Tubes Among Us: Analog Attack on Automatic Speaker Identification

Shimaa Ahmed and Yash Wani, *University of Wisconsin-Madison;* Ali Shahin Shamsabadi, *Alan Turing Institute;* Mohammad Yaghini, *University of Toronto and Vector Institute;* Ilia Shumailov, *Vector Institute and University of Oxford;* Nicolas Papernot, *University of Toronto and Vector Institute;* Kassem Fawaz, *University of Wisconsin-Madison*

https://www.usenix.org/conference/usenixsecurity23/presentation/ahmed-shimaa

This paper is included in the Proceedings of the 32nd USENIX Security Symposium.

August 9–11, 2023 • Anaheim, CA, USA

978-1-939133-37-3

Open access to the Proceedings of the 32nd USENIX Security Symposium is sponsored by USENIX.

Tubes Among Us: Analog Attack on Automatic Speaker Identification

Shimaa Ahmed[*1], Yash Wani[1], Ali Shahin Shamsabadi[†2], Mohammad Yaghini[3,4], Ilia Shumailov[4,5], Nicolas Papernot[3,4], and Kassem Fawaz[1]

[1]*University of Wisconsin-Madison*, [2]*Alan Turing Institute*, [3]*University of Toronto*, [4]*Vector Institute*, [5]*University of Oxford*

Abstract

Recent years have seen a surge in the popularity of acoustics-enabled personal devices powered by machine learning. Yet, machine learning has proven to be vulnerable to adversarial examples. A large number of modern systems protect themselves against such attacks by targeting 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 detection, 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 automatic speaker identification systems (ASI) that pack ML-based models to authenticate users based on their voiceprint [37, 51].

Speaker identification systems are vulnerable to an array of attacks such as speech synthesis [56, 59, 63], voice conversion [34, 48, 70], replay attacks [28], and adversarial examples [13, 19, 29]. 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 in the vision domain follow the same principle. 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.

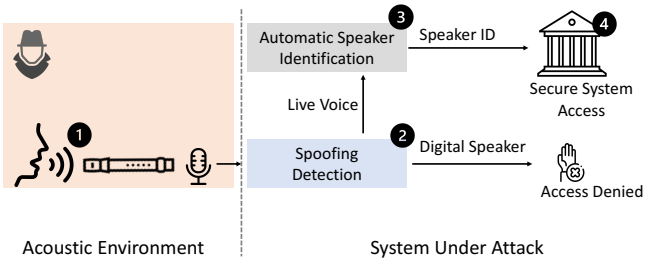


Figure 1: Overview of Mystique voice impersonation attack. Left: Acoustic environment fall's under the adversary's control. Right: the system under attack setup. ❶ The adversary speaks through an adversarially designed tube. ❷ A liveness detection model confirms the liveness of the captured voice. ❸ An automatic speaker identification model recognises the identity of the adversary as the target speaker. ❹ The secure system gives access to the adversary.

Current defenses leverage this observation and employ mechanisms to detect the digital attack artifacts in the input signal [46, 60, 62]. These defenses target either the (1) physical properties of the speaker *e.g.* their physical presence [39, 71] or (2) properties of the speech speakers produce *e.g.* the energy distribution of different harmonics [12, 20]. The resulting unified acoustic pipeline constrains the attacker when generating the attack samples, thus increasing the cost of the attack [32, 46, 62]. Generally speaking, the defense literature makes a basic assumption that *the attack source is not human*. In this paper, we challenge it by asking this question: *Is it possible to attack speaker identification systems using analog manipulation of the speech signal?*

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 iden-

[*]Corresponding Author: ahmed27@wisc.edu

[†]Work done partially while the author was at the Vector Institute.

tification 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 four 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 be a professional vocalist or have any impersonation experience. Third, the transform should directly impact the ASI model prediction. Fourth, the transform can be mathematically modeled to be incorporated in the attack optimization objective. Mystique exploits the acoustic resonance phenomenon to satisfy these conditions. Acoustic resonance is a physical transform where objects vibrate to specific frequencies. Acoustic resonance allows an object to act as a frequency 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, 58].

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 to find the optimal tube dimensions for a given target. 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 five spoofing detection baselines. We validate Mystique on standard speaker identification dataset, VoxCeleb, and on live speech by conducting a user study of 14 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 bypasses current acoustic defenses based on liveness and (presumably uniquely) identifying characteristics of the speaker, such as voice pitch.
- We demonstrate that, using commonly available plastic tubes, an attacker can change the properties of their speech in a systematic way and manipulate ML models. For example, an adversary can impersonate 500 other speakers using tubes. Moreover, Mystique is only 23% detectable by the best ASVspoof 2021 spoofing detection baseline that has 100% accuracy on classifying natural (*i.e.*, no tube) recordings as live.
- We run our attack on live speech to confirm its practicality. We perform a user study and show that the attack is successful over-the-air on live speech with 61.61% success rate. We conduct a human impersonation study as a baseline and find that its success rate is only 6.2%.
- We discuss a set of strategies to detect the attack and add a discussion of limitations and future work.

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 a signal that contains specific frequency components [22]. 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 f_i. A resonating object acts as a *filter* that magnifies the resonance frequencies, and filters out other frequencies in the excitation signal. The resonance vibrations encounter resistance and losses that define the filter sharpness—referred to as the quality factor Q. The filter's f_0 and Q are usually well 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, 22]. 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 $f_0 = c_{air}/\lambda$ (Hz), where c_{air} is the speed of sound in air, and λ 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 [38]. Thus, $f_0 = c_{air}/2L$, and $c_{air} = 20.05\sqrt{T}$ (m/s) in dry air [22], where T (°K) is the thermodynamic temperature. These equations, however, do not consider the tube diameter and air humidity. A more accurate equation is:

$$f_0 = \frac{c_{air}}{2(L+0.8d)}, \tag{1}$$

where d is the tube diameter, and $\Delta L = 0.8d$ is an empirical term derived from measurements [7].

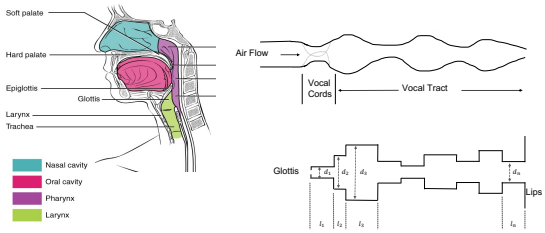
Quality Factor. The quality factor quantifies the acoustic losses inside the tube. There are two main sources of losses [22, 33]: radiation loss and wall loss. The radiation loss d_{rad} is the energy loss due to acoustic radiation outside the tube [22]: $d_{rad} = 2\pi A f_0^2 / c_{air}^2$, where A 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 [22]: $d_{wall} = \sqrt{\mu/\rho A f_0}$, where $\mu = 1.81 * 10^{-5} kg/ms$ is the air viscosity, and $\rho = 1.18 kg/m^3$ 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 [27]. Thus, the tube quality factor can be approximated by:

$$Q_0 = 1/(d_{rad} + d_{wall}). \quad (2)$$

2.2 Human Speech Modeling

Biological Characteristics. Humans generate speech using three main structures [52]: 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* [52]. 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* [52], and human speech production is 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 $G(f) = \mathcal{F}(g(t))$. The vocal tract can be modeled as a variable acoustic resonator $H_v(f)$ that filters the glottal excitation into speech $s(t) = \mathcal{F}^{-1}(H_v(f) \cdot G(f))$. 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. The different parts of the vocal tract are modeled as consecutive tubes [16], 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.



(a) Vocal Tract Structure. (b) Vocal Tract Model.

Figure 2: The vocal tract structure and model. (a) The structure including the glottis, the pharynx, the oral cavity, the nasal cavity, and the lips—adapted from AnatomyTool [1]. (b) Vocal tract parts modeled as consecutive tubes of different diameters.

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 [50], 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 is more secure against replay attacks, and more usable as it can be embedded within other tasks such as speech recognition in a seamless interaction.

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 by deep neural networks. These models capture the speaker's voice characteristics from a variable-length speech utterance $s(t)$ and map it to a vector (embedding) in a fixed-dimensional space. X-vector DNN [50, 51] 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 collects a speech utterance from the user, 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, i.e., spoofed. If the utterance is detected to be live, the spoofing detector feeds it to the ASI model which classifies the speaker identity and

grants the user access to the secure system. This system setup can be deployed for logical access applications such as phone banking services, voice assistants, and smart home devices.

Threat Model. We consider an adversary that wants to attack the ASI model to be identified as a target user. First, 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 by changing their *live* voice using physical objects. Our work introduces a systematic reproducible technique that allows the adversary to impersonate an arbitrary speaker's voice, in the eyes of the ASI model, without using a digital speaker. The attack is analog and only allows for the use of physical objects and natural sounds.

Second, the adversary performs an audio-only interaction with the system. Hence, they have complete control over the recording environment, as shown in Fig. 1. They have no access to the ASI model internals; *i.e.*, a black-box attack. The adversary can only query the ASI model on inputs of their choice and get the model's output scores and label. As such, the adversary needs no recordings of the victim's speech. They only know the victim is enrolled in the ASI model. Finally, the adversary impersonates the victim in the eyes of the ASI model to gain access to their protected accounts. The attack does not target human listeners explicitly.

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, validates the voice liveness, 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 speaks through a specifically designed tube to induce a targeted misclassification at the ASI system, effectively impersonating a target victim.

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 classifying the voice as 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:

$$\min_{p} \quad \mathrm{R}(\mathsf{ASI}(s'), y_t) \quad \text{s.t.} \quad s' = \mathrm{F}_{\text{tube}}(s, p), \tag{3}$$

where s is the original speech sample, p is the tube parametrization, y_t is the attack target label, R is the loss, $\mathrm{F}_{\text{tube}}(.)$ is the mathematical model of the tube, and $\mathsf{ASI}(.)$ 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 bypass spoofing detection and attack ASI with an analog attack.

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 Δf_i is defined by the quality factor Q_i (Eqn. (5)), where $i = 1, 2, \cdots, \lfloor f_s/f_0 \rfloor$ 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:

$$s_{out}(t) = \mathrm{F}_{\text{tube}}(s_{in}, p) = \mathcal{F}^{-1}(H_{res}(f) \cdot S_{in}(f)), \tag{4}$$

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, Eqn. 1 and 2 quantify the fundamental resonance parameters. The full harmonic range of f_i and Q_i are:

$$f_i = i \cdot f_0 = \frac{i \cdot c_{air}}{2(L + 0.8d)}; \quad Q_i = Q_0 / \sqrt[4]{i}, \tag{5}$$

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 $1/\sqrt[4]{i}$ 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 Sec. 6.1.

Resonance inside connected open-ended tubes happens when the acoustic impedance between the connected tubes equal an open-end impedance [53]. This condition is mapped to the following equation for each two tubes intersection:

$$A_1 \cdot \cot(2\pi f L_1/c_{air}) = A_2 \cdot \cot(2\pi f L_2/c_{air}), \qquad (6)$$

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 resonance model by measuring real tubes resonance and comparing it to our BPFs model. First, we excite the tube with a 3-second chirp signal [49] that exponentially spans the frequency range from 100 to 3700 Hz. Then, we play speech samples from VoxCeleb dataset and measure the similarity between tube and BPF output signals. We use the setup in Fig. 5 for recording.

Fig. 3 shows the Fast Fourier Transform (FFT), waveform, and cross-correlation plots for a tube of $L = 40.6$, $d = 3.45$ cm. Fig. 3a shows the FFT of the chirp output, which is effectively the tube's transfer function $H_{res}(f)$. The vertical dotted lines indicate the theoretical resonance frequencies, f_i, which align perfectly with the measurement. Fig. 3c shows the waveforms with the dynamic time warping (DTW) alignment, and Fig. 3d shows that the waveforms are highly correlated. We also measure the DTW alignment distance for a set of 6 tubes (Table 1), which is a measure of similarity. The distances are 0.027, 0.03, 0.025, 0.023, and 0.021. Thus, the tube and BPF waveforms are very similar for all evaluated tubes. Therefore, the BPF model is a realistic representation of the tube resonance. The attacker uses this model to obtain the tube parameters for a targeted attack.

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 [18, 23, 40]. Usually, training data has to be augmented with simulated environmental effects such as noise and echo [18, 23, 40]. 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 [4, 42]. 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 Appendix A.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 [43, 55]. In particular, ASI is sensitive to the pitch of the speech signal; therefore, applying the tube is expected to alter the classification.

4.3.1 Tubes Cause Pitch Shifts

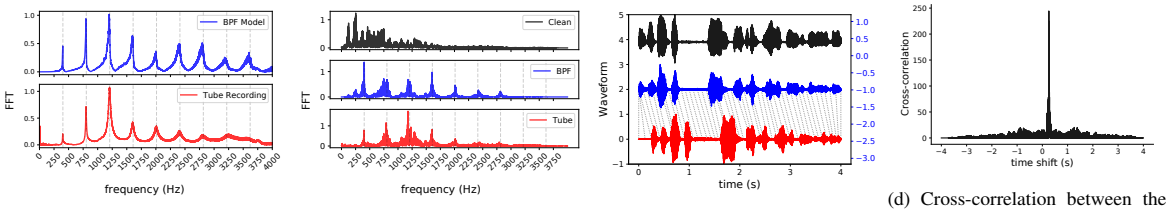
We build on the work of McAulay and Quatieri [31] who frame the pitch estimation as the solution of an unconstrained optimization of the mean square error between the Short-time Fourier transform (STFT) of a signal $s(t)$ and a sum of harmonics, parameterized by the pitch. In Appendix A.1, we show that the resonance effect of the tube translates to a constrained version of the same optimization problem. We then argue that given the smaller feasibility set of the constrained problem, its solution will inherently have filtered out frequencies. As a result, the estimated pitch will be different.

Validation. 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 [21] 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.

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 1.77×10^{-26} and 2.99×10^{-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 the tubes by the quality factor Q_0 and the fundamental frequency f_0. Although, for a single-tube configuration, the search space is small enough to be bruteforced within a few minutes, 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* (DE) [54]. Algorithm 1 describes our DE approach with *best2exp* strategy. The algorithm performs the tube parameters A search by picking three data samples from



(a) FFT of a chirp (b) FFT of a speech sample (c) Speech waveform (d) Cross-correlation between the tube and BPF waveforms

Figure 3: Resonance model validation of Tube 1 ($L = 40.6, d = 3.45$) vs its BPF model: (a) FFT of chirp, (b) FFT of a speech utterance, (c) speech wavefroms showing DTW alignment between tube and BPF signals, (d) cross-correlation between tube and BPF waveforms.

Algorithm 1 Differential Evolution

Input: s, y_t, pool size N, attack budget n, fitness function f, crossover parameter c, maximum iterations it, mutation proportion m
$A : N \times n = \text{random(pool)}$
for $i = 0$ **to** it **do**
  $A_{new} : N \times n = 0.0$
  for $j = 0$ **in** N **do**
    $r_1, r_2 = \text{sample-randomly}(A)$
    $l = A_{best} + m \times (r_1 - r_2)$
    $m = c > \text{random-mask-of-size}(n)$
    $a = l * m + A_j * (1 - m)$
    if $f(a, s, y_t) > f(A_j, s, y_t)$ **then**
      $A_{new,j} = a$
    else
      $A_{new,j} = A_j$
    end if
  end for
  $A = A_{new}$
end for

an underlying population and combining the best performing one with the difference between the other two. The algorithm is called differential, since the update step includes computing the difference between a pair of samples and stochastically appending it to the third. In the search algorithm, we set boundary conditions on the tube dimensions which defines the underlying population. We define the boundaries as: f_0 ranges from 50 Hz to 1 kHz, and its Q_0 ranges from 5 to 100, such that f_0 falls in the typical range of human voice pitch. We sample from this range using step size of 10 Hz for f_0 and 5 for Q_0. According to Eqn. (1) and Eqn. (2), the single tube length would range 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. The resultant f_i's are found from Eqn. (6). We set the population size $N = 100$, maximum iterations $it = 5$, and tolerance of 0.001. The attack is performed in a black-box fashion, requiring only the target class score of the ASI model. Thus, the fitness function $f(A, s, y_t)$ is the ASI model's score of the target label y_t when transformation A is applied on the user's utterance s. We find that within 100

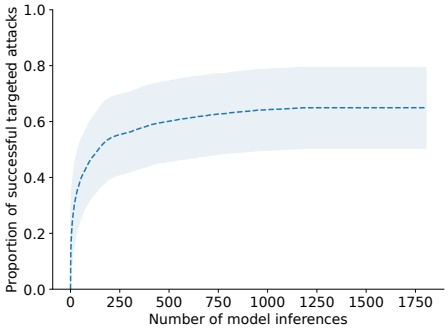


Figure 4: Average reachable target search performance across all of the participants with SpeechBrain model

model invocations, as is demonstrated in Fig. 4, we could find $46\% \pm 12$ of all possible reachable targets, whereas at 250 invocations it grows to $55\% \pm 14$. Despite relatively low performance, our DE algorithm enables the attacker to within minutes check with a reasonable probability if a user's utterance s can be transformed to impersonate a target y_t.

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 instantiate Mystique on a standard dataset, VoxCeleb, using the resonance filter model. We show that Mystique can successfully attack two ASI models. Using Mystique, each adversarial speaker successfully impersonates 500 targeted victims, on average. (Sec. 6.1)

Q2. Does Mystique's impersonation succeed in real-world? We build a physical recording setup and run Mystique over-the-air on VoxCeleb (Sec. 6.2). We also conduct a user study and evaluate Mystique on live speech. We show that Mystique's attack success rate over-the-air is 61% on a standard dataset and 61.61% on live speech. We also compare Mystique against human imperson-

ation as a baseline and find that most participants were not able to reliably impersonate a target speaker with a success rate of 6.2% on average. Finally, we show that Mystique is consistent over multiple trials.

Q3. How can a defender detect Mystique? We study different strategies to detect Mystique. We show that while the Mystique-generated and victim voiceprints are similar, the ASI model is less confident under Mystique. Further, we show that a human can discern samples generated from Mystique. Finally, we find that Mystique is successful against baseline spoofing detection, but not against a detector trained on Mystique's samples.

5.1 Datasets and ML Models

ASI Models. We evaluate two state-of-the-art ASI models: (1) the x-vector network [51] implemented by Shamsabadi et al. [45], and (2) the emphasized channel attention, propagation and aggregation time delay neural network (ECAPA-TDNN) [17], implemented by SpeechBrain.[1] Both models were trained on VoxCeleb dataset [15, 36, 37], a benchmark dataset for ASI. 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%.

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.

Spoofing Detection Models. We evaluate two spoofing detection techniques, (1) ASVspoof baselines and (2) Void. We consider two state-of-the-art baselines from the ASVspoof 2021 challenge[2] for physical access (PA) and logical access (LA) tasks. The PA task objective is to discriminate between live-human speech and replayed recordings via loud speakers, while the LA task objective is to differentiate between live speech and artificially generated speech using text-to-speech, voice conversion, or hybrid algorithms. The 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. Thus, the PA and LA tasks are designed to distinguish two different features of spoofed speech: loud speakers artifacts, and synthetic speech artifacts. We use the official implementation[3] employing the light CNN (LCNN) model [61]. However, each is trained on a task specific dataset from ASVspoof 2019 challenge: *bonafide* and *replayed* samples for the PA-LCNN model, and *bonafide* and *synthetic* samples for the LA-LCNN model. The second spoofing detection technique is Void (Voice Liveness Detection) [2], a recent high-performing system that uses spectral analysis to detect synthetic speech. It extracts 97 spectral features to train an SVM model. The key assumption is that live speech power is higher at low frequencies than at high frequencies, while the synthetic speech power is linearly spread out across the frequency range. This makes Void a good candidate for detecting Mystique since the resonance effect redistributes the speech power and amplifies the power at f_0 and its harmonics f_i as shown in Fig. 3 and Sec. A.1. We use Wenger et al.'s implementation [64], where they train three models on the ASVspoof dataset: (1) SVM, (2) Light CNN [26], and (3) a custom 5-layer CNN.

Live Human Impersonation. We conduct a user study to test Mystique on live speech, involving three stages.

– In the first stage, each participant records the first 50 utterances of the arctic dataset[4] using a microphone, without a tube. Since ASI is a text-independent task, we did not place any requirements or assumptions on the utterances' linguistic content. The use of the arctic dataset is an arbitrary choice. We then apply Mystique on these recordings to impersonate victims enrolled in the ASI models—speakers from VoxCeleb.

– In the second stage, we validate Mystique's success rate by conducting the attack over-the-air. We select three representative tubes that are common between the impersonation attacks of all participants. We ask each participant to speak each utterance through each tube and compare the live classification result to the one obtained from the filter. We ask the participants to maintain the same speaking style and not to press their lips against the tube opening as it creates non-linear transformations not captured by Mystique's model.

– In the third stage, we ask the participants to impersonate from 1 to 8 target speakers, based on their capacity. We select the targets from the successful impersonations using Mystique. Each participant watches videos of the target (celebrity) speaker till they feel confident about impersonating them, which took from 5 to 20 minutes each. Then, the participant is allowed five attempts to impersonate the target using their own words; i.e. they were not given a specific script to read.

We recruited 14 individuals[5] (7 males, 7 females, age:18-30). 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.

[1] SpeechBrain (https://github.com/speechbrain/speechbrain/) is an open-source state-of-the-art toolkit on Hugging Face

[2] https://www.asvspoof.org

[3] https://github.com/asvspoof-challenge/2021

[4] http://www.festvox.org/cmu_arctic/

[5] Two participants abstained from conducting the third stage of the study.

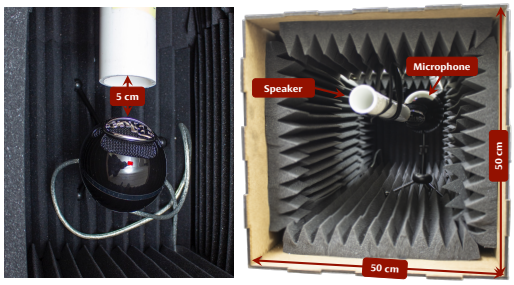


Figure 5: The recording setup: top view (left) and front view (right).

5.2 Physical Setup for the Attack

We design and implement a 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 tubes in this work. We conduct the single-tube experiments using six PolyVinyl Chloride (PVC) pipes purchased from a hardware store. Their dimensions are listed in Table 1. For the two-tube structures, we 3D printed the tubes for a fine-grained control over the tubes radii which impacts the resonance frequency (Eqn. 6). We used Formlab's Form 2[6] printer and Black Resin[7] material. We print the tubes with a 50 µm resolution for a smoother finish and a thickness of 2 mm, no support material was on the inside of the tube. The tubes are connected using High Density Fiberboard (HDF) rings at run time, for tube reusability, as shown in Fig. 15 in the appendix. We constructed three two-tube structures whose dimensions are in Table 2. For both sets, we select the tubes dimensions based on our observations from Sec. 6.1 experiment.

Recording Environment. We conducted the experiment in a $8 \times 3.6 \times 3.6$ m lab space. We built an audio chamber to prevent interference of the tube's input and output sounds, and 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,[8] 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-soundevice` library to automate the recordings[9].

[6] https://formlabs.com/3d-printers/form-2/
[7] https://formlabs.com/store/black-resin/
[8] https://www.bluemic.com/en-us/products/snowball/
[9] https://python-sounddevice.readthedocs.io/en/0.4.4/

6 Mystique's Evaluation Results

We conduct the following experiments to answer the three questions from Sec. 5 in detail.

6.1 Impersonation Attack at Scale

First, we test Mystique's impersonation attack feasibility on the full test set to address the first evaluation question. We run Mystique on the VoxCeleb (91 speakers) test set, representing the adversarial speakers, and find the range of successful impersonation attacks and the corresponding set of adversarial tubes. In this experiment, we consider structures of N-tubes, where N ≤ 2. Hence, the resonating frequencies depends on three parameters (degrees of freedom): the tubes lengths L_1, L_2 and the tubes cross-sectional area ratio: $ratio_A = (d_2/d_1)^2$.

For each adversarial speaker, Mystique attempts to impersonate every enrolled speaker in the ASI model; 7205 in SpeechBrain and 250 in X-Vector. Mystique searches for the BPF filters parameters that trick the ASI into identifying the adversarial utterance $\mathrm{F}_{\text{tube}}(s,p)$ as the target victim speaker, y_t, using the DE algorithm 1.

Fig. 6 shows the number of target ids from SpeechBrain that an attacker could impersonate using Mystique *theoretically*; i.e., the number of target ids where Mystique successfully finds a tube configuration that fools the model for each attacker. Since real-world requirements constrain the search and Mystique has little degrees of freedom, the algorithm might not find a tube for each source-target pair. Fig. 13 in appendix shows the same for the x-vector model. As the figure shows, by optimizing the tube dimensions, Mystique can successfully impersonate a wide range of victim speakers. Specifically, a speaker can impersonate 500 (out of 7205) target speakers on average on SpeechBrain model and 137 (out of 250) 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 ASI models. Next, we analyze the adversarial tube (BPF) parameters and the demographic distribution of the predictions to interpret how the attack works. We report three findings.

First, the attack is most effective when f_0 lies in the frequency range $f_0 \leq 400$ Hz with a high quality factor $Q_0 \geq 50$ as shown in Fig. 11. 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 [30]. Hence, this range of f_0 will have a stronger impact on the pitch, the significant spectrum, and the model prediction. Also, a high quality factor means a sharper filter; fine-grained selection.

Second, Mystique is 80% more successful on impersonating same-sex targets than cross-sex. Fig. 12 in appendix shows

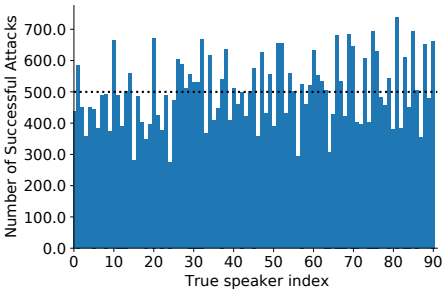


Figure 6: Successful impersonation attacks (out of 7205) on SpeechBrain model for each adversarial speaker from VoxCeleb. Dotted line shows the average number of successful attacks per speaker.

the prediction confusion matrix split by the attacker-victim speakers sex. The figures show that the cross-sex speakers submatrix is sparser than that of the same-sex.

Third, we find that Mystique impersonates different victims when optimizing for different utterances of the same speaker (attacker). Hence, the attack is not utterance (text) independent. We attribute this observation to two reasons: (1) ASI models are not perfect in separating the linguistic content and voice biometrics; the model prediction varies with the spoken utterance, (2) the attack's pitch shift and voice transformation is the resultant of Mystique's transformation applied on the spoken utterance original spectral content.

6.2 Over-the-air Attack

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, listed in Table 1, which have f_0, Q_0 in the most significant range—Fig. 11. We purchase them from the hardware store. While for the two-tube setting, we build three structures of 3D printed tubes as described in Sec. 5.2; their parameters are listed in Table 2.

We use the Pixel phone to simulate the speaker and play the VoxCeleb utterances over-the-air for all tubes. 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. We repeat the recordings 6 times to account for any environmental variations and to evaluate the attack reliability and consistency.

Single-Tube. Table 1 shows the number of successful attacks (impersonated targets) per tube and compares it to 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, depending on the attacker's spoken utterance. As discussed in Sec. 6.1, we found that different utterances sometimes lead to different impersonated victims per attacker-tube pair. Second, "Filter" columns (7 and 10) show the number of successful attacks using each tube's BPF model. The filter's successful attacks are on the same magnitude as the real tube. 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, and the linear BPF model (Eqn. 4, 5) is a reasonable approximation of the resonance effect. A more accurate model is to use wave simulation engines at the expense of increased computation complexity.

Finally, we assess the attack's reliability over multiple trials. We measure the model's predictions consistency rate—defined as the percentage of consistent predictions across six runs. Table 6 in appendix shows the consistency rate per tube, on average 84% of the predictions are consistent over six runs.

Two-Tube. Similarly, Table 2 shows Mystique's performance over-the-air using the two-tube configurations. The success rate is the percentage of matched successful attack between Filter and Real tubes impersonated identities. Mystique's targeted attack succeeds more than 50% of the time.

6.2.2 Live Impersonation Attack

We run Mystique on 14 participants natural recordings, 50 utterances each, and find the set of theoretically successful attacks (impersonated identities) per participant. Fig. 7 shows the number of successful attacks on the SpeechBrain model. Fig. 14 in the Appendix shows the same for the 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 evaluate the recordings on the ASI models and compare them to the BPF predictions. Table 3 reports the percentage of Mystique's BPF impersonation attacks that also succeeded over-the-air in the live recording of each participant. The average success rate ranges from

Tube	Tube Dimensions		Resonance Parameters		X-Vector False Predictions			SpeechBrain False Predictions		
	L (cm)	d (cm)	f_0 (Hz)	Q_0	Real	Filter	Match	Real	Filter	Match
1	40.6	3.45	402.16	58	158	141	64 (45.40%)	158	238	111 (46.64%)
2	61.3	4	270.70	68	123	194	75 (38.66%)	134	255	106 (41.57%)
3	87	5.2	191.48	77	202	242	101 (41.74%)	198	308	141 (45.8%)
4	99.4	3.45	170.89	64	325	174	77 (44.25%)	220	200	121 (60.5%)
5	120.3	5.2	140.20	79	190	167	95 (56.89%)	210	351	146 (41.6%)
6	154	5.2	110.36	76	176	108	63 (58.34%)	179	185	114 (61.62%)

Table 1: Evaluation of Mystique over-the-air for 40 speakers × 20 utterances: 800 total inferences. *Real*: # successful attacks of the real tube, *Filter*: # successful attacks of the corresponding filter model, *Match*: the number (percentage) of matched attacks between filter and real tube.

Tube	Tube Parameters (cm)				f_0 (Hz)	Attack Success Rate
	L_1	d_1	L_2	d_2		
7	9.53	2.1	10	1	853.1	66.6%
8	11.44	0.98	8.9	3.4	901.55	50%
9	14.53	2.1	10	1	600.4	100%

Table 2: Two-tube structures f_0 and attack success rate over-the-air.

ID	Gender	Tube3	Tube4	Tube6	Avg	Avg_{Cal}	Human
0	F	50.0	50.0	66.67	55.56	**65.0**	0/5
1	M	58.82	81.82	57.14	65.93	43.02	0/40
2	M	66.67	72.73	77.78	72.40	**72.58**	0/20
3	F	63.64	83.33	75.0	73.99	**78.7**	0/20
4	F	66.67	58.33	71.43	65.48	**73.15**	0/20
5	M	50.0	42.86	55.56	49.47	42.29	1/20 (5%)
6	M	46.15	54.55	80.0	60.23	**60.71**	6/25 (24%)
7	F	66.67	77.78	80.0	74.81	62.22	–
8	M	43.75	42.86	54.55	47.05	**52.06**	0/10
9	M	50.0	60.0	50.0	53.33	41.6	1/10 (10%)
10	F	66.67	62.5	80.0	69.72	**75.0**	5/20 (25%)
11	M	50.0	61.54	72.73	61.42	**69.17**	–
12	M	10.0	54.55	40.0	34.84	**35.56**	0/15
13	F	80	71.42	83.33	78.25	69.44	1/20 (5%)

Table 3: User study participants percentage (%) of successful over-the-air impersonation attacks with and without Mystique. **bold** values are enhanced by personalized calibration.

34.84% to 78.25%, showing that Mystique reliably launches over-the-air attacks. This result is significant—live human speech varies between recording sessions unlike *e.g.* VoxCeleb experiment with fixed recordings.

Moreover, we explore Mystique's personalization by fine-tuning the filter parameters to each participant's voice characteristics. Applying a voice envelope calibration to the filter gain increases Mystique's success rate, for most participants, up to 10%. However, it drops for a few participants, as shown in column 6 of Table 3. Thus, personalization is one way to further optimize Mystique, which we leave to future work. Additionally, we observe the same skew in the speaker's sex for successful attacks as in VoxCeleb (Fig. 12), where the cross-sex submatrix is sparse.

Finally, we evaluate the participants impersonation capabilities without using any tubes as a baseline for Mystique's performance. The last column in Table 3 shows the number

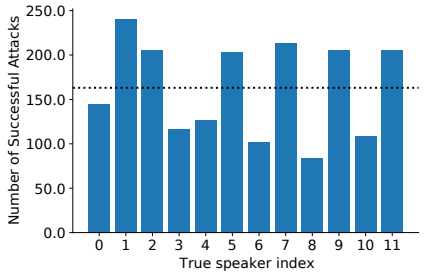


Figure 7: Number of successful attacks of the study participants recordings on SpeechBrain. The dotted line shows the average number per true speaker.

of times the participant was able to impersonate a target by the total number of trials, where for each target the participant performs 5 impersonation trials. Note that some participants were not willing to impersonate more than one target, thus the total number of trials is not the same for all of them. This study shows that most participants were not able to reliably impersonate a target speaker, where the average success rate is only 6.22%. Specifically, 7 participants did not succeed in any trials, 3 participants were able to impersonate one target one time and failed at the 4 other attempts for the same target, and only 2 participants could succeed more than once. We noticed they could capture the accent and pitch of the target. Participant 6 impersonated 3 (out of 5) targets for (3, 2, 1) trials for the same target, while participant 10 successfully impersonated 2 targets for (2, 3) times. Yet, Mystique significantly outperforms the strongest baseline; it impersonates 100+ victims with success rate of up to 78.7%.

6.3 Mystique's Robustness

We study different strategies to detect samples from Mystique, which include: comparing prediction confidence, human-based analysis, and state-of-the-art spoofing detection.

6.3.1 What is the ASI model confidence on Mystique?

The ASI model outputs the class (speaker id) with the highest prediction score. Here, we analyze the model's confidence of the predicted class, using the softmax score as a proxy.

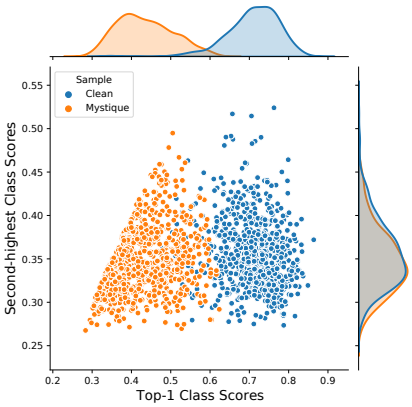


Figure 8: The distribution of BrainSpeech softmax scores for the top two classes on VoxCeleb clean and adversarial samples.

Fig. 8 shows the distribution of the model's top two classes confidence scores in case of clean (benign) and Mystique (adversarial) samples. The figure shows that the model is less confident of its top-1 class prediction on Mystique's samples; i.e., the gap between the top-2 scores decreases. This finding arises from Mystique's samples being out-of-distribution (OOD) samples with respect to the model's training data. Hence, this analysis suggests that Mystique's threat can be weakened if the ASI model is trained to reject samples of which it is not highly confident [44].

6.3.2 How similar is Mystique to the victim's voice?

Our second detection strategy assesses Mystique's spoofed speech similarity to the victim's speech. We analyze the attack-victim speech pairs that are successful over-the-air. Specifically, we visualize the attack in the problem space (audio) and evaluate the similarity in the embedding space.

Fig. 10 in Appendix B shows a sample of the attacker and victim waveforms and spectral content. The samples are not visually similar and do not exhibit high cross correlation. We attribute this result to speech being composed of two entangled characteristics: the speaker's voice and the linguistic content [3, 70]. The attack-victim samples are of different linguistic content. Note that VoxCeleb is composed of Youtube recordings of celebrities, and there is no transcript or one-to-one mapping of the linguistic content among different speakers. However, the ASI models learn a representation of the speech utterances that are presumably based on voice characteristics and invariant to the linguistic content.

Next, we analyze the ASI model embedding space similarity of the attack-victim pairs. Table 4 shows that the embeddings of attack and victim sample pairs exhibit high cosine similarity, compared with randomly selected non-victim speakers. Thus, regardless of the linguistic content, Mystique applies a transformation on the speech utterance that maps its embedding (voiceprint) towards the victim's voiceprint.

Cosine Similarity	Tubes					
	1	2	3	4	5	6
Attack-Victim	0.39	0.38	0.40	0.43	0.38	0.34
Attack-Non victim	0.07	0.07	0.08	0.08	0.08	0.08

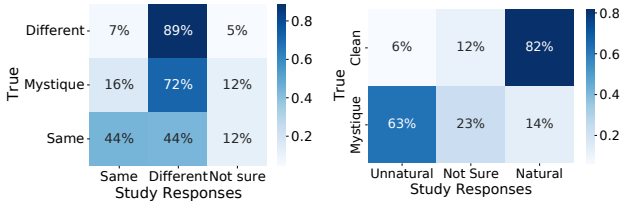
Table 4: Average cosine similarity score of the embeddings of Mystique's successful attack utterances and its victim speakers' utterances, compared to non-victim speakers similarity scores.

6.3.3 Does Mystique confuse humans as well?

Here, we investigate the similarity from humans point of view. Although Mystique is designed to attack ASI ML models by physically manipulating the spectral content of speech, we are curious whether it also confuses humans. To answer this question, we recruit participants to listen to two audio recordings and decide whether they belong to the same speaker, and also rate the audio quality as natural or unnatural. The study is approved by IRB and is conducted on the Prolific platform.

Study design. We recruited 151 participants, each compensated $1.4 for their effort, with an average completion time of 6 minutes. Each participant listens to 10 pairs of audio recordings from VoxCeleb; 3 pairs from each of the following cases: (a) the two recordings are clean and belong to the same speaker, (b) the two recordings are clean and belong to two different speakers of the same sex, and (c) one recording is generated by Mystique (attacker using a tube), while the other recording is of the corresponding victim's voice. The tenth pair is an attention check with two identical clean recordings. For each pair of recordings, we ask the participants two questions: (1) "*do they belong to the same speaker?*," and (2) "*how natural does the recording sound?*" on a 3-point Likert scale. We discard any responses that did not answer "same speaker" for the attention checker.

Results. Fig. 9 shows the distribution of responses. Fig. 9a shows that Mystique generated successful attacks on humans perception 16% of the time, and was able to confuse them 12% of the time. This result is interesting given that Mystique is not optimized to trick humans. The study also shows that the participants could distinguish different speakers voices with high probability (89%). However, they were confused on the "Same" speaker recordings. Here, 44% were labeled as same (different) speaker; *i.e.* not significantly better than random guessing. This result, supported by previous studies [25, 65], confirms that the voice identity perception can be challenging for humans, specially of unfamiliar speakers. Finally, Fig. 9b shows that 63% of Mystique's samples sound unnatural to the participants, yet 14% sound natural. These results show how ML models perceive speech differently than humans, creating the gap which Mystique and other attacks exploit.



(a) Similarity (b) Quality

Figure 9: The confusion matrix of the user study responses on the audio recording similarity and quality evaluation.

Model	Pre-trained		VoxCeleb		VoxCeleb$_{3\,tubes}$	
	EER	FAR$_0$	EER	FAR$_0$	EER	FAR$_0$
LA-LCNN	29.59	76.67	6.25	19.21	8.54	22.30
PA-LCNN	31.33	98.9	7.87	33.94	14.54	67.81
Void-SVM	62.12	97.70	36.70	91.64	48.3	96.12
Void-DNN	35.39	91.88	26.12	86.61	39.1	92.55
Void-LCNN	32.91	93	27.03	91	27.64	92.55

Table 5: Evaluation of five spoofing detectors on the user study recordings for: (1) pre-trained detectors, (2) fine-tuned on VoxCeleb clean and tube recordings, (3) fine-tuned on VoxCeleb **without** the three tubes used in the user study. Note: FAR$_0$ = FAR at FRR=0%.

6.3.4 Spoofing Detection

Finally, we assess Mystique's robustness against a set of five representative defenses and liveness detectors as detailed in Sec. 5.1. For the five evaluated models, we report the equal error rate (EER) which is the model's error rate when the false acceptance rate (FAR) and false rejection rate (FRR) are equal. A lower EER means a more accurate detector. We also report the FAR at 0% FRR, which sets the detector's operating threshold to correctly classify all live samples.

Table 5 reports the detectors performance on the recordings of 12 participants from the user study (Sec. 6.2.2). We label the no-tube speech as *bonafide* (live) and the tubes recordings as *spoofed*. The "Pretrained" column reports the performance of the pretrained models on ASVspoof dataset. All evaluated models are unable to distinguish the tube and no-tube samples. The best performing model (LA-LCNN) labels 76% of the tube samples as bonafide. Note that these models perform very well on the synthetic data from their original papers. For example, LA-LCNN's EER = 9.26% on the ASVspoof test set [68] and Void's EER < 12% on the curated synthetic dataset by Ahmed et al. [2] and Wenger et al. [64]. Yet, spoofing detectors do not generalize to Mystique's samples.

Next, we retrain the models on our VoxCeleb clean and tube recordings as the bonafide and spoofed samples respectively. The goal is to introduce Mystique's resonance effect to the model and label it as spoofed. The "VoxCeleb" column in Table 5 shows that the EER and FAR$_0$ dropped for all models, especially LA-LCNN with only 19% of the tube samples labeled as bonafide. Then, we retrain the models again on VoxCeleb but without the recordings from the tubes (3, 4, 6) that we use in the user study. The last column "VoxCeleb$_{3\,tubes}$" EER and FAR$_0$ values increase by 2% to 85% relative to the VoxCeleb column. When the exact tubes used by the participants in the test set are not part of the training data, the performance drops significantly. Thus these models overfit to their training distributions. Previous work [10, 35, 41] has reported the same observation; spoofing detectors hardly generalize to unseen transformations in the training data, which questions the security of voice-based authentication.

7 Discussion

Defenses. Having established a major vulnerability in spoofing detection systems leads to a question on how one stops such attacks. We show that defenses trained on samples of the resonance transform can detect Mystique and limit its effectiveness. 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. A better defense would have to incorporate properties of the medium and other modalities to rely on multiple factors, not just the speakers features.

Reproducibility. From formulating the original idea to completing the experiments of this paper, this work took around a year. 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 as reported in the 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 evaluation on the same day. Finally, when producing tubes with a 3D printer, the material on the inside of the tube should be smooth.

Limitations. Despite highlighting a flaw in current defenses design, there are a number of limitations in the current evaluation. First, we only considered simple tube structures, restricting the range of possible adversarial transformations. Second, we run the attack in a static recording environment, limiting its deployment in more practical situations where the adversary can be visually-observed or has partial control over the acoustic environment or experiences acoustic effects such as noise and interference. Third, we evaluated a small number of speakers and utterances, potentially underrating the over-

all attack performance. Fourth, the resonance effect sounds unnatural to humans, other transforms should be explored to have a more subtle impression on human listeners. Finally, Mystique needs access to the ASI model scores to perform the DE algorithm. However, Mystique can omit this requirement and perform an exhaustive search over all possible tube parameters, at the expense of the time complexity.

Future Work. We provide some directions to address Mystique's limitations. First, physical effects such as natural sounds and acoustic meta-materials should be explored to provide higher degrees of freedom and a less susceptible attack. Second, our evaluations suggest that the linguistic content can be optimized per each attacker-victim pair. Moreover, Table 3 suggests that attack personalization can boost its success rate. Third, Mystique can be made model-independent by performing the optimization on the estimated pitch as a proxy of the ASI model's decision as explained in Sec. 4.3, A.1.

8 Related Work

The literature on computer-based voice authentication is vast, and dates back to at least 1960s [20].

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 speech synthesis model on samples recorded from the victim speaker. The adversary uses this model to convert text into speech in the victim's voice [56, 59, 63]. Alternatively, voice conversion converts spoken utterances into the victim's voice [34, 48, 70]. In *replay* attacks, the adversary records the speaker's voice and replays the recorded speech [28]. Finally, many modern ML-based ASI models inherit the vulnerability to adversarial examples using standard gradient-based attacks [13, 19, 29].

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. [12] used a smartphone's magnetometer to detect the use of a loudspeaker. Blue et al. [9] tell electronic and human speakers apart by analyzing individual frequency components of a given speech sample. Yan et al. [69] calibrated individual speakers in the near field of the speakers to tell humans and electronic speakers apart.

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. [32] used an active radar to project a wave onto the face of the speaker and then detect shifts introduced to it from facial movement. Zhang et al. [71] analyzed hand movement to detect live speech by turning a smartphone into an active sonar.

Finally, there exists a class of defenses that restrict the attack surface by reducing attacker capabilities. Zhang et al. [72] used individual recordings from a stereo microphone to calculate time difference of arrival to detect replay attacks. Blue et al. [8] used two microphones to restrict the adversary to a 30 degree cone and protect against hidden and replay commands. Wang et al. [60] used correlates from a motion sensor to detect and reject hidden voice commands.

Physical Adversarial Examples. *Physical* adversarial examples are common in the vision domain, but have not been produced for acoustic tasks. Example adversarial objects include eyeware [14, 47], tshirts [66, 67], headwear [24, 73] and patches [57]. Although these objects were re-created in the real world, there is an important distinction here. These objects all apply perturbations that were initially designed for the digital space and then retrofitted with sophisticated machinery such at printers to realize them in the physical domain. Our attacks, on the other hand, directly restrict the search space of perturbations to those that can be easily realized physically. 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.

9 Conclusion

We demonstrate that a human adversary can reliably manipulate voice-based identification systems using physical tubes, *without access to the victim's speech*. Our attacks highlight acoustic intricacies that were largely ignored by prior literature, namely, the acoustic environment. 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. In this paper, we demonstrate that subjective nature of speech can be exploited to jeopardize the security of a critical system. Concretely, a fundamental question to consider in speaker identification is whether a person's identity can be accurately established despite the transformation of their voice.

Acknowledgment

This work was supported by DARPA (through the GARD program), the Wisconsin Alumni Research Foundation, the NSF through awards: CNS-1838733 and CNS-2003129, CIFAR (through a Canada CIFAR AI Chair), NSERC (under the Discovery Program and COHESA strategic research network), a gift from Intel, and a gift from NVIDIA. We also thank the Vector Institute's sponsors. Finally, we thank Bill Sethares, Andrew Allen, Nikunj Raghuvanshi, Hannes Gamper, and the reviewers for their fruitful discussions and recommendations.

References

[1] Anatomytool. "OpenStax AnatPhys fig.23.13 - The Esophagus - English labels" by OpenStax, license: CC BY. Source: book 'Anatomy and Physiology',https://openstax.org/details/books/anatomy-and-physiology.

[2] M. E. Ahmed, I.-Y. Kwak, J. H. Huh, I. Kim, T. Oh, and H. Kim. Void: A fast and light voice liveness detection system. In *29th {USENIX} Security Symposium ({USENIX} Security 20)*, pages 2685–2702, 2020.

[3] S. Ahmed, A. R. Chowdhury, K. Fawaz, and P. Ramanathan. Preech: A system for {Privacy-Preserving} speech transcription. In *29th USENIX Security Symposium (USENIX Security 20)*, pages 2703–2720, 2020.

[4] S. Ahmed, I. Shumailov, N. Papernot, and K. Fawaz. Towards more robust keyword spotting for voice assistants. In *31st USENIX Security Symposium*, 2022.

[5] A. M. Aljalal. Sound resonance in pipes with discrete fourier transform. *European Journal of Physics*, 36(5):055030, aug 2015.

[6] A. Allen and N. Raghuvanshi. Aerophones in flatland: Interactive wave simulation of wind instruments. *ACM Transactions on Graphics (TOG)*, 34(4):1–11, 2015.

[7] A. Bate. Lx.(i.) the end-corrections of an open organ flue-pipe; and (ii.) the acoustical conductance of orifices. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 10(65):617–632, 1930.

[8] L. Blue, H. Abdullah, L. Vargas, and P. Traynor. 2ma: Verifying voice commands via two microphone authentication. In *Proceedings of the 2018 on Asia Conference on Computer and Communications Security*, ASIACCS '18, page 89–100, New York, NY, USA, 2018. Association for Computing Machinery.

[9] L. Blue, L. Vargas, and P. Traynor. Hello, is it me you're looking for? differentiating between human and electronic speakers for voice interface security. In *Proceedings of the 11th ACM Conference on Security & Privacy in Wireless and Mobile Networks*, WiSec '18, page 123–133, New York, NY, USA, 2018. Association for Computing Machinery.

[10] S. Borzì, O. Giudice, F. Stanco, and D. Allegra. Is synthetic voice detection research going into the right direction? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 71–80, 2022.

[11] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[12] S. Chen, K. Ren, S. Piao, C. Wang, Q. Wang, J. Weng, L. Su, and A. Mohaisen. You can hear but you cannot steal: Defending against voice impersonation attacks on smartphones. In *2017 IEEE 37th International Conference on Distributed Computing Systems (ICDCS)*, pages 183–195, 2017.

[13] T. Chen, H. Luo, Y. Shen, F. Lin, and G. Xu. Spoofing speaker verification system by adversarial examples leveraging the generalized speaker difference. *Security and Communication Networks*, 2021:6664578, 2021.

[14] X. Chen, C. Liu, B. Li, K. Lu, and D. Song. Targeted backdoor attacks on deep learning systems using data poisoning, 2017.

[15] J. S. Chung, A. Nagrani, and A. Zisserman. Voxceleb2: Deep speaker recognition. In *INTERSPEECH*, 2018.

[16] J. R. Deller, J. H. L. Hansen, and J. G. Proakis. *Speech Production and Modeling*, pages 97–97. Wiley-IEEE Press, 2000.

[17] B. Desplanques, J. Thienpondt, and K. Demuynck. Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. *arXiv preprint arXiv:2005.07143*, 2020.

[18] H. Hu, T. Tan, and Y. Qian. Generative adversarial networks based data augmentation for noise robust speech recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5044–5048. IEEE, 2018.

[19] A. Kassis and U. Hengartner. Practical attacks on voice spoofing countermeasures, 2021.

[20] L. G. Kersta. Voiceprint identification. *The Journal of the Acoustical Society of America*, 34(5):725–725, 1962.

[21] J. W. Kim, J. Salamon, P. Li, and J. P. Bello. Crepe: A convolutional representation for pitch estimation. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 161–165, 2018.

[22] L. E. Kinsler, A. R. Frey, A. B. Coppens, and J. V. Sanders. *Fundamentals of acoustics*. John wiley & sons, 2000.

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

[24] S. Komkov and A. Petiushko. Advhat: Real-world adversarial attack on arcface face id system. *2020 25th International Conference on Pattern Recognition (ICPR)*, Jan 2021.

[25] N. Lavan, S. K. Scott, and C. McGettigan. Impaired generalization of speaker identity in the perception of familiar and unfamiliar voices. *Journal of Experimental Psychology: General*, 145(12):1604, 2016.

[26] G. Lavrentyeva, S. Novoselov, E. Malykh, A. Kozlov, O. Kudashev, and V. Shchemelinin. Audio replay attack detection with deep learning frameworks. In *Interspeech*, pages 82–86, 2017.

[27] J. Liljencrants. Tubes quality factor. http://www.fonema.se/qpipe/qpipe.htm.

[28] J. Lindberg and M. Blomberg. Vulnerability in speaker verification - a study of technical impostor techniques, 1999.

[29] S. Liu, H. Wu, H. yi Lee, and H. Meng. Adversarial attacks on spoofing countermeasures of automatic speaker verification, 2019.

[30] L. Mary and G. Deekshitha. *Searching Speech Databases: Features, Techniques and Evaluation Measures*. Springer, 2018.

[31] R. McAulay and T. Quatieri. Pitch estimation and voicing detection based on a sinusoidal speech model. In *International Conference on Acoustics, Speech, and Signal Processing*, pages 249–252 vol.1, 1990.

[32] Y. Meng, Z. Wang, W. Zhang, P. Wu, H. Zhu, X. Liang, and Y. Liu. Wivo: Enhancing the security of voice control system via wireless signal in iot environment. In *Proceedings of the Eighteenth ACM International Symposium on Mobile Ad Hoc Networking and Computing*, Mobihoc '18, page 81–90, New York, NY, USA, 2018. Association for Computing Machinery.

[33] M. J. Moloney and D. L. Hatten. Acoustic quality factor and energy losses in cylindrical pipes. *American Journal of Physics*, 69(3):311–314, 2001.

[34] L. Muda, M. Begam, and I. Elamvazuthi. Voice recognition algorithms using mel frequency cepstral coefficient (mfcc) and dynamic time warping (dtw) techniques, 2010.

[35] N. M. Müller, P. Czempin, F. Dieckmann, A. Froghyar, and K. Böttinger. Does audio deepfake detection generalize? *arXiv preprint arXiv:2203.16263*, 2022.

[36] A. Nagrani, J. S. Chung, W. Xie, and A. Zisserman. Voxceleb: Large-scale speaker verification in the wild. *Computer Science and Language*, 2019.

[37] A. Nagrani, J. S. Chung, and A. Zisserman. Voxceleb: a large-scale speaker identification dataset. In *INTERSPEECH*, 2017.

[38] C. J. Nederveen. Acoustical aspects of woodwind instruments. 1969.

[39] A. Owczarek and K. Ślot. Lipreading procedure for liveness verification in video authentication systems. In E. Corchado, V. Snášel, A. Abraham, M. Woźniak, M. Graña, and S.-B. Cho, editors, *Hybrid Artificial Intelligent Systems*, pages 115–124, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg.

[40] A. Pervaiz, F. Hussain, H. Israr, M. A. Tahir, F. R. Raja, N. K. Baloch, F. Ishmanov, and Y. B. Zikria. Incorporating noise robustness in speech command recognition by noise augmentation of training data. *Sensors*, 20(8):2326, 2020.

[41] A. Pianese, D. Cozzolino, G. Poggi, and L. Verdoliva. Deepfake audio detection by speaker verification. *arXiv preprint arXiv:2209.14098*, 2022.

[42] Y. Qin, N. Carlini, G. Cottrell, I. Goodfellow, and C. Raffel. Imperceptible, robust, and targeted adversarial examples for automatic speech recognition. In *International conference on machine learning*, pages 5231–5240. PMLR, 2019.

[43] J. Quiñonero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence. *Dataset shift in machine learning*. Mit Press, 2008.

[44] S. Rabanser, A. Thudi, K. Hamidieh, A. Dziedzic, and N. Papernot. Selective classification via neural network training dynamics. *arXiv preprint arXiv:2205.13532*, 2022.

[45] A. S. Shamsabadi, F. S. Teixeira, A. Abad, B. Raj, A. Cavallaro, and I. Trancoso. Foolhd: Fooling speaker identification by highly imperceptible adversarial disturbances. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6159–6163. IEEE, 2021.

[46] J. Shang, S. Chen, and J. Wu. Defending against voice spoofing: A robust software-based liveness detection system. In *2018 IEEE 15th International Conference on Mobile Ad Hoc and Sensor Systems (MASS)*, pages 28–36, 2018.

[47] M. Sharif, S. Bhagavatula, L. Bauer, and M. K. Reiter. A general framework for adversarial examples with objectives. *ACM Transactions on Privacy and Security*, 22(3):1–30, Jul 2019.

[48] B. Sisman, M. Zhang, S. Sakti, H. Li, and S. Nakamura. Adaptive wavenet vocoder for residual compensation in gan-based voice conversion. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 282–289, 2018.

[49] T. Smyth and J. S. Abel. Estimating waveguide model elements from acoustic tube measurements. *Acta Acustica united with Acustica*, 95(6):1093–1103, 2009.

[50] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur. Deep neural network embeddings for text-independent speaker verification. In *Interspeech*, pages 999–1003, 2017.

[51] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur. X-vectors: Robust DNN embeddings for speaker recognition. In *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, AB, Canada, April 2018.

[52] K. Stevens. Acoustic phonetics, cambridge, 1998.

[53] K. N. Stevens. *Acoustic phonetics*, volume 30. MIT press, 2000.

[54] R. Storn and K. Price. Differential evolution –a simple and efficient heuristic for global optimization over continuous spaces. *Journal of Global Optimization*, 11(4):341–359, 1997.

[55] M. Sugiyama and M. Kawanabe. *Machine learning in non-stationary environments: Introduction to covariate shift adaptation*. MIT press, 2012.

[56] P. Taylor. *Text-to-speech synthesis*. Cambridge university press, 2009.

[57] S. Thys, W. V. Ranst, and T. Goedemé. Fooling automated surveillance cameras: adversarial patches to attack person detection, 2019.

[58] N. Umetani, A. Panotopoulou, R. Schmidt, and E. Whiting. Printone: interactive resonance simulation for free-form print-wind instrument design. *ACM Transactions on Graphics (TOG)*, 35(6):1–14, 2016.

[59] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu. Wavenet: A generative model for raw audio, 2016.

[60] C. Wang, S. A. Anand, J. Liu, P. Walker, Y. Chen, and N. Saxena. Defeating hidden audio channel attacks on voice assistants via audio-induced surface vibrations. In *Proceedings of the 35th Annual Computer Security Applications Conference*, ACSAC '19, page 42–56, New York, NY, USA, 2019. Association for Computing Machinery.

[61] X. Wang and J. Yamagishi. A comparative study on recent neural spoofing countermeasures for synthetic speech detection. *arXiv preprint arXiv:2103.11326*, 2021.

[62] Y. Wang, W. Cai, T. Gu, W. Shao, Y. Li, and Y. Yu. Secure your voice: An oral airflow-based continuous liveness detection for voice assistants. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 3(4), Dec. 2019.

[63] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous. Tacotron: Towards end-to-end speech synthesis, 2017.

[64] E. Wenger, M. Bronckers, C. Cianfarani, J. Cryan, A. Sha, H. Zheng, and B. Y. Zhao. " hello, it's me": Deep learning-based speech synthesis attacks in the real world. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, pages 235–251, 2021.

[65] S. J. Winters, S. V. Levi, and D. B. Pisoni. Identification and discrimination of bilingual talkers across languages. *The Journal of the Acoustical Society of America*, 123(6):4524–4538, 2008.

[66] Z. Wu, S.-N. Lim, L. Davis, and T. Goldstein. Making an invisibility cloak: Real world adversarial attacks on object detectors, 2020.

[67] K. Xu, G. Zhang, S. Liu, Q. Fan, M. Sun, H. Chen, P.-Y. Chen, Y. Wang, and X. Lin. Adversarial t-shirt! evading person detectors in a physical world, 2020.

[68] J. Yamagishi, X. Wang, M. Todisco, M. Sahidullah, J. Patino, A. Nautsch, X. Liu, K. A. Lee, T. Kinnunen, N. Evans, et al. Asvspoof 2021: accelerating progress in spoofed and deepfake speech detection. *arXiv preprint arXiv:2109.00537*, 2021.

[69] C. Yan, Y. Long, X. Ji, and W. Xu. The catcher in the field: A fieldprint based spoofing detection for text-independent speaker verification. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, CCS '19, page 1215–1229, New York, NY, USA, 2019. Association for Computing Machinery.

[70] S. Yuan, P. Cheng, R. Zhang, W. Hao, Z. Gan, and L. Carin. Improving zero-shot voice style transfer via disentangled representation learning, 2021.

[71] L. Zhang, S. Tan, and J. Yang. Hearing your voice is not enough: An articulatory gesture based liveness detection for voice authentication. In *Proceedings of the*

2017 ACM SIGSAC Conference on Computer and Communications Security, CCS '17, page 57–71, New York, NY, USA, 2017. Association for Computing Machinery.

[72] L. Zhang, S. Tan, J. Yang, and Y. Chen. Voicelive: A phoneme localization based liveness detection for voice authentication on smartphones. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, CCS '16, page 1080–1091, New York, NY, USA, 2016. Association for Computing Machinery.

[73] Z. Zhou, D. Tang, X. Wang, W. Han, X. Liu, and K. Zhang. Invisible mask: Practical attacks on face recognition with infrared, 2018.

A Appendix

A.1 Proof: Tubes Cause Pitch shift

McAulay and Quatieri [31] 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_ℓ, ω_ℓ, and θ_ℓ represent the amplitudes, frequencies, and phases of the STFT peaks of the speech signal. Then, they find the value of ω_0 which fits $s[n]$ to $\tilde{s}[n,\omega_0]$ as:

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

where ω_0 is the signal pitch, $K(\omega_0)$ is the number of harmonics in the signal, $\widetilde{A}(k\omega)$ is the vocal tract envelope, and ϕ_k is the phase at each harmonic. Finally, the pitch is estimated by minimizing the mean squared error $\varepsilon(\omega_0) = P_s - \rho(\omega_0)$, where P_s is signal's power which is a constant. Therefore, we only need to minimize $-\rho(\omega_0)$, or equivalently:

$$\max \quad \rho(\omega_0) \tag{7}$$

where

$$\rho(\omega_0) = \sum_{k=1}^{K(\omega_0)} \widetilde{A}(k\omega_0) \left[\sum_{\ell=1}^{L} A_\ell \left| \operatorname{sinc}(\omega_\ell - k\omega_0) \right| - \frac{1}{2} \widetilde{A}(k\omega_0) \right]. \tag{8}$$

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:

$$\begin{aligned} &\max && \rho(\omega_0) && && (9)\\ &\text{s.t.} && A_\ell = 0 && \forall \ell \in \mathcal{L}, \forall \omega_0 \in [K(\omega_0)] \end{aligned}$$

Note that Eqn. (9) is a constrained version of Eqn. (7). We can solve the latter by maximizing the Lagrangian:

$$p(\omega,\eta) = \rho(\omega_0) - \sum_{k=1}^{K(\omega_0)} \sum_{\ell \in \mathcal{L}} \eta_{k\ell} A_\ell \tag{10}$$

where the matrix $\eta = [\eta_{k\ell}]_{K(\omega_0)\times|\mathcal{L}|}$ represents the Lagrange multipliers. Instead of directly maximizing Eqn. (10) and finding η, we re-write Eqn. (8) separating the components in and outside of $\mathcal{L}$:

$$\rho(\omega_0) = \rho_f(\omega_0) + \sum_{k=1}^{K(\omega_0)} \widetilde{A}(k\omega_0) \sum_{\ell \in \mathcal{L}} A_\ell \left| \operatorname{sinc}(\omega_\ell - k\omega_0) \right|. \tag{11}$$

where

$$\rho_f(\omega_0) = \sum_{k=1}^{K(\omega_0)} \widetilde{A}(k\omega_0) \Big[\sum_{\ell \notin \mathcal{L}} A_\ell \left| \operatorname{sinc}(\omega_\ell - k\omega) \right| - \frac{1}{2} \widetilde{A}(k\omega_0) \Big], \tag{12}$$

is the objective function for estimating the pitch of the filtered signal. Next, substituting Eqn. (11) in Eqn. (10):

$$p(\omega,\eta) = \rho_f(\omega_0) + \sum_{k=1}^{K(\omega_0)} \sum_{\ell \in \mathcal{L}} \left(\widetilde{A}(k\omega_0) \left| \operatorname{sinc}(\omega_\ell - k\omega_0) \right| - \eta_{k\ell} \right) A_\ell \tag{13}$$

Using the KKT conditions [11], we know for $p(\omega_0,\eta^*)$ to be the maximizer of Eqn. (13), the second term should vanish. Given $A_\ell > 0$, we should have that:

$$\eta_{k\ell} = \widetilde{A}(k\omega_0) \left| \operatorname{sinc}(\omega_\ell - k\omega_0) \right|. \tag{14}$$

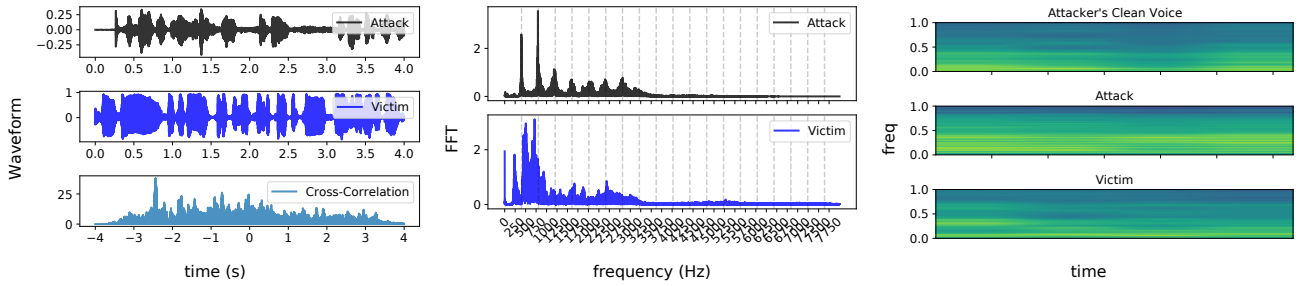
But that means $\rho_f(\omega_0) = p(\omega_0,\eta^*)$ is the exact solution to Eqn. (9), *i.e.*, the equality constraint holds perfectly.

Having established that the second optimization problem is a constrained version of the first, it follows that Ω, the feasibility set of Eqn. (7) is a subset of Ω_f, the feasibility set of Eqn. (9). Then, unless $\mathcal{L} = \emptyset$ (which trivially results in $\Omega = \Omega_f$), there exists $\omega_0 \in \Omega \setminus \Omega_f$ such that ω_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.

B Further Analysis of Mystique

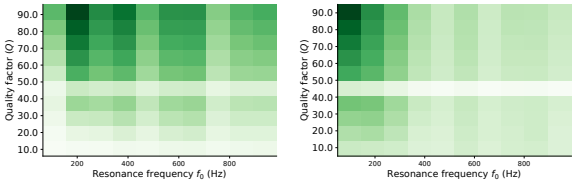
Model	Tubes						
	1	2	3	4	5	6	Avg
X-vector	87.53	85.8	82.45	76.47	84.22	85.95	83.74
SpeechBrain	88.88	84.31	83.69	82.56 2	80.38	85	84.14

Table 6: Mystique's over-the-air predictions consistency rate (%) across six repeated measurements.



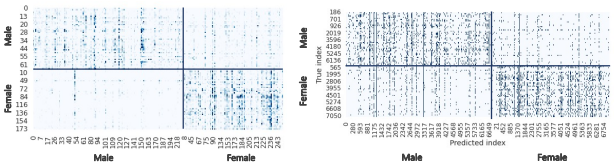
(a) Waveform (b) FFT (c) Spectrogram

Figure 10: Attack-victim pairs visualization when tube 1 ($L = 40.6, d = 3.45$ cm) is used: (a) the waveforms and their cross correlation, (b) FFT, and (c) spectrogram for a deeper look at the spectral content. along with the FFT of the BPF model applied to the chirp signal.



(a) X-Vector (b) SpeechBrain

Figure 11: Successful impersonations histogram using a single-tube configuration on (a) x-vecotr and (b) SpeechBrain. Most of them are generated by tubes that have Low f_0 and high Q_0 values.



(a) X-Vector (b) SpeechBrain

Figure 12: The confusion matrix of (a) x-vector and (b) SpeechBrain's predictions on Mystique attack split by the true (attacker) and predicted (impersonated) speakers sex. The cross-sex submatrix is sparse, indicating attack is more successful within same-sex speakers.

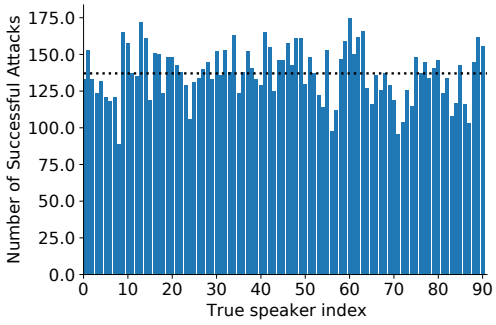


Figure 13: Number of successful impersonation attacks (out of 250) on x-vector model for each adversarial speaker from our VoxCeleb test set.

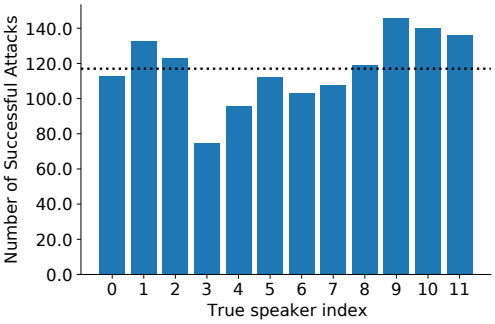
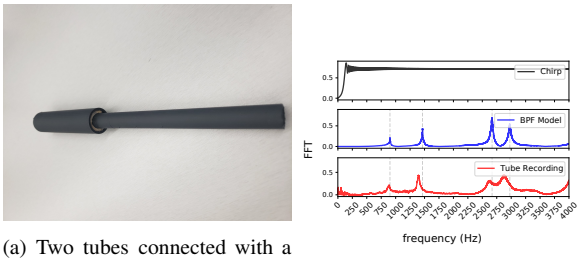


Figure 14: Number of successful attacks (false predictions) of the x-vector ASI model on the user study participants recordings.



(a) Two tubes connected with a HDF ring (b) FFT of a chirp

Figure 15: Two-Tube structure and resonance effect.