Pipe Overflow: Smashing Voice Authentication for Fun and Profit

Shimaa Ahmed*, Yash Wani†, Ali Shahin Shamsabadi†,
Mohammad Yaghini‡, Ilia Shumailov§,
Nicolas Papernot‡, Kassem Fawaz*

Abstract

Recent years have seen a surge of popularity of acoustics-enabled personal devices powered by machine learning. Yet, machine learning has proven to be vulnerable to adversarial examples. Large number of modern systems protect themselves against such attacks by targeting the artificiality, i.e., they deploy mechanisms to detect the lack of human involvement in generating the adversarial examples. However, these defenses implicitly assume that humans are incapable of producing meaningful and targeted adversarial examples. In this paper, we show that this base assumption is wrong. In particular, we demonstrate that for tasks like speaker identification, a human is capable of producing analog adversarial examples directly with little cost and supervision: by simply speaking through a tube, an adversary reliably impersonates other speakers in eyes of ML models for speaker identification. Our findings extend to a range of other acoustic-biometric tasks such as liveness, bringing into question their use in security-critical settings in real life, such as phone banking.

1 Introduction

As a primary mechanism for human communication, speech is a natural vehicle for human-computer interaction (HCI). Fueled by advancements in Machine Learning (ML), everyday devices and services accept speech as input; users can seamlessly control their smart devices and communicate with automated customer services. This convenience brought the need to authenticate users when speech is the primary interaction modality. Companies deploy speaker identification systems (ASI) that pack ML-based models to authenticate users based on their voiceprint [38, 52].

Similar to other ML-based solutions, speaker identification systems are vulnerable to an array of attacks. These attacks include speech synthesis [60, 63, 66], voice conversion [37, 50, 73], replay attacks [31], and adversarial examples [24, 15, 32]. The adversary generates and feeds the speaker identification system a speech sample to impersonate a target speaker. While the attack techniques differ, they share a common principle: the attacker manipulates the speech signal in the digital domain and potentially plays it through a speaker. Note that even physical adversarial examples follow the same principle, such as those in the vision or acoustic domains. Generating these examples requires obtaining a signal (such as a speech recording or a visual patch) by solving an optimization problem in the digital domain and later realizing it in the analog domain.

Current defenses leverage this observation and employ mechanisms to detect the digital attack artifacts in the input signal [48, 64, 65]. These defenses target either the (1) physical properties of the speaker e.g. their physical presence [42, 75] or (2) properties of the speech speakers produce e.g. the energy distribution of different harmonics [25, 14]. The resulting unified acoustic pipeline constrains the attacker when generating the attack samples, thus increasing the cost of the attack [35, 48, 65]. Generally speaking, the defense literature makes a basic assumption that the attack source is not human. In this paper, we challenge this assumption by asking this question: Is it possible to attack speaker identification systems using analog manipulation of the speech signal?

*University of Wisconsin-Madison
†Currently at The Alan Turing Institute, work done while at Vector Institute
‡Vector Institute and University of Toronto
§Vector Institute
Answering this question in the affirmative has critical implications on using ML to detect and identify human speakers. An analog transform of the speech signal to evade speaker identification challenges the *identifiability* assumption that underlies various acoustic tasks; human characteristics can no longer be uniquely identified from their speech. An attacker can control the propagation medium to affect the speaker identification task. Towards that end, we present **Mystique**, a live spoof attack, which enables analog transformations of speech signals. **Mystique** allows the attacker to transform their voice for inducing a targeted misclassification at the ASI system, effectively impersonating a target victim.

Realizing **Mystique** requires us to satisfy three conditions. First, the analog transform must occur on live speech. Second, an arbitrary speaker should be able to impersonate another arbitrary victim; *i.e.*, the attacker needs not to be a professional vocalist or have any impersonation experience. Third, the transform should directly impact the ASI model prediction. **Mystique** exploits the acoustic resonance phenomenon to satisfy these three conditions. Acoustic resonance is a physical transform where objects vibrate to specific frequencies. Acoustic resonance allows an object to act as a filter, amplifying some frequency components and dampening others.

**Mystique** uses hand-crafted tubes to apply the adversarial resonance transformation to the speaker’s voice. We chose tubes as our attack’s physical objects for two reasons. First, tubes are ubiquitous and inexpensive; they are available in hardware stores in different dimensions. Second, there is extensive literature on acoustic modeling of musical wind instruments, most of which have cylindrical or conical shapes. Note that the same methodology can be extended to arbitrary shapes using wave simulation and numerical analysis [6, 62].

To realize **Mystique**, we model the tube resonator as a band-pass filter (BPF) transform; the tube dimensions fully define the filter. Next, we develop a black-box optimization procedure over the filter parameters (tube dimensions) to trick the ASI model into recognizing the voice of a chosen target speaker. We apply an evolutionary algorithm (Sec. 4.4) that uses the ASI model score and label to find the optimal tube dimensions for a given target speaker. An adversary can use these parameters to realize a tube that would match their voice to a target speaker.

We perform extensive evaluation of **Mystique** on two state-of-the-art ASI models and two spoofing detection baselines. We validate **Mystique** on standard speaker identification dataset, VoxCeleb, and on live speech by conducting a user study of 12 participants. We build a physical recording setup, and evaluate **Mystique** physically. We confirm that **Mystique**’s adversarial tubes succeed in performing over-the-air impersonation.
attack in the real-world.
This paper makes the following contributions:

- We show that a human can directly produce analog audio adversarial examples in the physical domain. This adversary completely bypasses current acoustic defenses based on liveness and (presumably uniquely) identifying characteristics of the speaker, such as pitch.

- We demonstrate that, using commonly available plastic tubes, an attacker can change the properties of their speech in a deterministic way and manipulate ML models. We show that speaker identification and liveness models are vulnerable to our attacks. For example, an adversary can impersonate 500 other speakers using a tube. Moreover, on average, our attack is only 11% detectable by the ASVspoof 2021 spoofing detection baseline that has 97% accuracy on classifying our natural (i.e., no tube) recordings as live.

- We run our attack on live speech to confirm its feasibility and practicality. We perform a human study and show that the attack is successful over-the-air on live speech with 79% success rate. We show that our attacks can trick speaker identification models, suggesting that they should not be used in a safety-critical setting.

2 Acoustics Background

In this section, we introduce background concepts on acoustics and human speech modeling.

2.1 Acoustic Resonance

Resonance is a natural phenomenon in which objects vibrate when excited with periodic signals that contain specific frequency components [27]. These frequency components are referred to as the resonance frequencies, and they contain the fundamental frequency, \( f_0 \), (object’s natural frequency) and its harmonics. A resonating object acts as a filter that magnifies the resonance frequencies, and filters out all other frequencies in the excitation signal. In the real world, the resonance vibrations encounter resistance and losses that define the filter sharpness—referred to as the quality factor \( Q \). The filter frequency and quality factor are completely defined by the object’s shape and properties.

Acoustic resonance happens to sound waves that travel inside a hollow object, such as a tube, when it forms a standing wave [5, 27]. This phenomenon is observed in wind instruments musical notes. Similar to musical tones, human speech is produced by resonance inside the speaker’s vocal structure. In Mystique, we exploit this phenomenon and our understanding of the human speech to design a physical speech filter using tubes and perform targeted attacks on ASI.

**Resonance Frequency.** In (cylindrical) tubes, the fundamental resonance frequency \( f_0 = \frac{c_{\text{air}}}{\lambda} \) (Hz), where \( c_{\text{air}} \) is the speed of sound in air, and \( \lambda \) is the standing wave wavelength. For open-ended tubes, as in our use case, the fundamental mode \( \lambda = 2L \) where \( L \) is the tube length [10]. Thus, \( f_0 = \frac{c_{\text{air}}}{2L} \), and \( c_{\text{air}} = 20.05\sqrt{T} \) (m/s) in dry air [27], where \( T \) (°K) is the thermodynamic temperature. These equations, however, do not consider the tube diameter and air humidity. The resonance frequency of a real tube is lower than the frequency predicted by the ideal theory [7]. Thus, a correction term is added to the tube length to account for the tube diameter impact and the inertia of the standing wave outside the tube’s ends [21]. A more accurate equation is:

\[
f_0 = \frac{c_{\text{air}}}{2(L + 0.8d)},
\]

where \( d \) is the tube diameter. \( \Delta L = 0.8d \) is an empirical term derived from measurements [7], and is known as the end-correction term. However, this term is not consistent among different experiments [30].


Quality Factor. The quality factor quantifies the acoustic losses inside the tube. There are two main sources of losses \[36, 27\]: radiation loss and wall loss. The radiation loss \(d_{\text{rad}}\) is the energy loss due to acoustic radiation outside the tube \[27\]:

\[
d_{\text{rad}} = \frac{2\pi A f_0^2}{c_{\text{air}}},
\]

where \(A\) is the tube cross-sectional area. The wall losses happen because the air speed goes down to zero at the tube internal walls, hence, it leads to energy loss. Wall losses can be quantified by this damping factor \[27\]:

\[
d_{\text{wall}} = \sqrt{\frac{\mu}{\rho A f_0}},
\]

where \(\mu = 1.81 \times 10^{-5} \text{kg/ms}\) is the air viscosity, and \(\rho = 1.18 \text{kg/m}^3\) is the air density. There are other losses that are either hard to quantify, or environment dependent, or can be ignored compared to the radiation and wall losses \[2\]. Thus, the tube quality factor is:

\[
Q_0 = \frac{1}{(d_{\text{rad}} + d_{\text{wall}})}.
\]

2.2 Human Speech Modeling

Biological Characteristics. Humans generate speech using three main structures \[54\]: the lungs, the vocal folds (glottis), and the articulators as shown in Fig. 2a. The lungs produce airflow and control air pressure, this airflow in turn makes the vocal folds vibrate and modulate the passing air to produce sound (audible air vibrations)—referred to as the glottal excitation. The vocal folds physical shape controls the vibrations frequency, hence, it is considered the speech source \[54\]. The vibrating air passes through the articulators—referred to as the vocal tract—such as the pharynx, the oral cavity, the tongue, the nasal cavity, and the lips. The vocal tract forms a flexible airway that shapes the sound into the final distinctive speaker voice. The moving parts, such as the tongue and lips, change their position to produce different sounds and speech phonemes. Thus, the vocal tract is considered a linear acoustic filter \[54\]. Therefore, human speech production is studied and modeled as a sound source followed by an acoustic filter.

Source-Filter Model. The glottal excitation defines the voice pitch and can be modeled by an impulse train in the time domain \(g(t)\) and by harmonics in the frequency domain \(G(f) = F(g(t))\). The vocal tract can be modeled as a variable acoustic resonator \(H_v(f)\) that filters the glottal excitation into speech \(s(t) = F^{-1}(H_v(f) \cdot G(f))\). The resonator characteristics depends on the vocal tract size and shape; i.e. the speaker’s anatomy, and the speech phonemes vary with the tongue and lips movement \[3\]. The different parts of the vocal tract are modeled as consecutive tubes \[18\], as shown in Fig. 2b. The tubes are an acoustic resonator that amplifies certain frequencies and filters out others to shape the acoustic excitation into a specific voice and speech sound.

3 System and Threat Models

In this paper, we consider Automatic Speaker Identification (ASI)—a classification task that determines a speaker’s identity, based on their speech \[53\], from a set of enrolled speakers. Typically, the identification task can be text-dependent; i.e. the speaker has to say a predefined utterance, or text-independent; i.e. the speaker can say any utterance of their choice. Text-independent ASI provides better usability but also better security against replay attacks.

System Model. We consider a system that applies the ASI task for user identification and authentication. The system collects speech samples from its users during the enrollment phase to extract their voiceprint (speaker embeddings) and fine-tune the ASI model.

Modern ASI systems are based on speaker embedding output by deep neural networks. These models capture the speaker characteristics from a variable-length speech utterance \(s(t)\) and map it to a vector (embedding) in a fixed-dimensional space. X-vector DNN \[53, 52\] is a common ASI embedding network which consists of 3 stages: (1) feature extraction, (2) speaker embedding, and (3) classification. The first stage
extracts the mel-frequency cepstrum coefficients (MFCC) which reduce the dimensionality of the speech signal into a 2D temporal-spectral map, and applies voice activity detection (VAD) to filter out non-speech segments. Second, a time-delayed neural network (TDNN) maps the variable-length MFCC samples into a fixed-dimensional embedding (x-vectors) space. Finally, a softmax layer is applied on x-vectors to obtain the predicted identity of the speaker. The network is trained using a multi-class cross entropy objective.

During inference, the system asks the user to speak an utterance, and runs the ASI task to determine the user’s identity. The ASI task is the only access control mechanism deployed by the system. The system also applies a spoofing detection technique as a countermeasure against spoofing attacks; as we detail next in the threat model as well as Sec. 8.

Fig. 1 shows the system setup. The system runs a spoofing detector that determines whether the recorded utterance is from a live speaker or digitally produced: spoofed. If the utterance is detected to be live, the spoofing detector feeds it to the ASI model which classifies the speaker identity and gives the user access to the secure system. This system setup can be deployed for (1) logical access control applications such as phone banking services, voice assistants activation, and smart home devices, or (2) physical access control to secure buildings and spaces.

**Threat Model.** We consider an adversary that wants to attack the ASI model to be identified as a target user. We make the following assumptions about the adversary.

1. The adversary will not perform conventional spoofing techniques such as replay, speech synthesis, voice conversion, or digital adversarial examples to evade detection by the system’s spoofing detector. Note that spoofing detection techniques (Sec. 8) are based on the assumption that spoofed speech is always generated by a digital speaker, not a live human. Instead, the adversary will naturally impersonate the victim’s voice; they will change their live voice using physical objects. Our work introduces a systematic reproducible technique to give adversaries control of the identity they are able to spoof. Effectively, the adversary physically impersonates an arbitrary speaker’s voice without using a digital speaker. The attack is analog and only allows for the use of physical objects and natural sounds.

2. The adversary has no recordings of the victim’s speech.

3. The adversary has no access to the ASI model internals; i.e., this is a black-box attack. The adversary can only query the ASI model on inputs of their choice and get the model’s output score and label.
4 Attack Methodology

This section introduces our attack, Mystique, provides a theoretical intuition, and details its operation.

4.1 Overview

Fig. 1 displays Mystique’s system and attack flow. A microphone captures the speaker’s voice and feeds it to an ASI system. Mystique exploits the flawed assumption that spoof attacks must be generated from a digital speaker. The current ASI setup overlooks the acoustic environment attack vector. Mystique challenges these assumptions and performs an attack that is live by default. An attacker can speak through a specifically designed tube to induce a targeted misclassification at the ASI system, effectively impersonating a target victim.

Objectives. In Mystique, the adversary applies a transformation on their own voice to impersonate the victim’s voice; i.e., it is a live spoof attack. The transform has to satisfy four conditions: (1) analog transform on live speech, (2) an arbitrary speaker can impersonate another arbitrary victim; i.e., the attacker does not need not be a professional vocalist or have any impersonation experience, (3) the transform can be mathematically modeled to be incorporated in the attack optimization objective, and (4) the transform directly impact on the ASI model prediction.

Mystique exploits the acoustic resonance phenomenon to satisfy these four conditions. Acoustic resonance is a physical transform that is well-studied and modeled (Sec. 2.1) and has a direct impact on human voice formation, as explained in Sec. 2.2. Mystique uses hand-crafted tubes to apply the adversarial resonance transformation to the speaker’s voice. We chose tubes as our attack’s physical objects because: (1) tubes are ubiquitous and inexpensive, they can be easily found in hardware stores, plumbing pipes for example, in different dimensions, and (2) there exists extensive literature on acoustic modeling (Sec. 4.2) in musical wind instruments which are mainly of cylindrical or conical shapes. Note that the same methodology can be extended to arbitrary shapes using wave simulation and numerical analysis [6, 62].

Attack Description. The attack is as follows. The adversary models the tube resonator as a band-pass filter (BPF) transform (Sec. 4.2). The filter is fully defined by the tube dimensions. Next, the adversary runs an optimization function over the filter parameters (tube dimensions) to trick the ASI model into recognizing the voice of a chosen target speaker. In a black-box setting, we apply an evolutionary algorithm (Sec. 4.4) that uses the ASI model score and label to find the optimal tube dimensions for a given target speaker:

$$\min_p R(ASI(s'), y_t) \quad \text{s.t.} \quad s' = \text{tube}(s, p),$$

(5)

where $s$ is the original speech sample, $p$ is the tube parametrization, $y_t$ is the attack target label, $R$ is the loss, tube(.) is the mathematical model of the tube, and ASI(.) is the model under attack.

The adversary would then purchase the required tube, and speak through it to trick the system. Therefore, the adversary is able to systematically attack spoofing detection and ASI at once with an analog attack. In Sec. 4.2 we detail the mathematical model of the tubes resonance, and in Sec. 4.4 we explain Mystique’s optimization algorithm.

4.2 Modeling Resonance in Tubes

Modeling the filter corresponding to a particular tube is a key requirement for Mystique. We model the tube transfer function $H_{res}(f)$ as a sum of band-pass filters (BPFs), with a filter at each harmonic. The $i^{th}$ filter $H_i(f)$ is defined by its center frequency at the resonance harmonic $f_i$, and the filter width $\Delta f_i$ is defined by the quality factor $Q_i$ (Eqn. (7)), where $i = 1, 2, \cdots, [f_s/f_0]$ is the harmonic number, and $f_s$ is the speech sampling rate. The input speech signal $s_{in}(t)$ resonates at the tube’s fundamental frequency $f_0$ and its harmonics $f_i = i \cdot f_0$. Thus, the tube output speech signal is:

$$s_{out}(t) = \text{tube}(s_{in}, p) = \mathcal{F}^{-1}(H_{res}(f) \cdot S_{in}(f)),$$

(6)
where $\mathcal{F}^{-1}$ is the inverse Fourier transform, $S_{in}(f) = \mathcal{F}(s_{in}(t))$ is the input speech spectrum, $H_{res}(f) = \sum H_i(f)$ is the tube transfer function, and $p = (L, d)$ are the tube parameters. Note that $H_{res}(f)$ is parameterized by $p$, but we drop this parameterization to make the notation simpler. In Mystique, we adopt a simple two-pole band filter for $H_i(f)$.

**Single Tube.** Given a single tube with length and diameter parameters $p$, Eqs. [1] and [4] quantify the fundamental resonance parameters. The full harmonic range equations of $f_i$ and $Q_i$ are:

$$f_i = i \cdot f_0 = \frac{i \cdot c_{air}}{2(L + 0.8d)}; \quad Q_i = Q_0 / \sqrt{i},$$

where $i$ is a positive integer representing the harmonic number for open-ended tubes.

Our lab measurements revealed that there is about 1% mismatch between the theoretical (Eqn. [1]) and measured $f_0$. We attribute this mismatch to the end-correction term uncertainties and air humidity. Also, we estimated $Q_i$ empirically, as its change with $f_i$ depends on the dominating loss for a given tube. We found that $Q_i$ decays as $1/i, 1/\sqrt{i},$ or $1/\sqrt{i}$ give reasonable estimates and we decided to select the latter. We include both corrections in the filter formulation.

**Multiple Tubes.** Next, we extend the single tube model into a structure of multiple consecutive tubes of different lengths and radii to increase Mystique’s degrees of freedom and the set of possible filters. The extended structure can reach a wider range of spoofed identities, hence, it increases the attack success rate as shown in Section 6.1.

Resonance inside connected open-ended tubes happens when the acoustic impedance between the connected tubes equal an open-end impedance \[^{55}].\] This condition is mapped to the following equation for each two tubes intersection:

$$A_1 \cdot \cot(2\pi f L_1/c_{air}) = A_2 \cdot \cot(2\pi f L_2/c_{air}),$$

where $A_1$ and $A_2$ are the two tubes cross-sectional areas, $L_1$ and $L_2$ are their lengths. We solve this non-linear equation numerically to obtain the resonance frequencies $f_i$’s.

**Validation.** We validate the tube resonance model by measuring the resonance from real tubes and comparing it with our BPFs model. This measurement helps us verify that the model is reliable to be used in Mystique’s optimization objective. To observe the tube resonance, independent from the speakers voice and the linguistic content in speech, we use a chirp signal as the audio source \[^{51}].\] The chirp signal exponentially spans the frequency range from 100 Hz to 3700 Hz and is 3 seconds long. We play the chirp signal using a Pixel phone speaker and use the setup in Fig. 5 for recording.

Fig. 3 shows the Fast Fourier Transform (FFT) of the tube output signals for 3 different tube configurations: two single tubes and one two-tube structure. The vertical dotted lines indicate the theoretical resonance frequencies. The figure also plots the FFT of the BPF model applied to the same chirp signal (top plot in each figure). One can observe how the theoretical equations and BPF model match for all of the real tubes. We validate the tube resonance model by measuring the resonance from real tubes and comparing it with our BPFs model. This measurement helps us verify that the model is reliable to be used in Mystique’s optimization objective. In Mystique, we adopt a simple two-pole band filter for $H_i(f)$.

**4.3 Attack Intuition**

Speech technology applications such as speech recognition, speaker identification, and keyword spotting are highly sensitive to the acoustic environment. Models trained on clean speech recordings often fail in real world scenarios \[^{28} [13] [23].\] Usually, the training data has to be augmented with simulated environmental effects such as noise and echo \[^{28} [13] [23].\] The same applies for speech adversarial examples. Adversarial perturbations do not succeed over-the-air when the environmental variations are not considered in the optimization objective \[^{44} [4].\] Hence, one of the fundamental intuitions behind Mystique is that if the acoustic environment falls outside the expected distribution, the model predictions will become unreliable.

Still, one can wonder why a tube (resonator) has such a high impact on the ASI model’s performance. In Section 4.3.1 we theoretically show that tubes affect the estimated pitch. Next, we empirically validate that tube parameters are statistically significant predictors of pitch shifts between input and output signals. Such
pitch shifts introduce distribution shifts w.r.t the real-world utterance datasets used to train speech models. It has been well-established that such distribution shifts reduce model performance at inference time [45, 58]. In particular, as we will discover, ASI is sensitive to the pitch of the speech signal; therefore, applying the tube is expected to change the classification result.

### 4.3.1 Tubes Cause Pitch Shifts

McAulay and Quatieri [34] present a pitch estimation algorithm; it is based on fitting a speech signal to a sum of sinusoidal signals with frequencies being the harmonics of the fundamental frequency. First, they use the peaks of the Short-time Fourier transform (STFT) of a time domain signal \( s(t) \) to represent it as a sum of \( L \) sine waves:

\[
    s[n] = \sum_{\ell=1}^{L} A_\ell \exp[j(n\omega_\ell + \theta_\ell)].
\]

The values of \( A_\ell, \omega_\ell, \) and \( \theta_\ell \) represent the amplitudes, frequencies, and phases of the STFT peaks of the speech signal. Then, they try to find the value of \( \omega_0 \) which fits \( s[n] \) to \( \tilde{s}[n, \omega_0] \), defined as:

\[
    \tilde{s}[n, \omega_0] = \sum_{k=1}^{K(\omega_0)} \tilde{A}(k\omega) \exp[j(nk\omega_0 + \phi_k)],
\]

where \( \omega_0 \) is the signal pitch, \( K(\omega_0) \) is the number of harmonics in the signal, \( \tilde{A}(k\omega) \) is the vocal tract envelope, and \( \phi_k \) is the phase at each harmonic. Finally, the pitch is estimated by minimizing the mean squared error \( \epsilon(\omega_0) = P_s - \rho(\omega_0) \), where \( P_s \) is signal’s power which is a constant. Therefore, we only need to minimize \(-\rho(\omega_0)\), or equivalently maximize \( \rho(\omega_0) \):

\[
    \max \quad \rho(\omega_0) \quad \text{subject to} \quad \rho(\omega_0) = \sum_{k=1}^{K(\omega_0)} \tilde{A}(k\omega) \left[ \sum_{\ell=1}^{L} A_\ell \left| \text{sinc}(\omega_\ell - k\omega_0) \right| - \frac{1}{2} \tilde{A}(k\omega_0) \right].
\]

As discussed in Section 4.2, the tube results in a resonance effect, modeled as a set of bandpass filters at the resonance frequencies of the tubes. As such, some of the frequency components of \( s(t) \) will be dampened. We represent this effect as \( A_\ell = 0 \) for \( \ell \in \mathcal{L} \) as well as their submultiples \( \omega_0 \in [K(\omega_0)] \), where \( \mathcal{L} \) represents the set of non-resonant frequencies of the tube:

\[
    \max \quad \rho(\omega_0) \quad \text{s.t.} \quad A_\ell = 0 \quad \forall \ell \in \mathcal{L}, \forall \omega_0 \in [K(\omega_0)]
\]
Note that Eqn. (11) is a constrained version of Eqn. (9). We can solve the latter by maximizing the Lagrangian:

\[ p(\omega, \eta) = \rho(\omega_0) - \sum_{k=1}^{K(\omega_0)} \sum_{\ell \in \mathcal{L}} \eta_{k\ell} A_\ell \]  

(12)

where the matrix \( \eta = [\eta_{k\ell}]_{K(\omega_0) \times |\mathcal{L}|} \) represents the Lagrange multipliers. Instead of directly maximizing Eqn. (12) and finding \( \eta \), we re-write Eqn. (10) separating the components in and outside of \( \mathcal{L} \):

\[ \rho(\omega_0) = \rho_f(\omega_0) + \sum_{k=1}^{K(\omega_0)} \tilde{A}(k\omega_0) \sum_{\ell \in \mathcal{L}} A_\ell \left| \text{sinc}(\omega_\ell - k\omega_0) \right|, \]  

(13)

where

\[ \rho_f(\omega_0) = \sum_{k=1}^{K(\omega_0)} \tilde{A}(k\omega_0) \left[ \sum_{\ell \notin \mathcal{L}} A_\ell \left| \text{sinc}(\omega_\ell - k\omega_0) \right| - \frac{1}{2} \tilde{A}(k\omega_0) \right]. \]  

(14)

is the objective function for estimating the pitch of the filtered signal. Next, substituting Eqn. (13) in Eqn. (12):

\[ p(\omega, \eta) = \rho_f(\omega_0) + \sum_{k=1}^{K(\omega_0)} \sum_{\ell \in \mathcal{L}} \left( \tilde{A}(k\omega_0) \left| \text{sinc}(\omega_\ell - k\omega_0) \right| - \eta_{k\ell} \right) A_\ell \]  

(15)

Using the KKT conditions [12], we know for \( p(\omega_0, \eta^*) \) to be the maximizer of Eqn. (15), the second term should vanish. Given \( A_\ell > 0 \), we should have that:

\[ \eta_{k\ell} = \tilde{A}(k\omega_0) \left| \text{sinc}(\omega_\ell - k\omega_0) \right|. \]  

(16)

But that means \( \rho_f(\omega_0) = p(\omega_0, \eta^*) \) is the exact solution to Eqn. (11), i.e., the equality constraint holds perfectly.

Having established that the second optimization problem is a constrained version of the first, it follows that \( \Omega \), the feasibility set of Eqn. (9) is a subset of \( \Omega_f \), the feasibility set of Eqn. (11). Then, unless \( \mathcal{L} = \emptyset \) (which trivially results in \( \Omega = \Omega_f \)), there exists \( \omega_0 \in \Omega \setminus \Omega_f \) such that \( \omega_0 \) is a valid estimated pitch that has been filtered out by the tube. Therefore, we have shown that the tube will cause shifts in the estimated pitch.

**Validation.** To verify this intuition, we design an experiment to study the correlation between the pitch shift and the change in the classification result. We played samples from the VoxCeleb dataset through three tubes of different lengths (corresponding to different resonance frequencies). For each sample, we estimated the pitch of both signals (original and output) using CREPE [26] which provides a time-domain signal of the signal pitch. Given that the pitch varies in the duration of each utterance, we need to account for different speakers, utterances and original clip recordings to establish a generalized relationship between pitch shifts and tube parameters (diameter and length).

Using VoxCeleb metadata (speaker ID, and clip hash) this leads us to calculate the difference between average estimated pitch frequencies per audio clip. We regress this pitch difference using an ordinary least squares model with a design matrix containing tube parameters and 2060 audio samples. The linear regression model achieves an \( R^2 = 0.552 \). Therefore, the tube parameters explain at least 55% of the pitch shift variances. P-values achieved are \( 1.77 \times 10^{-26} \) and \( 2.99 \times 10^{-149} \) for length and parameter, respectively, which means that these tube parameters are good regressors of the shifts introduced by the tube in a variety of recording conditions, utterances and speakers.

**4.4 Mystique’s Algorithm**

In Sec. 4.2 we parameterize tubes by the quality factor \( Q_0 \) and the fundamental frequency \( f_0 \). Although, for a single tube, the search space is small enough to be bruteforced within a few hours, we find that in many cases we can speed up the attack using optimization. More precisely, we experiment with gradient-free non-convex optimization algorithm from a family of evolutionary algorithms called *differential evolution*.
Figure 4: Average reachable target search performance across all of the participants with a Huggingface model

Evolutionary algorithms were used extensively in the past to attack machine learning and DE, in particular, proved effective at discovering adversarial examples for both vision and language models. DE performs the search by picking three data samples from an underlying population and combining the best performing one with the difference between the other two.

In the search algorithm, we set boundary conditions on the tube dimensions. We allow the tube length to range from 0.1 m to 3 m, and the diameter from 1 cm to 15 cm. Hence, based on Eqn. (1), \( f_0 \) ranges from 50 Hz to 1 kHz, and its quality factor \( Q_0 \), Eqn. (4), ranges from 5 to 100. The attacks are performed in a black-box fashion, requiring only per-class probabilities of the acoustic model. We find that within 100 model invocations, as is demonstrated in Figure 4, we could find 46% ± 12 of all possible reachable targets, whereas at 250 invocations it grows to 55% ± 14. Despite relatively low performance, our GA enables the attacker to within minutes check with a reasonable probability if a user can be matched with a given target. We further explain the results and list the underlying algorithm with hyperparameters in Appendix B.

5 Experimental Setup

We design an experimental setup, comprising speech datasets, ASI models, spoofing detection models and a physical measurement setup to evaluate our proposed attack, Mystique; our evaluation answers the following questions:

Q1. How well does Mystique perform as an impersonation attack on ASI models?

We validate the feasibility of Mystique on VoxCeleb test set using the resonance filter model. We evaluate two ASI models and show that Mystique can attack both of them. Mystique can achieve 500 successful targeted attacks, on average, for each adversarial speaker in the test set.

Q2. Does Mystique’s impersonation succeed in real-world?
We build a physical recording setup and run Mystique over-the-air using a standard dataset (VoxCeleb). We also conduct a user study and evaluate Mystique on live speech. We show that Mystique achieves up to 61% success rate on standard dataset and 79% on live speech.

Q3. How effective is spoofing detection against Mystique?

We validate the undetectability of Mystique using two spoofing detection models, which fail to discriminate utterances by Mystique as “not-live”.

5.1 Datasets and ML Models

ASI Models. We analyze the impersonation performance of Mystique against two state-of-the-art ASI models: (1) the x-vector network implemented by Shamsabadi et al., and (2) the emphasized channel attention, propagation and aggregation time delay neural network (ECAPA-TDNN), implemented by SpeechBrain. Both models were trained on the VoxCeleb dataset, a benchmark dataset for the ASI task. The x-vector network is trained on 250 speakers using 8 kHz sampling rate. ECAPA-TDNN is trained on 7205 speakers using 16 kHz sampling rate. Both models report a test accuracy within 98-99%.

Spoofing Detection Models. We consider two different state-of-the-art spoofing detection models: baselines from the ASVspoof 2021 challenge for logical access (LA) and physical access (PA) tasks. The ASVspoof 2021 PA task objective is to discriminate between live-human speech and replay attack via loud speakers. The attack is thus physical; replaying recordings over-the-air. The ASVspoof 2021 LA task objective is to differentiate between live speech and artificially generated speech using text-to-speech, voice conversion, or hybrid algorithms. The ASVspoof 2021 LA task considers only logical attacks; i.e. the adversary feeds the spoofed utterance digitally to the ASI model and does not play it over-the-air. The two tasks are separated and countermeasures on one of them would not necessary succeed on the other. We use the official implementation of ASVspoof 2021 baselines. These baselines are trained using an end-to-end RawNet2 deep neural network on ASVspoof 2019 training data for the LA and PA tasks.

Evaluation Dataset. Both ASI models are trained on VoxCeleb. Thus, we use VoxCeleb as our test dataset. We select a subset of 91 speakers, 45 female and 46 male speakers, that are common in the training dataset of both models. We select 20 random utterances per speaker on which both models achieve 100% accuracy.

User Study. We conduct a user study to test the performance of Mystique on live user utterances across three representative tubes. The user study involves two stages. In the first stage, 12 participants record a set of 50 utterances (Appendix) using a microphone, without a tube. We then pass these recordings through the filters representing each tube, and obtain the classification result for each filtered recording. In the second stage, we ask each participant to speak each utterance through each tube to compare the live classification result with the one obtained from the filter. We do not provide the participants with any additional instructions.

We recruited 12 individuals (6 male, 6 female, age:18-30) to conduct our study. We obtained IRB approval from our institution to conduct the study; we collected no personal information, obtained informed consent from each participant, and followed health protocols. We use the ASI models described above, without retraining as to mimic a realistic attacker, which would attack black-box models. We use the physical setup, described below, to conduct the user study.

5.2 Physical Setup for the Attack

We design and implement a physical measurement setup to conduct the attack over the air. Fig. 5 visualizes our setup which comprises tube(s), a recording device, and the recording environment.
Tubes. We use two sets of tube in this work. We conduct single-tube experiments using PolyVinyl Chloride (PVC) pipes, purchased from a hardware store. The dimensions of these tubes are listed in Table 2. The diameters of the tubes represent those of popular plumbing pipes: 3.45cm, 4cm, and 5.2cm. We used different lengths to generate resonance frequencies within the fundamental frequency range of human speakers. Our evaluation in Sec. 6.1 shows that these frequencies are more successful in changing the classification.

For two-tube experiments, we 3D printed four tubes using Formlab’s Form 2
cia Black Resin material. The 3D printer enables fine-grained control over the tube radius and thickness. We printed the tubes with a 50 µm resolution for a smoother finish and a thickness of 2 mm. We ensured no support material was on the inside of the tube. The dimensions (length, diameter) of the four tubes are: (9.53cm, 2.1cm), (10cm, 1cm), (11.44cm, 0.98cm), and (14.53cm, 2.1cm). We used these tubes to construct three two-tube devices, as listed in Table 2. We connect the tubes with High Density Fiberboard (HDF) connectors cut on a 150Watt 10.6 µm CO2 laser cutter; we secure the connection with clay at the edge of the connector.

Recording Environment. We conducted the experiment in a lab space with dimensions 8 × 3.6 × 3.6 m. We built an audio chamber to isolate the experiment from the background noise and speech interference from adjacent rooms; this helps unify the acoustic environment throughout the experiments. The chamber is a wooden box lined with acoustic panels to absorb the noise and minimize reverberation. We attached floating suspension loops to the chamber’s ceiling to hold the tube in the air as shown in Fig. 5. Suspending the tube minimizes its surface mechanical vibrations. We used a Blue snowball microphone, placed as Fig. 5 to capture the tube output signal. The setup is inspired by the design of musical instruments measurement environments. We use a Google Pixel 2 phone as a digital speaker to play sound over-the-air. The recording is controlled by a MacBook Pro laptop. We used python-sounddevice library to automate the recordings.

\[https://formlabs.com/3d-printers/form-2/\]
\[https://formlabs.com/store/black-resin/\]
\[https://python-sounddevice.readthedocs.io/en/0.4.4/\]
6 Mystique’s Evaluation Results

We describe the evaluation results, which provide detailed answers to the questions in Section 5.

6.1 Impersonation Attack at Scale

First, we test Mystique’s impersonation attack feasibility on the full test set to address our first evaluation question from Sec. 5. We run Mystique on the VoxCeleb test set, representing the adversarial speakers, and find the range of successful impersonation attacks and the corresponding set of adversarial tubes. We consider structures of N-tubes, where N ≤ 2. Hence, the resonating frequencies depend on three parameters (degrees of freedom): the tubes lengths $L_1$, $L_2$ and the tubes cross-sectional area ratio: $\text{ratio}_A = \left(\frac{d_2}{d_1}\right)^2$.

We apply the filter transform to the 91 speakers in our evaluation dataset and test them against the ASI models. We define the search space boundaries as: the fundamental resonance frequency $f_0$ ranges from 50 Hz to 1 kHz and its quality factor $Q_0$ ranges from 5 to 100, such that $f_0$ falls in the typical range of human voice. We sample from this range using $f_0$ step size of 10 Hz and $Q_0$ step size of 5. According to Eqn. (1), Eqn. (4) of a single-tube, the tube dimensions range is: the length ranges from 10 cm to 3 m, and the diameter ranges from 1 cm to 15 cm, which is a practical range. For two-tube structures, each tube length can range from 5 cm to 120 cm with 5 cm step size, and the areas ratio ranges from 1 to 10 with step size of 1, and $f_i$’s are found from Eqn. (8).

Fig. 6 shows SpeechBrain model’s number of successful attacks (false predictions) for each true (adversarial) speaker. Fig. 10 in appendix shows the same for the x-vector model. As the figure shows, by controlling the tube dimensions, Mystique can reach a wide range of successful impersonation attacks. Specifically, a speaker can impersonate 500 target speakers on average on SpeechBrain model and 137 on x-vector model. Recall that the models are initially 100% accurate on the selected evaluation dataset. Hence, this experiment shows that Mystique is capable of forming an adversarial impersonation attack on speaker identification models.

Next, we analyze the false predictions in terms of the successful filter (tube) parameters and the predictions distribution to interpret how the attack works. We found that:

1. The attack is most effective when $f_0$ lies in the lower frequency range $f_0 \leq 400$ Hz with a high quality factor $Q_0 \geq 50$ as shown in Fig. 7, Fig. 11 in appendix. The figure shows the false predictions histogram at different filter parameters. This observation matches our intuitions from Sec. 4.3. The significant $f_0$ range falls within the typical human pitch range. An adult woman pitch range is 165 to 260 Hz on average, and an adult man’s is 85 to 155 Hz. Moreover, low frequency speech range carries more information than the higher frequency range [33]. Hence, this range of $f_0$ will have a stronger impact on the pitch and the significant spectrum, and thus the model prediction. Also, a high quality factor means a sharper filter; fine-grained selection.

2. When the model mis-identifies a speaker, it will more likely predict a same-sex speaker with 80% chance. Fig. 8 shows the prediction confusion matrix split by the true and predicted speakers sex. As the figure shows, the cross-sex speakers submatrix is sparse compared to the same-sex submatrix. The same applies for the x-vector model in Fig. 12.

6.2 Over-the-air Attack

Next, we validate Mystique’s impersonation attack over-the-air using our physical setup in Fig. 5 to answer the second evaluation question. We conduct this experiment on VoxCeleb as a standard dataset for ASI—Sec. 6.2.1 and also on live speech from our user study participants—Sec. 6.2.2.

6.2.1 Standard Dataset Evaluation

Because of the physical resources (mainly run-time) limitations, we select a subset of the evaluation speakers to form the adversarial speakers set. We also select a subset of the possible tube dimensions to run the over-the-air attack. Specifically, we randomly select 40 speakers, 20 males and 20 females, out of the 91 speakers dataset. There are 20 utterances for each speaker; a total of 800 four-second long utterances. The subset is balanced and representative of the full dataset. For the single-tube setting, we select 6 random
tubes of various dimensions that have $f_0, Q_0$ in the most significant range—Fig. 7. We purchase them from the hardware store. While for the two-tube setting, we build three structures from the four 3D printed tubes in Sec. 5.2. The selected tubes parameters are listed in Table 1 and 2.

We use the Pixel phone to simulate the speaker and play the VoxCeleb utterances over-the-air for each tube(s) configuration. We record the tube output sound using the physical setup. We place the speaker on a separate tripod to allow acoustic propagation only through the air; i.e., no sound is transmitted to the microphone via vibrations through the recording table. We allow a 3 sec silence between consecutive utterances till resonance effect passes off. We repeat the recordings 6 times to account for any environmental variations.

Table 1 shows the number of successful attacks (false predictions) per each real tube and compares it with the successful attacks using the filter model. First, “Real” columns (6 and 9) report the number of successful attacks of the 40 speakers using the real tubes. Each speaker can impersonate up to 5 speakers identities on average using an individual tube. We found that different utterances sometimes lead to different false predictions per speaker-tube pair. Second, “Filter” columns (7 and 10) show the number of successful attacks using the same tube’s filter model. The filter successful attacks are on the same magnitude as the real tube attacks. Finally, the “Match” columns (8 and 11) show the matching rate between the real and simulated tubes attacked identities. The match rate ranges from 38.7% to 61.62%, 48% on average. Hence, Table 1 confirms that speaking through a tube forms a real and effective attack on the ASI task. Mystique’s optimization objective and resonance model are successful over-the-air. They do not map the reality 100%, yet, they do not overestimate the attack power. Note that Mystique models resonance as a linear BPF filter, which is an approximation. A more accurate model is to use wave simulation engines at the expense of increased computation complexity.

Table 2 shows the attack success rate over-the-air for the two-tube structures. Mystique’s targeted attack succeeds more than 50% of the time.
Figure 7: False predictions histogram on SpeechBrain. Low $f_0$ and high $Q_0$ values result in more false predictions.

6.2.2 Live Speech Attack

Finally, we run Mystique on the 12 participants natural recordings, 50 utterances each, and find the set of successful attacks (impersonated identities) per participant. Fig. 9 shows the number of successful attacks on SpeechBrain model. Fig. [13] in appendix shows the same for x-vector model. An arbitrary speaker can impersonate 163 (117 for x-vector) target identities on average using a single tube.

Next, we ask the participants to speak the same 50 utterances through three of our tubes. We record the tube output and evaluate the recordings on the ASI models. Table 3 reports the percentage of successful physical live attacks over-the-air for each participant. Column 5 shows the average success rate of the 3 tubes for each participant where the lowest is 47.05%.

We can improve this success rate by fine-tuning the filter parameters to each participant voice characteristics. When we apply a voice envelope calibration to the filter gain, we observed a relative increase of up to 21% in Mystique’s success rate. The last column in Table 3 shows the average success rate per participant when the filter is calibrated to their estimated voice envelope. Thus, filter personalization can further increase Mystique’s efficacy in real-world. Finally, we observed the same skew in successful attack speakers sex as reported for VoxCeleb dataset in Fig. 8 the cross-sex submatrix is sparse.

6.3 Spoofing Detection

Finally, we evaluate the effectiveness of spoofing detection against our attack, Mystique, to answer the last question in Sec. 5. We utilize the two spoofing detection models described in Section 5.1 namely ASVspoof 2021 LA and ASVspoof 2021 PA baselines. The baselines are trained on the LA and PA training partitions of the ASVspoof 2019 dataset.

Table 4 reports the spoofing detection results of our recordings from the 12 participants (50 recordings per each speaker) in 4 different settings (no-tube, tube 3, tube 4, and tube 6). The table shows the percentage of recordings that are classified as live by each model. Our results demonstrate that state-of-the-art spoofing detection models are unreliable against Mystique. First, the PA model fails to generalize beyond its training
Male
Female
Male
Female

Figure 8: SpeechBrain’s prediction confusion matrix split by the true and predicted speakers sex. The cross-sex submatrix is sparse.

<table>
<thead>
<tr>
<th>Tube</th>
<th>Tube Dimensions</th>
<th>Resonance Parameters</th>
<th>X-Vector False Predictions</th>
<th>SpeechBrain False Predictions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Tube</td>
<td>Dimensions</td>
<td>Resonance</td>
<td>Parameters</td>
</tr>
<tr>
<td>1</td>
<td>40.6</td>
<td>3.45</td>
<td>402.16</td>
<td>58</td>
</tr>
<tr>
<td>2</td>
<td>61.3</td>
<td>4</td>
<td>270.70</td>
<td>68</td>
</tr>
<tr>
<td>3</td>
<td>87</td>
<td>5.2</td>
<td>191.48</td>
<td>77</td>
</tr>
<tr>
<td>4</td>
<td>99.4</td>
<td>3.45</td>
<td>170.89</td>
<td>79</td>
</tr>
<tr>
<td>5</td>
<td>120.3</td>
<td>5.2</td>
<td>140.20</td>
<td>76</td>
</tr>
</tbody>
</table>

Table 1: Physical evaluation of Mystique over-the-air for 40 speakers, 20 utterances each; i.e. 800 total inferences. Real: number of successful attacks of the real tube, Filter: number of successful attacks of the corresponding filter model, Match: the number (percentage) of matched attacks between filter and real tube.

environment. Although the accuracy of the ASVspoof 2021 PA baseline is above 95% on the development partition of ASVspoof 2019 dataset, this PA baseline classifies all of our no-tube recordings as not-live. As the PA baseline fails in the no-tube setting, its classification results on the tube-based recordings cannot be considered reliable. Second, the ASVspoof 2021 LA baseline correctly classifies 64% of our no-tube recordings as live. Additionally, it classifies more than 83% of the tube-based recordings as live. Thus, it does not detect digital artifacts in the tube recordings. These results highlight that spoofing detection is unreliable and over-fitted to its training data distribution.

7 Discussion

Lessons and Insights. In Section 6 we demonstrate that an attacker can fool a whole family of defenses that rely on non-human features for their protection. Contrary to existing literature, we demonstrate that physical human attacks exist and can be performed without much effort or knowledge. Our attacks are realizable using commodity pipes, available at hardware stores. Importantly, not only do we demonstrate theoretical attack existence, but also show that such attacks trick models in the real world with real users. It is worth noting that current literature on acoustic adversarial examples has had some struggles performing their attacks in the real world and attributed it to the environmental noise [41]. Our attacks on the other
Table 2: Two-tube structures dimensions and attack success rate over-the-air.

<table>
<thead>
<tr>
<th>Tube</th>
<th>Tube Parameters</th>
<th>Success Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>9.53 2.1 10 1</td>
<td>853.1 66.6%</td>
</tr>
<tr>
<td>8</td>
<td>11.44 0.98 8.9 3.4</td>
<td>901.55 50%</td>
</tr>
<tr>
<td>9</td>
<td>14.53 2.1 10 1</td>
<td>600.4 100%</td>
</tr>
</tbody>
</table>

Table 3: User study participants percentage (%) of successful over-the-air attacks.

<table>
<thead>
<tr>
<th>ID</th>
<th>Tube3</th>
<th>Tube4</th>
<th>Tube6</th>
<th>Avg</th>
<th>Calibrated Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>50.0</td>
<td>44.44</td>
<td>66.67</td>
<td>53.70</td>
<td>65.0</td>
</tr>
<tr>
<td>1</td>
<td>58.82</td>
<td>81.82</td>
<td>57.14</td>
<td>65.93</td>
<td>65.93</td>
</tr>
<tr>
<td>2</td>
<td>66.67</td>
<td>72.73</td>
<td>77.78</td>
<td>72.40</td>
<td>72.58</td>
</tr>
<tr>
<td>3</td>
<td>63.64</td>
<td>83.33</td>
<td>75.0</td>
<td>73.99</td>
<td>78.70</td>
</tr>
<tr>
<td>4</td>
<td>66.67</td>
<td>58.33</td>
<td>71.43</td>
<td>65.48</td>
<td>73.15</td>
</tr>
<tr>
<td>5</td>
<td>50.0</td>
<td>42.86</td>
<td>55.56</td>
<td>49.47</td>
<td>49.47</td>
</tr>
<tr>
<td>6</td>
<td>46.15</td>
<td>54.55</td>
<td>80.0</td>
<td>60.23</td>
<td>60.71</td>
</tr>
<tr>
<td>7</td>
<td>66.67</td>
<td>77.78</td>
<td>80.0</td>
<td>74.81</td>
<td>74.81</td>
</tr>
<tr>
<td>8</td>
<td>43.75</td>
<td>42.86</td>
<td>54.55</td>
<td>47.05</td>
<td>52.06</td>
</tr>
<tr>
<td>9</td>
<td>50.0</td>
<td>60.0</td>
<td>50.0</td>
<td>53.33</td>
<td>53.33</td>
</tr>
<tr>
<td>10</td>
<td>66.67</td>
<td>62.5</td>
<td>80.0</td>
<td>69.72</td>
<td>75.0</td>
</tr>
<tr>
<td>11</td>
<td>50.0</td>
<td>61.54</td>
<td>72.73</td>
<td>61.42</td>
<td>69.17</td>
</tr>
</tbody>
</table>

hand do work in the real world.

**Tubes in the limit.** Although we use tubes to perform Mystique, the tubes are just an example of a whole family of attacks that utilize the environment to shape the signal. Current literature understands acoustic modelling well [6], and there are readily available solutions that allow for generation of objects with given acoustic properties [62].

**Defenses.** Having established a major vulnerability in spoofing detection systems leads to a question on how one stops such attacks. The immediate defense would consider a variant of adversarial training [67], where the training set is augmented with samples after applying the resonance filters. However, it is not clear whether such a defense approach is reliable, or even desirable. An attacker can simply use objects with different filter profile to render the defense unsuccessful; the defender cannot predict what filter the attacker would deploy. Second, given that the tube is effectively changing the characteristics of the speech (such as pitch), such a defense might break the natural accuracy of the task. Unfortunately, it is not clear how one solves the problem given the systemic problem underlying the task—*what is said is not what is heard*, because the signal changes as it moves through space. What is clear is that the solution would have to incorporate properties of the medium, not just the speakers features.

**Reproducibility.** From formulating the original idea to completing the experiments described in this paper, this project took around a year. Although we converged on the method relatively quickly, it took significant amount of effort to match theory with practice. For reproducibility, below we make a note of the things that slowed us down significantly and required non-trivial debugging. First, the use of Bluetooth or Wifi operated devices introduces significant problems because of occasional variable lag and interference. Second, during the theoretical and practical matching, it is important to isolate the setup as much as possible. In our case, matching $f_0$ and $Q$ without the acoustic chamber was extremely challenging. Third, distance to the microphone and its’ directionality matters—nothing should be blocking the opening of the tube, as otherwise it leads to additional echo and changes the filter. The same observation is reported in resonance measurements literature [7]. Fourth, experiments ran on different days lead to different results, because of a change in speed of sound with temperature and humidity – its best to conduct hardware calibration and the
The dotted line shows the average number per true speaker.

Table 4: The performance of ASVspoof 2021 logical access and physical access baselines on our user study recordings.

<table>
<thead>
<tr>
<th>Recording</th>
<th>Logical access</th>
<th>Physical access</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Tube</td>
<td>64%</td>
<td>0%</td>
</tr>
<tr>
<td>Tube 3</td>
<td>92%</td>
<td>0%</td>
</tr>
<tr>
<td>Tube 4</td>
<td>93%</td>
<td>0%</td>
</tr>
<tr>
<td>Tube 6</td>
<td>83%</td>
<td>0%</td>
</tr>
</tbody>
</table>

Evaluation on the same day. Fifth, when producing tubes with a 3D printer, the material on the inside of the tube should be smooth for best matching with theory. Furthermore, when printing tube structures, it’s important to position it in a way to not get any of the support material on the inside of the pipes.

**Limitations.** Despite our study clearly highlighting a flaw in the design of current defenses, there are a number of limitations in the current evaluation. First, we only considered simple tube structures, restricting the range of possible adversarial transformations. For example, we observed that using a single tube, it is highly likely to perform attacks in-sex than across-sex. Second, we ran our user study inside of an ‘acoustic vacuum’, restricting the impact of the environment, underestimating how hard it would be to perform the attacks in practice. Third, we only considered a small number of utterances, potentially underrating the overall attack performance.

8 Related Work

The literature on computer-based voice authentication is vast, and dates back to at least 1960s [25].
Attacks on ASI. We start by describing the four most common attacks: (1) speech synthesis, (2) voice conversion, (3) replay attacks and (4) adversarial examples. In speech synthesis, an adversary trains a (possibly few-shot) speech synthesis model on samples recorded from the target speaker. The adversary then uses this model to convert any text into speech in the target speaker’s voice [60, 63, 66]. Alternatively, a voice conversion model can be trained to convert any spoken utterance into the target speaker’s voice [37, 50, 73]. In replay attacks, the adversary records the speaker’s voice and replays the recorded utterance to grant access into the secure system [31]. Despite being the simplest, replay attacks are often the strongest [68]. Finally, since many of modern ASI models rely on machine learning components, they inherit the vulnerability to adversarial examples using standard gradient-based attacks [32, 24, 15].

Defenses against Acoustic Attacks. What these attacks have in common is that the adversarially-generated sample would need to be generated, and transmitted digitally and reproduced through a (digital) speaker. Defense mechanisms, therefore, include (1) detecting the electronic footprint of the digital speaker (known as spoofing detection), or (2) verifying that the speaker is a live human. Spoofing detection relies on patterns extracted from the acoustic signal to classify it as a legitimate or fake sample. Chen et al. used a smartphone’s magnetometer to detect the use of a loudspeaker [14]. Blue et al. tell electronic and human speakers apart by analyzing individual frequency components of a given speech sample [10]. This is possible because the authors assumed a non-human adversary who is remote and incapable of changing the electronic speaker properties. Yan et al. calibrated individual speakers in the near field of the speakers to tell humans and electronic speakers apart [72].

Second, liveness detection leverages other sensing modalities such as visual, acoustic and EM signals to determine the liveness of the acoustic signal. Meng et al. used an active radar to project a wave onto the face of the speaker and then detect shifts introduced to it from facial movement [35]. Owczarek and Slot employed a video feed to correlate audio with facial movement to protect against spoofing attacks [42]. Zhang et al. analyzed hand movement to detect live speech by turning a smartphone into an active sonar [75].

In the above, the security of the overall authentication system is based on the spoofing and liveness detection systems, as well as an independent speaker verification system. Usually, each of these systems is studied, designed, and evaluated separately. To evade them simultaneously, we create adversarial examples for the audio domain under the constraint of being physically reproducible by humans in the real world. Finally, there exists a class of defenses that restrict the attack surface by reducing attacker capabilities. Zhang et al. used individual recordings from a stereo microphone to calculate time difference of arrival [74] to detect replay attacks. Blue et al. used two microphones to restrict the adversary to a 30 degree cone and protect against hidden and replay commands [9]. Wang et al. used correlates from a motion sensor to detect and reject hidden voice commands [64]. Since we assume a human adversary in this paper, they would bypass such defenses by interacting with the system as is intended.

Physical Adversarial Examples. Such physical adversarial examples are common in the vision domain, but have not been produced for acoustic tasks. Example adversarial objects include eyeware [16, 49], tshirts [69, 70], headwear [29, 76] and patches [61]. Although these objects were re-created in the real world, there is an important distinction to be drawn here. These objects all apply perturbations that were initially designed for the digital space and which were later retrofitted with sophisticated machinery such as printers for both paper and clothing to realize them in the physical domain. Our attacks, on the other hand, require little to no preparation because they directly restrict the search space of perturbations to those that can be easily realized physically: all that one needs to do is get a tube, cut it to an appropriate length, and speak through it. Most importantly, our attacks target a different property of the physical world—we use the environment to shape the signal, rather than exploit errors in the ML model. An alternative in the computer vision world would be to change the physics of light passing through the environment such that a camera sensor “sees” a different picture. An example of such attack would be to spray heavy aerosols to increase light dispersion or to increase temperature to make air less dense and diffract light in a different way.
9 Conclusion

We demonstrate that a human adversary can reliably manipulate voice-based identification systems using physical tubes. Our attacks highlight acoustic intricacies that were largely ignored by prior literature, namely, the acoustic environment. Additionally, we highlight that practically all current defenses assume that the adversary is non-human and focus on verifying this assumption. Our human-produced attacks show that this assumption does not hold in the first place. Our work is largely motivated by the recent public concern over systems that differentiate individuals using ML [13, 46]. We focus on one of the biometric markers: speech. We demonstrate that models differentiating individuals through speech are vulnerable and should not be used in security-critical applications.

To better understand the fundamental issue with acoustic tasks, one needs to question an assumption made when applying ML: the existence of objective labels [8]. Although this question is discussed in the philosophy literature [20], it also has implications to the reliability of deployed systems. Models learned from subjective historical assessments encoded through labels in the data inherit these assessments’ biases and vulnerabilities. Minority groups often get marginalised because of subjective labels in ML training sets [22]. In this paper we demonstrate that subjective nature of speech can be exploited to jeopardize the security of a critical system. Concretely, for the speaker identification, we should ask whether a human speaking through a pitch-shifting medium must still be identified as their true selves? In other words, is it realistic to expect invariance to medium changes?

Acknowledgments

This work was supported by DARPA (through the GARD program), CIFAR (through a Canada CIFAR AI Chair), by NSERC (under the Discovery Program, and COHESA strategic research network), by a gift from Intel. We also thank the Vector Institute’s sponsors.

References


Lindasalwa Muda, Mumtaj Begam, and I. Elanvazuthi. Voice recognition algorithms using mel frequency cepstral coefficient (mfcc) and dynamic time warping (dtw) techniques, 2010.


A Appendix

A.1 X-vector model Figures

We provide here the evaluation figures of x-vector model. The figures discussion is within the main paper body.

Figure 10: Number of false predictions (successful attacks) of the xvector ASI model using two-tube structure. The dotted line shows the average number of false predictions per true speaker.

A.2 User study utterances.

We use the first 50 utterances of the arctic dataset[^8] for our live experiment recordings. Here is the list of utterances.

1. "Author of the danger trail, Philip Steels, etc."
2. "Not at this particular case, Tom, apologized Whittemore."
3. "For the twentieth time that evening the two men shook hands."

[^8]: http://www.festvox.org/cmu_arctic/
Figure 11: False predictions histogram using a single tube on x-vector model. Low $f_0$ and high $Q_0$ values result in higher number of false predictions.

4. "Lord, but I’m glad to see you again, Phil."
5. "Will we ever forget it."
6. "God bless ’em, I hope I’ll go on seeing them forever."
7. "And you always want to see it in the superlative degree."
8. "Gad, your letter came just in time."
9. "He turned sharply, and faced Gregson across the table."
10. "I’m playing a single hand in what looks like a losing game."
11. "If I ever needed a fighter in my life I need one now."
12. "Gregson shoved back his chair and rose to his feet."
13. "He was a head shorter than his companion, of almost delicate physique."
14. "Now you’re coming down to business, Phil, he exclaimed."
15. "From that moment his friendship for Belize turns to hatred and jealousy."
16. "There was a change now."
17. "I followed the line of the proposed railroad, looking for chances."
18. "Clubs and balls and cities grew to be only memories."
19. "It fairly clubbed me into recognizing it."
20. "It was my reports from the north which induced people to buy."
21. "I was about to do this when cooler judgment prevailed."
22. "To my surprise he began to show actual enthusiasm in my favor."
23. "Their forces were already moving into the north country."
24. "I had faith in them."
25. "They were three hundred yards apart."
26. "Since then some mysterious force has been fighting us at every step."
<table>
<thead>
<tr>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 12: X-vector model predictions confusion matrix split by the true and predicted speakers' sex. Cross-sec submatrix is sparse, indicating attack is more successful within same-sex speakers.

27. "He unfolded a long typewritten letter, and handed it to Gregson."
28. "Men of Selden's stamp don't stop at women and children."
29. "He stopped, and Philip nodded at the horrified question in his eyes."
30. "She turned in at the hotel."
31. "I was the only one who remained sitting."
32. "Yes, it was a man who asked, a stranger."
33. "We'll have to watch our chances."
34. "The ship should be in within a week or ten days."
35. "I suppose you wonder why she is coming up here."
36. "Meanwhile I'll go out to breathe a spell."
37. "It seemed nearer to him since he had seen and talked with Gregson."
38. "Her own betrayal of herself was like tonic to Philip."
39. "He moved away as quietly as he had come."
40. "The girl faced him, her eyes shining with sudden fear."
41. "Gregson was asleep when he re-entered the cabin."

B Differential evolution

We ran differential evolution with best2exp strategy, population size of 100, maximum of 5 iterations and tolerance of 0.001. Results for individual utterances are shown in Figures 14 and 15.
Figure 13: Number of successful attacks (false predictions) of the x-vector ASI model on the user study participants recordings. The dotted line shows the average number of successful attacks per true speaker.

Figure 14: Search performance over 1–4 different utterances.

Figure 15: Search performance over 5–8 different utterances.
Algorithm 1 Differential evolution

**Input:** $x$, $y$, pool size $N$, attack budget $n$, fitness function $f$, crossover parameter $c$, maximum iterations $i$, mutation proportion $m$

$A : N \times n = \text{random(pool)}$

for $i = 0$ to $i$

$A_{\text{new}} : N \times n = 0.0$

for $j = 0$ in $N$

$r1 = \text{sample-randomly}(A)$

$r2 = \text{sample-randomly}(A)$

$t = A_{\text{best}} + m \times (r1 - r2)$

$m = c > \text{random-mask-of-size}(n)$

$a = t \ast m + A_j \ast (1 - m)$

if $f(a) > f(A_j)$ then

$A_{\text{new},j} = a$

else

$A_{\text{new},j} = A_j$

end if

end for

$A = A_{\text{new}}$

end for