) is involved in the conversation [35]. Similarly, when making calls from Skype to mobile or landline phones, the portion of the call that passes through an unencrypted Public Switched Telephone Network (PSTN) makes it susceptible to basic eavesdropping attacks [28]. Moreover, all of these end-to-end encryption models require users to trust the application servers and service providers' infrastructure, leaving a loophole in the entire system anyway. On the other hand, several attacks on cellular telephony, including Voice over LTE (VoLTE) [56] and 5G voice transmission [45], show the possibility of voice eavesdropping.

The lack of trust in the network and application infrastructure has led to an alternative approach to voice data privacy – elimination of sensitive information before transmission. Some techniques show ways to prevent automatic speaker identification [15, 19, 22, 77] from the data, and some other techniques aim to prevent automatic speech recognition [16, 18, 80] from the shared data. It is a promising concept that aims to understand the capabilities of AI-enabled attacks on data privacy and propose a countermeasure by pre-processing the voice data to mask sensitive features. Unfortunately, these approaches are computationally heavy and require an offline pre-processing stage, preventing their real-time application in an online voice conversation. Attempts at real-time applications are caught in the three-way tradeoff between acceptable latency for voice communication, computation capacity of a laptop or a smartphone, and efficiency of attainable feature masking. Moreover, the solution by design relies on online processing and the user device, which itself can be compromised through malware [30, 31]. We ask the question: ***Is it possible to develop a zero-trust real-time system to prevent both automated voice monitoring and speaker profiling?***

In this paper, we introduce *VoiceSecure*, an innovative hardware-software module designed to provide real-time modification on voice data that protects user privacy from automatic speech recognition and speaker profiling while keeping speech natural for the human listener. This process requires the elimination of machine-identifiable voice personalization features and components that are essential for automatic speech recognition. We call this process 'feature-redaction' of the voice data. This ensures users can enjoy the same audio call experience without worrying about privacy breaches over voice calls. We envision the *VoiceSecure* module as an external microphone, capturing speech audio, applying feature-redaction in real-time, and transmitting it to the user's device via either an audio jack or Bluetooth connection. Figure 1 shows an application scenario.

Developing this system presents several challenges, primarily due to the limited capabilities of the hardware module. Current speech anonymization and masking techniques [19, 77, 80] either only protect speaker identity or speech content and are often too computationally intensive for real-time processing on resource-constrained devices, such as in live calls. Table 1 summarizes these methods, detailed in Section 11. Therefore, a new, computationally efficient feature-redaction technique must be designed. This approach needs to jointly optimize to thwart both automatic speaker identification and speech recognition while ensuring real-time performance. It is crucial that the feature-redaction process operates seamlessly, preserving the natural flow of communication without disrupting the call experience. Maintaining this imperceptibility is essential to uphold the naturalness of the communication.

To address this challenge, we start by understanding the inherent differences between how the human brain perceives speech sounds and how machines such as speaker identification and speech recognition models process speech signals [39, 39, 85]. Specifically, *VoiceSecure* adopts a multi-step approach. Initially, *VoiceSecure* eliminates inaudible frequency components from the speech, leveraging the human tendency to focus on the high-energy frequency and ignoring low-energy components

in understanding speech content. 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 are not noticeable by humans. *VoiceSecure* then design a reinforcement learning agent to control these variations in real-time based on the input speech signal to optimally prevent speech from automated monitoring while keeping it intelligible to humans.

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 propose a set of signal processing-based modifications for speech feature-redaction leveraging inherent properties present in human perception.
- We design a reinforcement learning model to control speech modifications in real-time to maximize speech privacy while maintaining the naturalness of the speech.
- We implement *VoiceSecure* on an off-the-shelf microcontroller, and evaluate its performance under practical scenarios.

2 Adversary Model

Our adversary model assumes a highly skilled adversary capable of accessing and analyzing voice calls, which often contain sensitive information. The adversary could extract distinctive voiceprints from the audio and conduct replay or mimicry attacks on speaker verification systems commonly used for authentication. Additionally, the intercepted speech could be exploited to infer sensitive information, such as financial details during a bank call or health issues during a medical consultation. Once linked to specific speakers, this sensitive information could lead to targeted and malicious attacks.

We assume that an adversary could access the audio through various methods, such as exploiting vulnerabilities in the communication channel, compromising the user's device, or infiltrating the service provider, as illustrated in Figure 2. Existing research [45, 56] highlights vulnerabilities in GSM and LTE networks that attackers might exploit to intercept sensitive voice data. Additionally, our model accounts for scenarios where an adversary might compromise the user's device to directly access raw audio samples. We also consider the possibility of an attacker infiltrating the service provider, or situations where the service provider itself may access the audio data [3, 27]. This comprehensive threat model covers both network, device, and server-level attacks, ensuring robust protection of speech privacy throughout the communication process. It is crucial to note here that we are not preventing attackers from accessing the speech data, our focus is on preventing automated monitoring systems from analyzing the speech content using advanced speaker verification and speech recognition techniques.

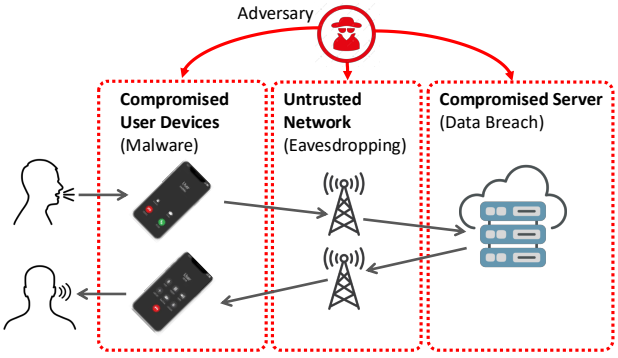

Figure 2: *VoiceSecure* assumes that a skilled attacker can access audio by compromising the user devices, exploiting vulnerabilities in the communication channel, or infiltrating the data server.

Moreover, our adversary model assumes a black-box scenario, allowing the attacker to employ any state-of-the-art model for speaker verification and speech recognition. We make no assumptions regarding the specific models or algorithms the attacker may use, leaving them free to apply advanced machine-learning techniques or pre-trained models to analyze intercepted audio. This approach ensures that our system is resilient against a wide array of potential adversarial methods, protecting speech privacy against a broad spectrum of strategies.

In addressing these concerns, our proposed solution must fulfill three criteria:

(1) Speech Anonymity: The solution must prevent the adversary from inferring speaker identities and sensitive content from the speech audio.

(2) 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.

(3) Security: The solution should apply feature-redaction 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, whether during traditional phone calls or through advanced teleconferencing applications. However, our solution extends beyond just audio calls, offering protection in scenarios where adversaries gain access to audio content uploaded by users on social media platforms. While existing

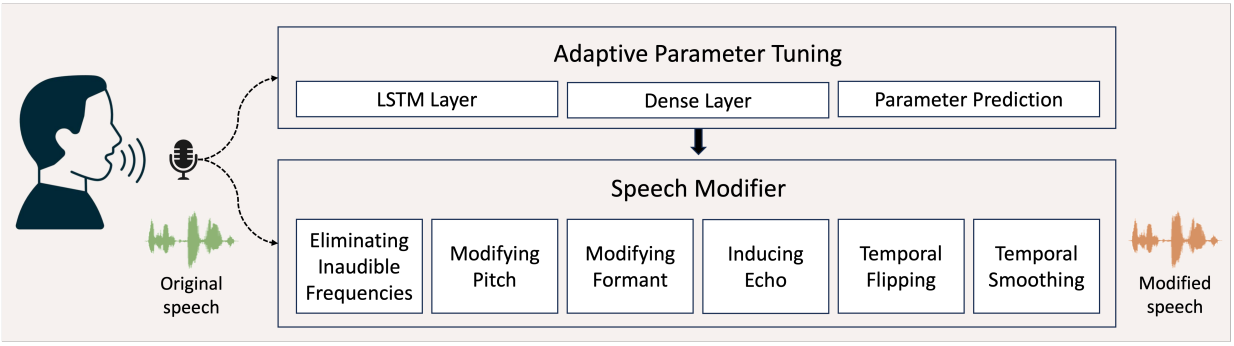

Figure 3: Design Overview: *VoiceSecure* comprises of two modules. The "Adaptive Parameter Tuning module" listens to the speech audio and predicts the optimal parameters for speech modification, and then the "Speech Modifier module" applies a set of speech modifications based on the given parameters in real-time to generate a modified speech.

solutions may address similar scenarios, we emphasize that *VoiceSecure* provides a versatile and adaptable solution applicable across a wide range of applications.

Although *VoiceSecure* is developed primarily for English, its design is inherently language-agnostic. This is due to the significant phoneme overlap among languages and the universal nature of psychoacoustic principles that guide speech perception. As such, the underlying mechanisms of *VoiceSecure* can be applied to other languages with minimal adaptation, making it suitable for multilingual environments.

VoiceSecure does not account for scenarios where a human attacker listens to the speech audio to identify the speaker or access sensitive content. This situation is less relevant to our application scenario, which focuses on protecting users against mass surveillance. Assigning human operators to listen to all user conversations is both inefficient and impractical at scale. However, we acknowledge that our system does not safeguard speech privacy from humans. Our primary goal is to protect speech against automated monitoring without affecting the flow of communication.

We also acknowledge the risk posed by local eavesdroppers such as compromised smart home devices or intelligent voice assistants. These devices, if compromised, could potentially capture speech before transformation is applied. To mitigate this, we envision *VoiceSecure* as a lightweight, pluggable module that can be integrated directly into local devices, including phones, laptops, or hardware peripherals, ensuring that transformation occurs at the source before any data is transmitted or accessed. Additionally, users can complement this protection with existing privacy controls and device-level settings to limit unwanted data collection from local endpoints.

3 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 not to interfere with user experience. *VoiceSecure* targets to apply feature-redaction before getting accessed by the device software, mitigating the possibility of any software-based spoofing attacks.

We envision *VoiceSecure* as a compact low-power microphone module, similar to any off-the-shelf headphones, that can seamlessly integrate with existing devices such as phones, laptops, and tablets via audio jack or Bluetooth. This module captures speech audio, applies carefully designed speech modifications in real-time, and transmits the altered audio to the connected device, ensuring no disruption to the natural call experience. Figure 3 presents a systematic overview of *VoiceSecure*, in which the microphone captures the speech signal, the adaptive parameter tuning module predicts the modification parameters, and the speech modifier uses these parameters to generate a modified speech.

The development of *VoiceSecure* is divided into three phases:

Designing Speech Modifications: This phase involves understanding how humans perceive speech signals and how this differs from machine-based processing. Leveraging this understanding, we design a set of speech alterations that effectively deceive automatic speaker identification and speech recognition systems. These modifications are carefully crafted to maintain perceptual similarity for human listeners while disrupting the accuracy of automated systems, ensuring both privacy and usability.

Adaptive Parameter Tuning: The primary objective of this module is to adaptively adjust modification parameters in real-time based on the input speech signal to achieve maximum speech privacy while preserving the natural flow of communication for humans.

***VoiceSecure* Module Development:** This phase involves the implementation of designed modifications on a microcontroller that can apply these modifications in real-time and can seamlessly integrate with existing call devices.

4 Fundamental Concepts and Primers

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

4.1 Human Speech Basics

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. A phoneme is the smallest unit of sound in a language that can distinguish meaning. 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. Figure 4 shows a spectrogram of the speech and its corresponding formants.

Beyond recognition and verification, speech and acoustic signals have been explored in diverse domains, including health monitoring [14, 59], security and privacy [54, 61, 62], acoustic sensing [9, 10, 25, 26], and spatial perception [17, 41, 60, 67, 78]. These efforts expand the utility of sound beyond communication, opening up novel directions at the intersection of acoustics and human-computer interaction.

4.2 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 [20, 63]. Deep neural networks

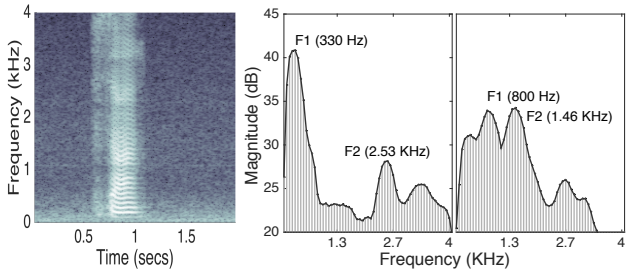

Figure 4: a) The spectrogram of the spoken consonant 's' followed by the vowel 'a' recorded with a microphone, b) The locations of the first two formants (F1 and F2) for the vowel sound *'i'* and *'a'* [55].

are trained to compute hierarchical features from the speech and learn robust speaker representations that are agnostic to the speech content, recording channel, and ambient noise.

4.3 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.

5 Designing Speech Modifications

Designing effective speech modifications is a pivotal aspect of *VoiceSecure*, as it seeks to obscure critical information from automated speaker identification and speech recognition systems while maintaining intelligibility and naturalness for human listeners. Simply adding noise or indiscriminately altering speech signals would not only prevent automatic systems from extracting sensitive information but could also disrupt communication for users. Therefore, *VoiceSecure* leverages a more sophisticated approach by carefully crafting speech modifications that strike a delicate balance between security and naturalness.

To achieve this, *VoiceSecure* capitalizes on the inherent differences between human auditory perception and how automated systems process speech signals. While automated systems analyze speech features such as pitch, formants, and spectral patterns, the human brain processes speech in a fundamentally different way, heavily relying on auditory cues that are not necessarily tied to the raw features in the signal. By exploiting these differences, *VoiceSecure* ensures that

modifications can obscure speech from automated systems while preserving its intelligibility for human listeners.

5.1 Psychoacoustic Speech Perception

Human auditory perception is marked by complex non-linearities that distinguish how we hear sounds from how they might be physically represented or recorded. The field of psychoacoustics investigates these unique perception processes, revealing that the brain does not process speech simply as a series of sound waves, but instead as a set of complex, context-dependent features [39, 85]. Further, research shows that the brain has specialized mechanisms for processing speech, which are distinct from how non-speech sounds are perceived [42]. Drawing from extensive psychoacoustic literature, *VoiceSecure* leverages these insights to design speech modifications that evade automated systems while maintaining the naturalness of the audio for human listeners.

The key psychoacoustic principles foundational to design *VoiceSecure*'s speech modifications are as follows [39].

Fundamental Frequency Perception: Humans can tolerate slight variations in fundamental frequencies, perceiving two complex tones as distinct yet similar even when their frequencies are slightly mistuned [39]. This tolerance allows for subtle pitch alterations without significantly affecting intelligibility.

Closure Principle: The brain uses the closure principle to fill in missing auditory information, maintaining sound continuity even when segments of speech are missing. This principle allows for small disruptions in the signal without compromising speech comprehension.

Speech Signal Reversal: Reversing small speech segments still maintains intelligibility because speech is processed in phoneme-sized blocks. This property allows for the introduction of small distortions that are imperceptible to listeners but can confuse automatic systems.

Haas Effect: The Haas effect demonstrates that slight delays between two versions of a sound arriving at the ear can influence its perceived spatial and temporal attributes [75]. By exploiting this phenomenon, we can create echoes that deceive speaker identification systems without affecting human perception.

These insights underscore the complex nature of human auditory perception and provide invaluable insights into the development of *VoiceSecure*.

5.2 Implementing Speech Modifications

Building on the psychoacoustic principles outlined above, *VoiceSecure* employs a multi-faceted approach to speech

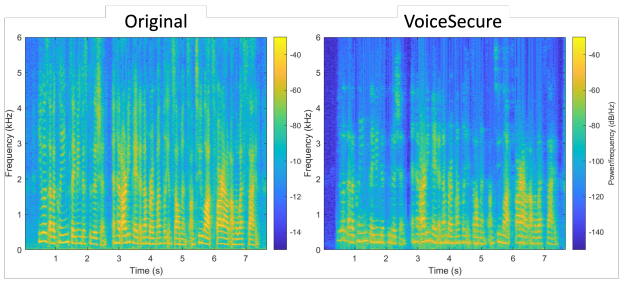

Figure 5: The spectrogram comparison between the original speech and the speech modified by *VoiceSecure* reveals that the modified speech maintains a high degree of perceptual similarity.

modification that targets both the human auditory system and automated recognition systems.

Eliminating Inaudible Frequency Components: *VoiceSecure* starts by removing inaudible low-amplitude frequencies from the speech signal, capitalizing on the human tendency to focus on higher-energy voiced regions for comprehension. These low-amplitude frequencies do not contribute significantly to speech intelligibility and can be discarded without perceptibly affecting the human listener's understanding of the message [85].

Manipulating Pitch and Formants: Automatic speaker recognition systems rely heavily on the pitch and formant structures of speech. *VoiceSecure* subtly modifies these features to confuse speaker identification systems. While pitch alterations can be detected by humans, the range of modifications is designed to remain within a perceptible range, preserving intelligibility while disrupting recognition by automated models [39].

Introducing Temporal Distortions: Temporal distortions, such as flipping small time-domain speech windows, are introduced based on psychoacoustic principles [39]. These distortions can disorient automatic speech recognition systems, which typically depend on continuous temporal features while remaining undetectable to the human listener.

Utilizing the Haas Effect: By introducing slight delays between the two versions of the speech signal, *VoiceSecure* generates echoes that act as background noise for automatic systems. These echoes effectively obscure the origin of the speech, complicating automatic speaker recognition processes. However, these modifications remain imperceptible to human listeners, as the delay is minimal and does not significantly affect the naturalness of the speech [75].

Selecting Modification Windows: Drawing from the closure principle, *VoiceSecure* selects small speech windows (approximately 50 milliseconds) for modification. These windows can be altered with minimal perceptible effect on human listeners, as small variations within these intervals go unnoticed by the brain. By strategically selecting modification windows with appropriate intervals between them, *VoiceSe-*

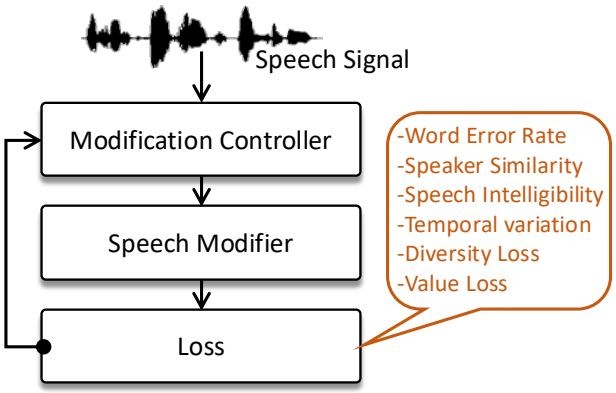

Figure 6: The training process of the Adaptive Parameter Tuning module involves feeding in a speech signal, applying speech modifications, and optimizing a set of loss functions to fine-tune the modification parameters.

cure ensures natural speech flow and intelligibility.

Combining Time Windows: To prevent abrupt transitions between modified speech windows, which could introduce perceptible artifacts, *VoiceSecure* uses cubic spline fitting to smooth out the transitions. This technique ensures that modifications are seamlessly integrated into the speech, avoiding unnatural jumps or glitches that might otherwise disrupt the user experience. Figure 5 shows the spectrogram of speech before and after applying the designed modifications. It clearly highlights that the modified speech maintains a high degree of perceptual similarity to the original audio.

While randomly applying these modifications could result in significant distortions of the speech signal, altering the perception of the speaker's identity or disrupting the natural flow of conversation, *VoiceSecure* carefully tunes these parameters. This meticulous adjustment ensures that the modifications are effective at deceiving automated systems while remaining imperceptible and non-intrusive for humans.

6 Adaptive Parameter Tuning

VoiceSecure employs a suite of speech modification techniques, each controlled by a distinct set of parameters. These parameters dictate key aspects of speech transformation, including the minimum energy in audible frequencies, the degree of pitch and formant shifts, and the timing and strength of added echoes. The optimal effectiveness of *VoiceSecure* hinges on dynamically determining the ideal combination of these parameters in real-time, based on the characteristics of the input speech. This approach is essential for protecting user privacy against speaker identification and automatic speech recognition (ASR) systems while preserving audio quality for humans.

A fixed set of parameters, though potentially effective in certain scenarios, poses limitations. Primarily, static modifications lack adaptability, allowing adversaries to potentially reverse-engineer these patterns and restore the original speech content. To address these challenges, we developed a machine learning model that dynamically predicts optimal modification parameters, adjusting in real-time to the unique properties of each speech segment. This model processes each incoming speech window, predicts the optimal modification parameters, and passes these to the modifier to transform the speech accordingly. This approach continuously achieves an optimal balance between security and intelligibility, safeguarding the speech signal from speaker identification and ASR systems without compromising naturalness or clarity for human listeners.

Our model's training process is guided by a loss function designed to balance speech privacy and intelligibility preservation. The goal is to modify speech in a way that degrades automated model performance (e.g., speaker identification or ASR), while ensuring the output remains intelligible to humans. We formulate the loss as follows:

$$\min_{p \in P} -(\text{WER}(\tilde{x},x) - \text{Similarity}(\tilde{x},x) + \text{STOI}(\tilde{x},x))$$

Here, x represents the original speech signal, and p denotes the set of modification parameters. The function $f(x,p)$ transforms x into a modified version $\tilde{x}$, designed to reduce recognizability while preserving intelligibility. The loss incorporates three components: The term $WER(\tilde{x},x)$ the word error rate between the original and modified speech, $Similarity(\tilde{x},x)$ which captures the cosine similarity in speaker identity, and $STOI(\tilde{x},x)$, the Short-Time Objective Intelligibility metric, which quantifies how understandable the speech remains to human listeners.

We aim to maximize WER and minimize speaker similarity to ensure privacy from automated monitoring systems, while maximizing STOI to preserve human intelligibility. This formulation discourages excessive modifications that would degrade intelligibility, and likewise discourages leaving the speech unchanged, which would preserve intelligibility but fail to reduce recognizability. Together, these terms guide the model to learn transformations that effectively protect both content and identity while maintaining clear and understandable speech.

6.1 Model Architecture and Framework

Designing and training a model to dynamically predict modification parameters for speech presents a new challenge. The model must process each speech window, predict optimal modification parameters, apply these through a speech modifier function, and evaluate the modified speech in terms of Word Error Rate, speaker similarity, and Short-Time Objective Intelligibility. However, the non-differentiable nature of the speech modifier and evaluation metrics makes conventional backpropagation infeasible, highlighting the need for an alternative approach.

Instead, our problem aligns naturally with a reinforcement learning framework, where each speech window represents a state, modification parameters serve as actions, the speech modifier acts as the environment, and rewards reflect our objective of maximizing WER, STOI, and minimizing speaker similarity. This approach allows the model to iteratively refine its predictions through trial and error, bypassing the non-differentiability challenge. Figure 6 illustrates the training process for our system.

There are various reinforcement learning strategies to address such challenges, including Q-learning, policy gradient approaches, and hybrid techniques such as actor-critic. While Q-learning is effective for discrete action spaces, it struggles to handle continuous actions efficiently, which are central to our task. Policy gradient methods, on the other hand, directly optimize actions but struggle due to high variance in gradient estimates, leading to unstable training. Among these strategies, the actor-critic framework emerged as the most suitable for our system. In this architecture, the actor predicts the modification parameters for each speech window, while the critic evaluates the value of each state, providing feedback to guide the actor toward better actions. The actor-critic framework's ability to balance exploration and exploitation, coupled with its effectiveness in handling continuous action spaces, makes it ideal for predicting smooth and dynamic modification parameters. This approach enables the model to progressively refine its strategies while optimizing for privacy and intelligibility.

To enhance the model's contextual awareness, we incorporated a long short-term memory (LSTM) layer into the Actor-Critic framework, allowing the model to capture temporal dependencies between windows. Additionally, each state includes the past 'N' speech windows and corresponding 'N' actions, equipping the model with historical information necessary for making well-informed parameter predictions. This architectural choice enables the model to consider both current and past speech features along with previous modifications, resulting in smoother, temporally coherent transformations across windows. Figure 7 illustrates the model architecture, which takes as input the current speech window (1024 raw samples), two past speech windows (2×1024), and two past action predictions (2×6). The architecture consists of a single LSTM layer with 256 units, followed by two output layers: one of size 6 for action prediction and another of size 1 for the value estimate. We set the window size to 1024, as larger windows increase computational cost and risk introducing perceptual distortions. Similarly, we use two past windows to provide sufficient context for smoother modification, while avoiding the overhead associated with longer temporal dependencies. Our system is trained to generalize across all users, eliminating the need for user-specific model retraining.

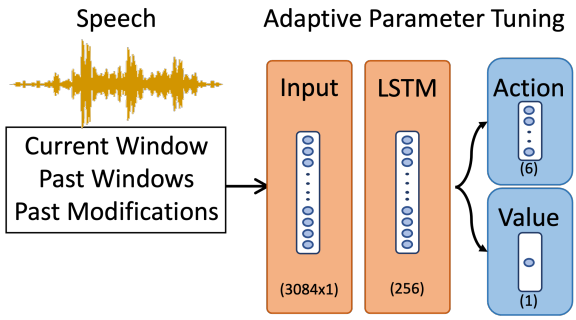

Figure 7: The Adaptive Parameter Tuning model consists of an input layer that processes the current speech window along with past windows and past actions. This is followed by an LSTM layer and two output layers: an action layer for predicting modification parameters and a value layer for tuning the model.

6.2 Customized Loss for Robust Learning

The model's training loss comprises multiple components designed to encourage both security and naturalness. The primary loss term, a combination of Word Error Rate (WER), speaker similarity, and Short-Time Objective Intelligibility (STOI), serves as the main objective, which we aim to minimize. This primary loss is formulated as:

$$\mathcal{L}_p = -\left(\text{WER}(\tilde{x},x) - \text{Similarity}(\tilde{x},x) + \text{STOI}(\tilde{x},x)\right)$$

where x represents the original speech signal and $\tilde{x}$ denotes the modified speech signal.

To ensure that the model does not converge on a fixed set of parameters for an entire speech signal, we introduce a temporal variation loss. This component encourages diversity in the parameters across windows, helping to avoid a static, easily reversible transformation pattern. The temporal variation loss is computed as the sum of the squared differences between consecutive predicted parameters:

$$\mathcal{L}_t = -\frac{1}{T}\sum_{t=1}^{T} \parallel p_{t+1} - p_t \parallel^2$$

where p_t and p_{t+1} are the predicted parameters at consecutive time steps t and $t+1$. By promoting temporal variation, this loss ensures the model generates dynamic transformations over time, preventing predictable or static patterns. However, large variations can result in perceptible distortions or "jitters" in the speech signal.

To mitigate this, we introduce a target-change constraint to regulate the magnitude of variations in the parameters. This constraint minimizes excessive shifts, preserving the naturalness and consistency of the audio while maintaining diversity. The target-change loss can be formulated as:

$$\mathcal{L}_t = \frac{1}{T}\sum_{t=1}^{T} \parallel\mid p_{t-1} - p_t \mid - \Delta_{target} \parallel^2$$

Where Δ_{target} is the maximum allowable change in parameters between consecutive windows. This constraint limits the degree of modification, ensuring the speech remains intelligible.

We also apply a diversity loss to prevent random oscillations and promote the exploration of a broader parameter space. By incorporating the mean parameters of past N windows $\overline{p}$, this loss encourages diversity and randomness in the predicted modifications. The mean-action loss is given by:

$$\mathcal{L}_d = -\frac{1}{T}\sum_{t=1}^{T} \| p_t - \overline{p}_t \|^2$$

Finally, we use a value loss for the critic model, which ensures that the critic accurately estimates the expected reward for each state. This value loss is calculated as the squared error between the predicted reward and the actual reward, facilitating stable learning in the actor-critic setup. The value loss is formulated as:

$$\mathcal{L}_v = \| R_{pred} - R_{actual} \|^2$$

The final loss function combines these components, producing a robust training objective that optimally balances privacy preservation with intelligibility. The overall training objective is then:

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_p + \lambda_2 \mathcal{L}_t + \lambda_3 \mathcal{L}_d + \lambda_4 \mathcal{L}_v$$

Where λ are hyperparameters that control the relative importance of each loss component. This final loss function ensures that the model effectively balances privacy and intelligibility in the generated speech. We train our model on the LibriSpeech 40-speaker dataset [46], using X-vector and DeepSpeech to compute losses based on speaker similarity and word error rate, respectively. To evaluate the robustness and generalizability of our system, we test it across four diverse speech datasets (Table 3) and compare its performance against three state-of-the-art speaker identification and speech recognition systems (Table 2). Sections 8 and 9 detail the experimental setup and present the evaluation results.

7 *VoiceSecure* Prototype Development

We developed a compact microphone module that can be easily integrated with existing devices through either an audio jack or Bluetooth, making it highly versatile for diverse communication platforms. This module captures live speech audio, applies frame-by-frame modifications in real time, and transmits the modified output to any connected device. To perform speech processing, we implemented the modifications in Python.

To optimize for lightweight, efficient deployment, we trained our adaptive speech modification model in PyTorch and converted it to ONNX format [21], reducing its size to 12 MB while preserving functionality. This ONNX format is compatible with a range of hardware, allowing for real-time performance on resource-constrained devices.

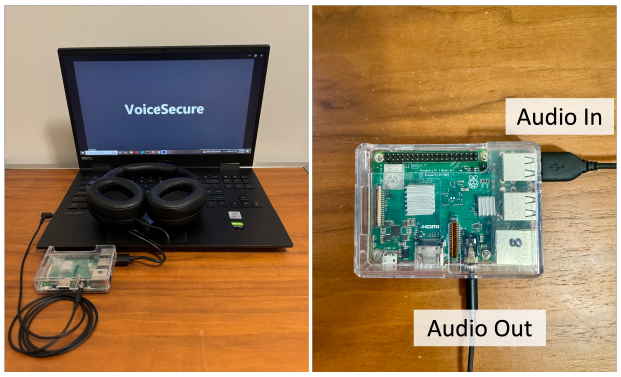

Figure 8: *VoiceSecure* module: (left) connected to commercial headphones during an audio call; (right) close-up of the module showing two audio ports.

For prototyping, we used a Raspberry Pi 3 B+ [48] equipped with a 1.4GHz quad-core processor and 1GB of RAM. With a connected commercial microphone, the Raspberry Pi enables seamless integration via both audio jack and Bluetooth, providing a flexible platform for implementing *VoiceSecure*'s feature-redaction techniques in compact, standalone modules. Figure 8 illustrates our module, with a commercial microphone connected to the system (left) and the processing module performing real-time modifications (right).

8 Experimental Setup

8.1 State-of-the-art Speech Models

To develop and assess *VoiceSecure*, we employed a selection of state-of-the-art models for both speaker verification and speech recognition, as listed in Table 2. For speaker verification, we used X-Vector [63], ECAPA-TDNN [20], and I-Vector [72]. **X-Vector** is a neural network architecture designed to extract speaker embeddings, capturing specific vocal characteristics for accurate speaker verification across diverse acoustic conditions. **ECAPA-TDNN** builds on this by adding channel attention and aggregation mechanisms, enhancing its ability to distinguish between speakers even in challenging audio environments. Additionally, we used the **I-Vector** as a baseline, providing a compact representation of speaker characteristics and enabling benchmarking against the more advanced neural models.

For speech recognition, we incorporated DeepSpeech [6], Whisper [53], and Wav2Vec 2.0 [8] to evaluate transcription accuracy. **DeepSpeech** is an end-to-end model based on a recurrent neural network (RNN) that translates speech spectrograms into text, offering reliable transcription

Speaker Verification	Speech Recognition
X-Vector [63]	DeepSpeech [6]
ECAPA-TDNN [20]	Whisper [53]
I-Vector [72]	Wav2Vec2 [8]

Table 2: List of state-of-the-art speaker verification models and speech recognition systems used for evaluation.

DataSet	# Speakers	# Utterances	Usage
LibriSpeech [46]	40	2703	Train
LibriSpeech [46]	900	27000	Test
VoxCeleb1 [43]	1211	6054	Test
CommonVoice [7]	-	2945	Test
VCTK [70]	110	5000	Test

Table 3: List of commonly used speech datasets for evaluation.

for continuous speech. **Whisper**, developed by OpenAI, leverages a transformer architecture for high-accuracy transcription across languages and accents, with robustness in noisy, real-world conditions. Finally, **Wav2Vec 2.0**, a self-supervised model trained directly on raw audio, excels in transcribing speech under low-resource and noisy conditions. Together, these models provide a comprehensive framework to evaluate *VoiceSecure* effectiveness in both protecting speaker identity and speech content across various use cases.

8.2 State-of-the-art Speech DataSets

Table 3 lists the datasets used to train and evaluate the performance of *VoiceSecure*. Initially, we train our adaptive speech modification controller using the **LibriSpeech** [46] 40-speaker dataset. For testing both speaker verification and speech recognition, we then use the **LibriSpeech** [46] 919-speaker dataset. Additionally, we use the **VoxCeleb1** [43] 1211-speaker dataset and the **VCTK** [70] dataset to evaluate speaker verification performance. For evaluating speech recognition accuracy, we utilize the **CommonVoice** [7] and **VCTK** [70] datasets.

8.3 Baseline Methods for Comparison

To benchmark the performance of *VoiceSecure*, we selected two widely used signal processing-based anonymization and masking techniques: McAdams [47] and VoiceMask [51]. These methods were chosen due to their low computational requirements and applicability on resource-constrained devices, which makes them suitable for real-world deployment. MicPro [76] is a recent anonymization method designed specifically to obscure speaker identity. However, it is not designed to protect against speech recognition. In contrast, *VoiceSecure* is designed to preserve both speaker identity and speech content from automated systems. We excluded generative model-based approaches such as Vcloak [19] and Smack [80], which, despite their impressive privacy-preserving capabilities, require significantly more computational resources, making them impractical for lightweight, real-time deployments on edge devices. For McAdams and VoiceMask comparison, we follow the configuration guidelines provided in the GitHub implementations [73, 82]. Specifically, we set the McAdams coefficient to 0.8 and the VoiceMask warping factor to 0.1 based on the parameters used in typical evaluations of these baseline methods. This setup allows us to assess the effectiveness of *VoiceSecure* in comparison to these established techniques.

8.4 Evaluation Metrics

To evaluate the performance of *VoiceSecure*, we employ the following metrics commonly used in the field:
(1) MisMatch Rate (MMR): measures the rate at which the automatic speaker verification (ASV) model incorrectly verifies speaker identity, reflecting the model's robustness in preserving speaker privacy.
(2) Equal Error Rate (EER): represents the point at which the false acceptance rate (FAR) equals the false rejection rate (FRR).
(3) Word Error Rate (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.
(4) Short-Time Objective Intelligibility (STOI): assesses the intelligibility of speech signals by evaluating the similarity between the original and modified speech signals, helping to ensure that modifications maintain natural communication.
(5) Latency: measures the time delay introduced by *VoiceSecure*, critical for real-time applications and user experience.

9 Evaluation

This section summarizes *VoiceSecure*'s overall performance, evaluated across six advanced speaker verification and speech recognition systems using four widely-used speech datasets. We compare *VoiceSecure* 's results to two popular signal-processing-based anonymization and masking methods and also assess performance against environmental noise to demonstrate the robustness of identification and recognition models in noisy conditions. Our results indicate that *VoiceSecure* achieves a 52% Word Error Rate and a 33% Speaker MisMatch Rate while preserving 72% speech intelligibility.

9.1 Comparison with Baseline Methods

This section compares *VoiceSecure* with two baseline methods across various speaker verification and speech recognition models. As shown in Figure 9, *VoiceSecure* surpasses McAdams in safeguarding speaker identity (achieving a higher mismatch rate and equal error rate) and in concealing speech content (higher word error rate). While in terms of speech protection, *VoiceSecure* performs similarly to VoiceMask, it achieves 12% greater intelligibility, as displayed in Figure 9. The overall results demonstrate

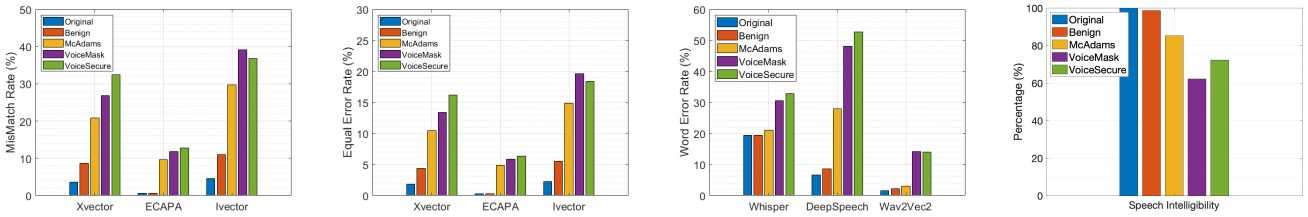

Figure 9: Performance comparison of *VoiceSecure* against baseline methods across four metrics: speaker mismatch rate, equal error rate, word error rate, and speech intelligibility evaluated on three speaker verification and speech recognition models.

Method	LibriSpeech	VoxCeleb	VCTK
McAdams	20%	33%	97%
VoiceMask	26%	37%	64%
VoiceSecure	33%	40%	92%

Table 4: Performance comparison of methods across multiple datasets in terms of mismatch rate.

Method	LibriSpeech	VCTK	CommonVoice
McAdams	27%	68%	83%
VoiceMask	48%	100%	92%
VoiceSecure	52%	106%	92%

Table 5: Performance comparison of methods across multiple datasets in terms of Word Error Rate (WER). Note that WER above 100% is not an error, and according to the official JiWER documentation, values above 100% can occur in cases with significant insertions or discrepancies [2].

VoiceSecure's balance between privacy protection and speech quality, making it a robust solution for real-world applications where intelligibility and security are both crucial.

9.2 Comparison across Diverse Datasets

To assess the robustness of *VoiceSecure* in different settings, we tested our system on a variety of commonly used speech datasets. Table 4 demonstrates that *VoiceSecure* outperforms existing methods, consistently achieving higher or comparable mismatch rates, effectively protecting speaker identity across all tested datasets. Similarly, Table 5 shows that *VoiceSecure* surpasses other methods in obscuring speech content across three distinct datasets by consistently achieving higher word error rates. These results highlight that *VoiceSecure* is highly generalizable to diverse scenarios and effective for unknown speakers without the need for additional training.

9.3 Comparison 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 10, *VoiceSecure* consistently achieves a higher mismatch rate and word error rate, indicating its superior efficacy in concealing sensitive information from automated systems. This highlights the point that these state-of-the-art speech models are highly robust against environmental noise, and simple addition of noise is not enough to protect speech.

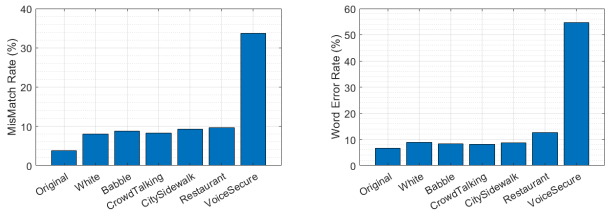

Figure 10: Performance comparison of *VoiceSecure* under various benign noise conditions, shown in terms of speaker mismatch rate (left) and word error rate (right).

9.4 Subjective Quality

We conducted a user study to evaluate the real-world performance of *VoiceSecure* in preserving speech naturalness and intelligibility during phone conversations. The primary goal of the study was to determine whether the modified speech remains perceptually identical to the original speaker's voice, ensuring a natural communication experience.

The study was approved by our Institutional Review Board (IRB) and carried out in accordance with established ethical guidelines. We recruited 22 adult participants (ages 18–35) via departmental email. All of whom were regular users of voice calls and affiliated with a university setting. Participation was voluntary, driven by general interest in speech technology, and no monetary compensation was provided. All participants provided informed consent prior to the study. No personally identifiable information (PII) was collected or stored, and all responses were anonymized before analysis to ensure confidentiality.

During the study, participants were presented with 15 pairs of short audio clips, each consisting of two different utterances from the same speaker. One utterance was unmodified (clean), while the other was modified using either *VoiceSecure* or a baseline method (McAdams or VoiceMask). For each pair, participants rated two aspects using a 10-point Likert scale: Speaker Confidence, indicating how confident they were that both clips were from the same speaker, and Clarity, reflecting the perceived intelligibility and naturalness of the modified audio. Each session lasted approximately 20 minutes.

Figure 11 summarizes the subjective ratings collected during our user study. As shown in Figure 11 (left), *VoiceSecure* achieves higher average speaker confidence scores compared

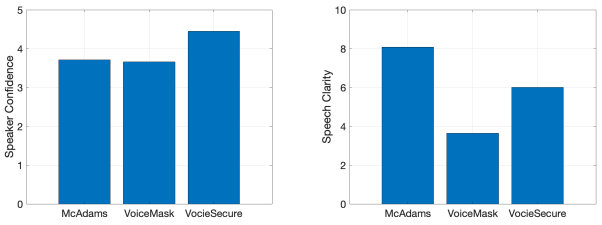

Figure 11: Subjective comparison of *VoiceSecure* against McAdams and VoiceMask, evaluated in terms of confidence in speaker similarity (left) and perceived speech clarity (right).

to both McAdams and VoiceMask. Figure 11 (right) shows that *VoiceSecure* outperforms VoiceMask in speech clarity but slightly trails McAdams. To assess the statistical significance of these trends, we performed paired two-tailed t-tests on the speaker confidence and speech clarity scores between *VoiceSecure* and the two baseline methods. Each participant rated 15 examples, and we aggregated these responses into a single mean score per method per participant, resulting in 22 independent samples for comparison. This approach satisfies the independence assumption required for the paired t-test. For speech clarity, *VoiceSecure* significantly outperformed VoiceMask ($p<0.001$) and also differed significantly from McAdams ($p<0.001$). For speaker confidence, although *VoiceSecure* received higher mean scores than both McAdams and VoiceMask, the differences were not statistically significant ($p=0.076$ and $p=0.014$, respectively), indicating that perceived speaker identity remained comparable across methods. This result is consistent with our objective evaluation, where McAdams achieves higher intelligibility at the expense of privacy, while *VoiceSecure* aims for a more balanced approach. In contrast, VoiceMask offers lower intelligibility and weaker privacy protection relative to *VoiceSecure*. In general, these findings highlight the ability of *VoiceSecure* to maintain speech clarity while preserving speech privacy, making it a good candidate for real-world application.

9.5 Performance across Different Speakers

In this section, we assess the performance of *VoiceSecure* across a diverse range of speakers, which is essential for ensuring the system's robustness. Using the LibriSpeech dataset with 919 speakers, we perform two evaluations. First, for each speaker, we calculate the mean cosine similarity between speaker embeddings of the original and modified speech. Figure 12 (left) presents a cumulative distribution function (CDF) of cosine similarity, showing that *VoiceSecure* consistently achieves lower similarity, making it more challenging for speaker identification models to recognize the speaker. Second, we compute the mean Word Error Rate (WER) for each speaker. Figure 12 (right) illustrates that *VoiceSecure* effectively hinders automated speech recognition across speakers, consistently achieving high WER and thereby supporting robust privacy protection across diverse voices.

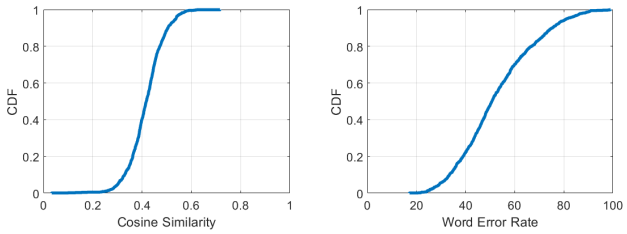

Figure 12: The CDF plot illustrating cosine similarity between speaker embeddings (left) and word error rate (right) across a diverse range of speakers.

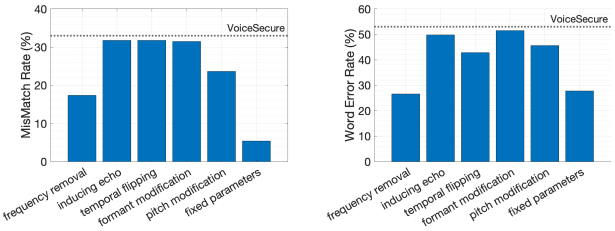

Figure 13: Ablation study showing the impact of individual transformations and fixed-parameter approach in terms of both speaker mismatch rate (left) and word error rate (right).

9.6 Ablation Study

VoiceSecure applies a combination of speech transformations to protect user speech from automated monitoring. To understand the individual contribution of each transformation, we conduct an ablation study using the LibriSpeech dataset. In this study, we systematically disable one transformation at a time while keeping the others active.

The transformations examined include: inaudible frequency removal, pitch modification, formant shifting, temporal flipping, and echo addition. We evaluate the impact of each component by measuring the degradation in *VoiceSecure*'s effectiveness in terms of speaker identification and speech recognition. Speaker Mismatch Rate (MMR) is computed using X-vector, and Word Error Rate (WER) is measured using DeepSpeech.

Figure 13 presents the results of this analysis. The dotted line represents the baseline performance of *VoiceSecure* with all transformations enabled. The figure highlights the unique contribution of each transformation and demonstrates how their combined effect enables robust speech protection. Inaudible frequency removal is the most impactful component, contributing the largest gains in both WER and MMR, because it removes frequency cues that are informative to both ASR and speaker embedding models. Pitch modification plays an important role in speaker anonymization, as it disrupts vocal traits critical for speaker identification. Temporal flipping affects ASR more than speaker identification, suggesting its role is primarily in content disruption. Echo addition and formant have minimal effect on privacy but improve perceptual smoothness, helping the modified speech remain natural to human listeners.

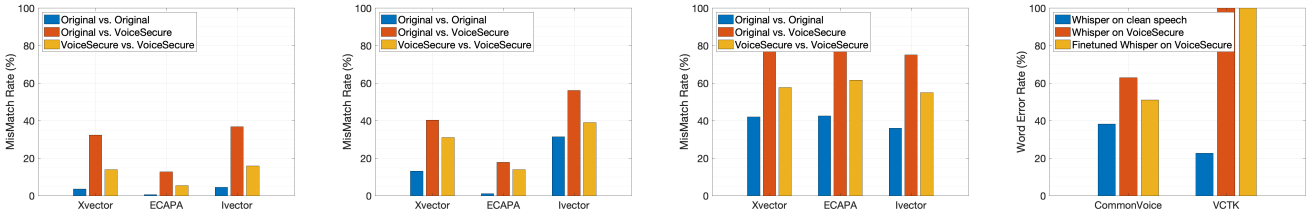

Figure 14: Adversarial evaluation results. (a–c) Speaker Mismatch Rate (MMR) across LibriSpeech, VoxCeleb, and VCTK using X-vector, ECAPA, and i-vector under different enrollment–test configurations. (d) Word Error Rate (WER) of Whisper-base before and after fine-tuning on *VoiceSecure* modified speech, evaluated on CommonVoice and VCTK.

In addition, we compare our reinforcement learning-based adaptive method with a fixed-parameter approach to assess its effectiveness in enhancing speech privacy. For this comparison, we modify samples from the LibriSpeech dataset using both methods and evaluate the resulting audio using the same MMR and WER metrics. As shown in Figure 13, the adaptive tuning approach consistently outperforms the fixed-parameter approach, offering improved protection against both speaker identification and ASR systems.

9.7 Latency

This section evaluates the latency performance of *VoiceSecure*, which is crucial for real-time operation. In this test, we configured *VoiceSecure* to listen to an incoming speech and apply modifications in real-time on a frame-by-frame basis. We record data for 30 minutes and compute the time taken to process each frame. Figure 15 presents the cumulative distribution function (CDF) of 28,125 frames, showing that *VoiceSecure* achieves a median latency of 25 milliseconds. The red line represents the maximum acceptable latency threshold for audio communication systems, confirming that *VoiceSecure* operates within the required limits for seamless, real-time interaction [58, 66].

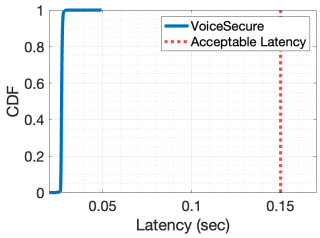

Figure 15: The CDF plot of latency showcases the distribution of processing delays across frames, with a marked line indicating the acceptable latency threshold for real-time audio applications.

10 Adversarial Evaluation

To assess the robustness of *VoiceSecure* against adaptive adversaries, we conducted a targeted adversarial evaluation simulating realistic threat scenarios. Specifically, we modeled a knowledgeable attacker who has access to our transformation pipeline and aims to circumvent speech protection by adapting automated systems for speaker identification and speech recognition.

Speaker Identification Attack: In this evaluation, we simulate a knowledgeable adversary attempting to defeat speaker anonymization by enrolling and testing speaker samples that have been modified by *VoiceSecure*. The goal is to assess whether speaker embeddings remain discriminative despite having access to the modified samples.

We use three standard speech datasets: LibriSpeech, CommonVoice, and VCTK. We process the audio from each dataset through the *VoiceSecure* pipeline to generate protected versions. Speaker identification performance is evaluated in three configurations: using clean speech for both enrollment and testing, using clean speech for enrollment and *VoiceSecure*-modified speech for testing, and using *VoiceSecure*-modified speech for both enrollment and testing. These scenarios allow us to evaluate both worst-case and adaptive attacker behavior. We employ X-vector, ECAPA-TDNN, and i-vector, three widely used speaker identification models.

Figure 14(a), (b), and (c) show the speaker mismatch results for LibriSpeech, VoxCeleb, and VCTK datasets, respectively. As expected, clean-clean configurations yield the lowest mismatch rates, reflecting strong speaker identity preservation. However, when clean speech is used for enrollment and *VoiceSecure*-modified speech for testing, the mismatch rate increases sharply across all models and datasets, indicating that speaker embeddings from clean and transformed audio are no longer well aligned. In the adaptive setting, where both enrollment and testing are done on protected speech, the mismatch rate drops slightly but still remains higher than the clean baseline. These results demonstrate that while a knowledgeable attacker can reduce the mismatch rate slightly by adapting to the transformation, *VoiceSecure* still provides strong protection against speaker identification attacks.

ASR Adaptation Attack: In this evaluation, we simulate an adversary attempting to defeat speech content obfuscation by adapting an automatic speech recognition (ASR) model to *VoiceSecure*. This scenario models a determined attacker who has access to a large amount of protected speech and fine-tunes their ASR system accordingly.

We begin by applying *VoiceSecure* to speech samples from the LibriSpeech dataset, 27000 speech samples with durations ranging from 3 to 30 seconds. These transformed

audio samples are then used to fine-tune the open-source Whisper-base ASR model. The goal is to reduce the WER by learning to decode the protected speech, effectively reversing the impact of *VoiceSecure*. After fine-tuning, we evaluate the model on the CommonVoice and VCTK speech datasets protected using *VoiceSecure*. We compare the WER of the fine-tuned model against two baselines: (1) the Whisper-base model evaluated on clean speech, and (2) the Whisper-base model evaluated on protected speech.

Figure 14(d) presents the WER across these three conditions. While fine-tuning on protected speech leads to a modest reduction in WER compared to the unadapted baseline, the performance remains substantially worse than the clean-speech case. This indicates that even with access to thousands of protected examples, the adapted ASR model cannot fully recover intelligible transcriptions. These results demonstrate that *VoiceSecure* effectively impairs speech recognition even under adversarial adaptation, preserving speech content privacy in the face of realistic ASR retraining attacks. Moreover, collecting a sufficiently large corpus of protected speech with ground-truth transcriptions is non-trivial. An attacker would either need to manually transcribe protected audio using human annotators or employ our system to generate protected samples along with corresponding labels, both of which introduce significant logistical and operational barriers.

Discussion: While *VoiceSecure* demonstrates strong resilience to adversarial attacks, its robustness can be further enhanced by applying the transformation multiple times. However, doing so naively may degrade intelligibility, so a more effective approach is to train an adaptive controller that dynamically determines the number of passes and parameter configurations at runtime. This would make it significantly harder for attackers to curate consistent samples for fine-tuning or enrollment. We leave this adaptive reprocessing strategy as promising future work to strengthen defense against increasingly sophisticated threats.

11 Related Work

Attacks on Verification Models

Existing voice anonymization and masking methods are primarily based on signal processing (SP), voice conversion (VC), and voice synthesis (VS). Signal processing methods [47, 69] modify speaker-related features such as formants, pitch, and tempo, but often degrade audio quality by overlooking intelligibility and naturalness. Voice conversion [49, 50, 64, 79] and voice synthesis [24, 29, 32, 40] methods convert speech into a different sounding voice, achieving anonymity but sacrificing the user's natural vocal identity, making them less suitable for use cases that require both privacy and naturalness.

Adversarial methods have also been proposed to fool ASV systems by adding imperceptible perturbations [15, 22, 36, 37, 84], but these are typically generated via slow, iterative processes and are impractical for real-time deployment. More recent systems like FAPG [77] and VCloak [19] offer improved intelligibility and naturalness, yet primarily focus on anonymizing speaker identity and fail to protect speech content from ASR systems. In contrast, *VoiceSecure* is designed to protect both speaker identity and speech content from automated monitoring, all while preserving the naturalness of speech for human listeners.

Adversarial Attacks on Recognition Models

Adversarial attacks on automatic speech recognition involve creating examples that are transcribed differently by machines while appearing natural to humans [74]. Early efforts designed obfuscated commands unintelligible to humans but effective against GMM-based ASR systems [12], while CommanderSong [81] embedded adversarial perturbations into music to fool DNN-based models. Gradient-based methods further optimized adversarial audio using CTC loss [13]. Researchers have also looked at ways to improve the practicality and stealthiness of adversarial audio examples. Metamorph [16] studied mechanisms to enhance the survival of the adversarial audio examples in over-the-air transmission, while Schonherr et al. [57] 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 [52]. While effective, most of these attacks assume white-box access to the target model. To address this, black-box approaches have been explored using signal pre-processing [5] or local model approximations, as in Devil's Whisper [18]. However, none of these methods generate real-time perturbations capable of simultaneously deceiving both ASR and ASV systems.

Recently, SMACK [80] proposed semantic adversarial perturbations by altering pitch, tone, and phoneme duration to fool both ASR and ASV systems. However, it is 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.

12 Conclusion

VoiceSecure presents an innovative microphone module that protects user privacy from automated speaker identification and speech recognition while preserving natural audio quality. It combines signal-processing-based modifications with a reinforcement learning model that adaptively tunes parameters in real time. Implemented on an off-the-shelf microcontroller, *VoiceSecure* integrates easily with existing devices. Evaluations show that it outperforms existing methods in safeguarding voice data from automated monitoring while maintaining human intelligibility.

13 Acknowledgments

This work was partially supported by NSF CAREER Award 2238433. We also thank our industry partners for their continued support of the iCoSMoS Laboratory at UMD.

14 Ethical Consideration

This study was conducted with strict adherence to ethical guidelines to ensure participant privacy, transparency, and voluntary engagement. We recruited 22 adult participants, aged between 18 and 35, all of whom were regular users of voice calls and affiliated with a university setting. No personally identifiable information (PII) was collected or stored at any stage of the study. All responses were anonymized before analysis to maintain confidentiality. Our study is approved by the IRB. Participants were presented with pairs of short audio samples from the same speaker: one unmodified and the other modified using either *VoiceSecure* or a baseline method, and were asked to evaluate speech in terms of clarity and similarity to the original speaker. Each session lasted approximately 20 minutes, and no monetary or material compensation was provided. All participants were informed about the nature of the study and voluntarily consented to participate. When implementing and testing our system, we refrained from targeting any specific commercial or open-source systems running online. All evaluations were conducted offline using controlled environments on desktops to ensure that our research does not interfere with or disrupt any external systems.

Our work aims to empower users by protecting their speech from automated monitoring, thereby mitigating the risks of mass surveillance and safeguarding privacy. While we acknowledge the dual-use nature of such technology, enhancing privacy for legitimate users but potentially being misused by malicious actors to evade detection, this is not a significant concern. Our system is designed to preserve intelligibility for human listeners, ensuring that authorities with proper legal authorization can still monitor and identify malicious individuals through traditional means. This approach strikes a balance between privacy protection and accountability.

15 Open Science

In compliance with open science principles, we publicly release our source code, pre-trained models, and dataset to facilitate reproducibility and foster further research in the field.[1].

References

[1] About end-to-end encryption. https://faq.whatsapp.com/820124435853543. Last accessed 07 April 2023.

[2] jiwer. https://github.com/jitsi/jiwer. Last accessed 22 January 2025.

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

[4] Voip adoption statistics for 2023 & beyond. https://wisdomplexus.com/blogs/voip-adoption-statistics-2023-beyond/, Oct 2024. Last accessed 14 November 2024.

[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] Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, et al. Deep speech 2: End-to-end speech recognition in english and mandarin. In *International conference on machine learning*, pages 173–182. PMLR, 2016.

[7] Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber. Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*, 2019.

[8] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020.

[9] Yang Bai, Nakul Garg, and Nirupam Roy. Spidr: Ultra-low-power acoustic spatial sensing for micro-robot navigation. In *Proceedings of the 20th Annual International Conference on Mobile Systems, Applications and Services*, pages 99–113, 2022.

[10] Yang Bai, Irtaza Shahid, Harshvardhan Takawale, and Nirupam Roy. Scribe: Simultaneous voice and handwriting interface. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, pages 1–31, 2024.

[11] Saikat Basu, Jaybrata Chakraborty, Arnab Bag, and Md Aftabuddin. A review on emotion recognition using speech. In *2017 International conference on inventive communication and computational technologies (ICICCT)*, pages 109–114. IEEE, 2017.

[1] https://doi.org/10.5281/zenodo.15603263

[12] 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.

[13] 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.

[14] Justin Chan, Antonio Glenn, Malek Itani, Lisa R Mancl, Emily Gallagher, Randall Bly, Shwetak Patel, and Shyamnath Gollakota. Wireless earbuds for low-cost hearing screening. In *Proceedings of the 21st Annual International Conference on Mobile Systems, Applications and Services*, pages 84–95, 2023.

[15] 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.

[16] 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.

[17] Tuochao Chen, Malek Itani, Sefik Emre Eskimez, Takuya Yoshioka, and Shyamnath Gollakota. Hearable devices with sound bubbles. *Nature Electronics*, pages 1–12, 2024.

[18] 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.

[19] 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.

[20] 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.

[21] ONNX Runtime developers. Onnx runtime. https://onnxruntime.ai/, 2021. Version: x.y.z.

[22] 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.

[23] Matt Egan and Sean Lyngaas. Nearly all att cell customers' call and text records exposed in a massive breach. https://www.cnn.com/2024/07/12/business/att-customers-massive-breach/index.html, July 2024. Last accessed 14 November 2024.

[24] 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.

[25] Nakul Garg, Yang Bai, and Nirupam Roy. Owlet: Enabling spatial information in ubiquitous acoustic devices. In *The 19th Annual International Conference on Mobile Systems, Applications, and Services (MobiSys '21), June 24–July 2, 2021, Virtual, WI, USA*. ACM.

[26] Nakul Garg, Harshvardhan Takawale, Yang Bai, Irtaza Shahid, and Nirupam Roy. Structure assisted spectrum sensing for low-power acoustic event detection. In *Proceedings of Cyber-Physical Systems and Internet of Things Week 2023*, pages 278–284. 2023.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] Smith K Khare, Victoria Blanes-Vidal, Esmaeil S Nadimi, and U Rajendra Acharya. Emotion recognition and artificial intelligence: A systematic review (2014–2023) and research recommendations. *Information fusion*, 102:102019, 2024.

[34] Ege Cem Kirci, Maria Apostolaki, Roland Meier, Ankit Singla, and Laurent Vanbever. Mass surveillance of voip calls in the data plane. In *Proceedings of the Symposium on SDN Research*, pages 33–49, 2022.

[35] 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.

[36] 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.

[37] 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.

[38] Yael Mathov, Tal Ben Senior, Asaf Shabtai, and Yuval Elovici. Stop bugging me! evading modern-day wiretapping using adversarial perturbations. *Computers & Security*, 121:102841, 2022.

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

[40] 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.

[41] Ayushi Mishra, Yang Bai, Priyadarshan Narayanasamy, Nakul Garg, and Nirupam Roy. Spatial audio processing with large language model on wearable devices. *arXiv preprint arXiv:2504.08907*, 2025.

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

[43] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. Voxceleb: a large-scale speaker identification dataset. *arXiv preprint arXiv:1706.08612*, 2017.

[44] Ellen Nakashima and Josh Dawsey. Chinese hackers said to have collected audio of american calls. https://www.washingtonpost.com/national-security/2024/10/27/chinese-hackers-cellphones-trump/, Oct 2024. Last accessed 14 November 2024.

[45] 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.

[46] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE, 2015.

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

[48] Raspberry Pi. Raspberry pi 3 model b+. https://www.raspberrypi.com/products/raspberry-pi-3-model-b-plus/. Last accessed 14 November 2024.

[49] 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.

[50] 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.

[51] Jianwei Qian, Haohua Du, Jiahui Hou, Linlin Chen, Taeho Jung, Xiang-Yang Li, Yu Wang, and Yanbo Deng. Voicemask: Anonymize and sanitize voice input on mobile devices. *arXiv preprint arXiv:1711.11460*, 2017.

[52] 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.

[53] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR, 2023.

[54] Nirupam Roy, Haitham Hassanieh, and Romit Roy Choudhury. Backdoor: Making microphones hear inaudible sounds. In *Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services*, pages 2–14, 2017.

[55] Nirupam Roy and Romit Roy Choudhury. Listening through a vibration motor. In *Proceedings of the 14th Annual International Conference on Mobile Systems, Applications, and Services*, pages 57–69, 2016.

[56] 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.

[57] 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.

[58] Maulik Shah. What is voip latency and why does it matter and how to fix it. https://tragofone.com/what-is-latency-business-voip-fix-low-latency/#:~:text=Acceptable%20limits%20of%20latency%20in%20VoIP,-VoIP%20systems%20use&text=Industry%20experts%20suggest%20that%20latency,are%20within%20acceptable%20latency%20limits., Jan 2024. Last accessed 14 November 2024.

[59] Irtaza Shahid, Khaldoon Al-Naimi, Ting Dang, Yang Liu, Fahim Kawsar, and Alessandro Montanari. Towards enabling dpoae estimation on single-speaker earbuds. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 246–250. IEEE, 2024.

[60] Irtaza Shahid, Yang Bai, Nakul Garg, and Nirupam Roy. Voicefind: Noise-resilient speech recovery in commodity headphones. In *Proceedings of the 1st ACM International Workshop on Intelligent Acoustic Systems and Applications*, pages 13–18, 2022.

[61] Irtaza Shahid and Nirupam Roy. " is this my president speaking?" tamper-proofing speech in live recordings. In *Proceedings of the 21st Annual International Conference on Mobile Systems, Applications and Services*, pages 219–232, 2023.

[62] Irtaza Shahid and Nirupam Roy. Poster: Preventing fake news through live speech signature. In *Proceedings of the 21st Annual International Conference on Mobile Systems, Applications and Services*, pages 567–568, 2023.

[63] 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.

[64] 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.

[65] Matt Storey. Advancing voice surveillance: A call to action for financial institutions. https://www.steel-eye.com/news/advancing-voice-surveillance-a-call-to-action-for-financial-institutions, September 2024. Last accessed 14 November 2024.

[66] Electronic Office Systems. What is considered an acceptable latency for voip calls? https://www.electronicofficesystems.com/2023/10/09/what-is-considered-an-acceptable-latency-for-voip-calls/, Oct 2023. Last accessed 14 November 2024.

[67] Harshvardhan Takawale and Nirupam Roy. Learning speaker-listener mutual head orientation by leveraging hrtf and voice directivity on headphones. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1171–1175. IEEE, 2024.

[68] Ben Underwood and Hossein Saiedian. Mass surveillance: A study of past practices and technologies to predict future directions. *Security and Privacy*, 4(2):e142, 2021.

[69] 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.

[70] Christophe Veaux, Junichi Yamagishi, Kirsten MacDonald, et al. Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit. *University of Edinburgh. The Centre for Speech Technology Research (CSTR)*, 6:15, 2017.

[71] Kannan Venkataramanan and Haresh Rengaraj Rajamohan. Emotion recognition from speech. *arXiv preprint arXiv:1912.10458*, 2019.

[72] Pulkit Verma and Pradip K Das. i-vectors in speech processing applications: a survey. *International Journal of Speech Technology*, 18:529–546, 2015.

[73] VoicePrivacy2020. Voice privacy challenge 2020. https://github.com/Voice-Privacy-Challenge/Voice-Privacy-Challenge-2020/, 2020. Last accessed 14 November 2024.

[74] 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.

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

[76] 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.

[77] 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.

[78] Alan Xu and Romit Roy Choudhury. Learning to separate voices by spatial regions. In *International Conference on Machine Learning*, pages 24539–24549. PMLR, 2022.

[79] 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.

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

[81] 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.

[82] yuunin. time-invariant-anonymization. https://github.com/yuunin/time-invariant-anonymization, 2020. Last accessed 14 November 2024.

[83] Tao Zhang and Zhenhua Tan. Survey of deep emotion recognition in dynamic data using facial, speech and textual cues. *Multimedia Tools and Applications*, pages 1–40, 2024.

[84] 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.

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

USENIX Security '25 Artifact Appendix: Preventing Automated Monitoring on Voice Data

Irtaza Shahid, Nirupam Roy

University of Maryland, College Park
{irtaza, niruroy}@umd.edu

A Artifact Appendix

A.1 Abstract

We present the artifact for the paper, titled "For Human Ears Only: Preventing Automated Monitoring on Voice Data". It includes the datasets, code, and models necessary for training and reproducing the major results presented in the paper. As voice communication becomes an essential part of modern life, the exposure of sensitive information through audio calls presents significant privacy risks. Malicious actors can gain access to this data by compromising user devices, exploiting communication channels, or attacking data servers, making it vulnerable to automated monitoring systems that can identify speakers and extract speech content. To address these privacy concerns, we introduce VoiceSecure, the first microphone module designed to prevent automated monitoring of speech while preserving its natural sound for humans. By leveraging the principles of human auditory perception, VoiceSecure employs a set of speech modifications that are adaptively tuned in real-time to obscure speaker identity and speech content, without compromising the audio quality for humans.

This artifact supports the goals of open science and reproducibility by providing all necessary components to replicate the core results of the paper. It includes source code, pre-trained models, processed datasets, evaluation metrics (WER, MMR, STOI), and scripts to generate all key figures and tables. The artifact allows evaluators to apply VoiceSecure transformations, measure their impact on speech recognition and speaker identification systems, and verify that the transformations preserve intelligibility—thereby validating the paper's major claims.

A.2 Description & Requirements

This section lists all the information necessary to recreate the same experimental setup we have used to develop our system.

A.2.1 Security, privacy, and ethical concerns

Our artifact presents no known security, privacy, or ethical risks to evaluators or their systems. It does not require administrative access, disabling of any security mechanisms, or the use of potentially harmful scripts or binaries.

A.2.2 How to access

The complete artifact is permanently archived and accessible via Zenodo at the following DOI: https://doi.org/10.5281/zenodo.15603263 This archive includes:

- VoiceSecure source code
- Pre-trained models
- Scripts for evaluation, metrics computation, and figure generation
- Instructions for training on custom datasets
- Precomputed results used in the paper
- A detailed Readme to guide setup and execution

A.2.3 Hardware dependencies

None. The artifact is designed to run on standard consumer-grade hardware. No GPU is required. All evaluations can be performed using a CPU machine with at least 8 GB of RAM.

A.2.4 Software dependencies

The artifact requires:

- **Python 3.8**, with packages listed in the provided requirements.txt
- **MATLAB R2021a or later**, with the following toolboxes: Audio, Signal Processing, and 5G Toolboxes.

A.2.5 Benchmarks

The following datasets and models were used in the experiments presented in the paper:

- **Speech Datasets:** Open-source corpora including LibriSpeech, VoxCeleb, CommonVoice, and VCTK.

- **User Study Responses:** Collected from human listeners to evaluate perceptual intelligibility using STOI.
- **Speech Recognition:** Whisper, DeepSpeech, and Wav2Vec2.
- **Speaker Verification:** x-vector, ECAPA-TDNN, and i-vector models.

All benchmark data (except raw VoxCeleb audio due to licensing) and model outputs—including speaker embeddings, mismatch rates (MMR), WER, and STOI scores—are included in the artifact package. Pre-trained speaker verification models are bundled with the artifact. Open-source speech recognition models are supported and installed via dependencies specified in **requirements.txt**.

A.3 Set-up

This section provides detailed steps to install and configure the environment necessary for evaluating the VoiceSecure artifact. Following the instructions below, evaluators will be able to run a basic test to verify the correct installation and functionality of all required components. Please note that the full setup process may take approximately 2 hours.

A.3.1 Installation

1. **Download the artifact files** from Zenodo: https://doi.org/10.5281/zenodo.15603263 The download includes two zip files: **Data2.zip (23 GB)**, and **VoiceSecure_Artifacts_Scripts.zip (1.5 GB)**.
2. **Unzip both files into a common directory**. Your directory structure will look like:
   - Data2/
   - ScriptForApplyingVoiceSecure/
   - ScriptForTrainingModel/
   - ScriptForComputingMetrics/
   - ScriptsForCompiledResults/
   - ScriptForDataSetCreation/
   - Trained_Model/
   - Testing_Installation/
   - requirements.txt
   - README.md
3. **Install Miniconda (if not already installed):**
   bash Miniconda3-latest-Linux-x86_64.sh
   source /miniconda3/bin/activate
4. **Create and activate a conda environment:**
   conda create –name py38 python=3.8.18
   conda activate py38
5. **Install Python dependencies:**
   pip install -r requirements.txt
6. **MATLAB Dependencies:** MATLAB (R2021a or later) with the Audio, Signal Processing, and 5G toolboxes.
7. **To run DeepSpeech-based evaluations:**
   - Clone from the GitHub: https://github.com/SeanNaren/deepspeech.pytorch
   - Navigate to the repo
   - pip install -r requirements.txt
   - Ensure this repo is added to your environment path.

A.3.2 Basic Test

To verify that the installation has been completed successfully, we provide a testing script located in the `Testing_Installation/` directory. This script performs checks to ensure that all major components of the system are functional, including the VoiceSecure model, the speaker embedding models (x-vector, ECAPA), and the ASR models used for computing word error rates (Whisper, DeepSpeech, Wav2Vec2). To run the test, execute the script `TestPythonInstallation.py`. If the setup is correct, the script will print seven confirmation messages, each beginning with the word `"Functional"`, corresponding to the various components being tested. At the end, it will display the message `"Installation complete"`, indicating that all necessary modules are working as expected.

A.4 Evaluation workflow

VoiceSecure is a speech transformation method designed to obfuscate speaker identity and speech content while preserving intelligibility for human listeners. To evaluate its effectiveness, we perform three key types of evaluation:

1. **Speaker Verification:** We use three state-of-the-art speaker verification models (X-Vector, ECAPA-TDNN, and i-Vector) to extract speaker embeddings from both the original and VoiceSecure-modified speech. We then compute the speaker mismatch rate (MMR), which quantifies how often the modified speech is misidentified as originating from a different speaker. A **higher mismatch rate** indicates stronger anonymization, which is the desired outcome.
2. **Word Error Rate (WER):** To assess the impact of VoiceSecure on speech recognition accuracy, we evaluate the modified speech using three automatic speech recognition (ASR) models: Whisper, DeepSpeech, and Wav2Vec2. We compute the WER for both original and modified speech. A **higher WER** reflects greater disruption to ASR systems, which is the intended effect.

3. **Speech Intelligibility:** We measure intelligibility using the Short-Time Objective Intelligibility (STOI) metric, which correlates strongly with human perceptual scores. In this context, a **higher STOI score** is desirable, as it indicates that the modified speech remains understandable to human listeners despite the applied transformations.

Our artifact includes all necessary scripts to apply VoiceSecure modifications, compute speaker embeddings, mismatch rate, WER, and STOI scores. It also provides original and transformed speech samples, along with pre-computed embeddings, mismatch rates, WERs, and STOI scores. In addition, we include MATLAB scripts that generate the key figures and tables from the paper using the pre-evaluated data.

A.4.1 Major Claims

(C1): VoiceSecure achieves a 52% Word Error Rate, 33% Speaker Mismatch Rate, and 72% intelligibility, demonstrating its ability to protect privacy while preserving human understanding. (Section 9)

(C2): Compared to existing baselines, VoiceSecure offers a better trade-off between privacy and intelligibility, outperforming McAdams in both speaker anonymization and ASR obfuscation, and exceeding VoiceMask by 12% in intelligibility. (Section 9.1, Figure 9)

(C3): Subjective evaluations confirm that VoiceSecure maintains perceived speech clarity while enhancing privacy, making it suitable for real-world deployment. (Section 9.4, Figure 11).

A.4.2 Experiments

(E1): [60 human-minutes + 20 compute-hours] This experiment evaluates VoiceSecure's effectiveness in anonymizing speaker identity and obfuscating speech content, supporting major claim C1.
Preparation: Ensure all dependencies are installed. Confirm the availability of the pre-trained VoiceSecure model and evaluation datasets.
Execution: Run the scripts in `ScriptForApplyingVoiceSecure/` to apply the transformation and generate modified speech samples. Then execute `ComputeSpeakerEmbeddings.py` and `ComputeWER.py` to compute and store speaker embeddings and word error rates, respectively. For this experiment, we recommend using the LibriSpeech dataset along with the X-Vector and DeepSpeech for speaker embeddings and word error rates, respectively. As this process is computationally intensive, we also provide pre-modified speech samples as well as pre-computed embeddings and word error rates. Finally, run the scripts in `ScriptForComputingMetrics/` to compute the speaker mismatch rate, mean word error rate, and speech intelligibility scores.
Results: The experiment outputs speaker mismatch rates, word error rates, and intelligibility scores, supporting claim 1.

(E2): [60 human-minutes + 30 compute-hours] This experiment supports major claim C2 by comparing VoiceSecure against baselines across various state-of-the-art speaker verification and speech recognition systems in terms of privacy protection and intelligibility.
Preparation: Ensure `Data2/LibriSpeech_Dev/` contains speaker embeddings (for three models), word error rates (for three models), and intelligibility scores for both original and processed speech samples (noise, McAdams, VoiceMask, and VoiceSecure).
Execution: Use the scripts in 'ScriptForComputingMetrics/' to compute and store MMR, WER, and STOI for all methods. Then run the 'ScriptsForCompiledResults/Plotting_Compiled_Results' MATLAB script to generate comparison figures.
Results: This reproduces Figure 9 in the paper, demonstrating that VoiceSecure achieves a superior trade-off between privacy and intelligibility compared to existing methods.

(E3): [10 human-minutes + 0.1 compute-hour] This experiment supports major claim C3 by analyzing listener feedback on perceived speech intelligibility after transformation.
Preparation: Navigate to `Data2/UserStudy/` and verify that the listener response data is present.
Execution: Run the provided MATLAB script 'ScriptsForCompiledResults/Plotting_UserStudy_Results' to process user study responses and generate aggregated perceptual scores.
Results: The experiment reproduces the user study results (Figure 11) in the paper, validating that VoiceSecure maintains high perceived intelligibility while offering strong privacy protections.

A.5 Notes on Reusability

The scripts in `ScriptForApplyingVoiceSecure/` allow users to apply VoiceSecure-based modifications to any speech dataset, enabling privacy protection across diverse use cases. Additionally, the script in `ScriptForTrainingModel/` enables training the model from scratch on custom datasets, allowing adaptation to different domains or data conditions.

A.6 Version

Based on the LaTeX template for Artifact Evaluation V20231005. Submission, reviewing and badging methodology followed for the evaluation of this artifact can be found at https://secartifacts.github.io/usenixsec2025/.