EchoVib: Exploring Voice Authentication via Unique Non-Linear Vibrations of Short Replayed Speech

S Abhishek Anand
University of Alabama at Birmingham
anandab.official@live.com

Jian Liu
University of Tennessee, Knoxville
jliu@utk.edu

Chen Wang
Louisiana State University
chenwang1@lsu.edu

Maliheh Shirvanian
Visa Research
mshirvan@visa.com

Nitesh Saxena
University of Alabama at Birmingham
saxena@uab.edu

Yingying Chen
Rutgers University
yingche@scarletmail.rutgers.edu

ABSTRACT

Recent advances in speaker verification and speech processing technology have seen voice authentication being adopted on a wide scale in commercial applications like *online banking and customer care support* and on devices such as *smartphones* and *IoT voice assistant systems*. However, it has been shown that the current voice authentication systems can be ineffective against voice synthesis attacks that mimic a user's voice to high precision. In this work, we suggest a paradigm shift from the traditional voice authentication systems operating in the *audio domain* but susceptible to speech synthesis attacks (in the same audio domain). We leverage a motion sensor's capability to pick up phonatory vibrations, that can help to uniquely identify a user via voice signatures in the *vibration domain*. The user's speech is *played/echoed* back by a device's speaker for a *short duration* (hence our method is termed *EchoVib*) and the resulting non-linear phonatory vibrations are picked up by the motion sensor for speaker recognition. The uniqueness of the device's speaker and its accelerometer results in a device-specific fingerprint in response to the echoed speech. The use of the vibration domain and its non-linear relationship with audio allows EchoVib to resist the state-of-the-art voice synthesis attacks, shown to be successful in the audio domain.

We develop an instance of EchoVib using the onboard loudspeaker and the accelerometer embedded in smartphones, as the authenticator, based on machine learning techniques. Our evaluation shows that even with the low-quality loudspeaker and the low-sampling rate of accelerometer recordings, EchoVib can identify users with an accuracy of over 90%. We also analyze our system against state-of-art-voice synthesis attacks and show that it can distinguish between the morphed and the original speaker's voice samples, correctly rejecting the morphed samples with a success rate of 85% for voice conversion and voice modeling attacks. We believe that using the vibration domain to detect synthesized speech attacks is effective due to the hardness of preserving the unique phonatory vibration signatures and is difficult to mimic due to the non-linear mapping of the unique speaker and accelerometer response in the vibration domain to the voice in the audio domain.

KEYWORDS

Voice Echo Fingerprint, Vibration Domain, Voice Imitation Resistance

ACM Reference Format:
S Abhishek Anand, Jian Liu, Chen Wang, Maliheh Shirvanian, Nitesh Saxena, and Yingying Chen. 2021. EchoVib: Exploring Voice Authentication via Unique Non-Linear Vibrations of Short Replayed Speech. In *Proceedings of the 2021 ACM Asia Conference on Computer and Communications Security (ASIA CCS '21), June 7–11, 2021, Hong Kong, Hong Kong*. ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3433210.3437518

ASIA CCS '21, June 7–11, 2021, Hong Kong, Hong Kong

ACM ISBN 978-1-4503-8287-8/21/06…$15.00
https://doi.org/10.1145/3433210.3437518

1 INTRODUCTION

Voice-based technology has gained popularity in many mobile and online commercial systems with increased use in personal [52] and business contexts, to interact online with their customers [38]. Financial institutions like Barclays [2], HSBC [3], and Wells Fargo [7] utilize voice authentication for mobile and online banking customers [44]. Several voice authentication applications (Nuance FreeSpeech customer authentication platform, Nuance VocalPassword voice biometric authentication system, virtual home assistance devices such Amazon Echo and Google Home, Voice Password, and Voice Screen Lock) also offer voice recognition capability. Personal mobile devices such as smartphones, laptops and wearables utilize voice authentication for unlocking the devices or launching on-board voice assistants as an additional, easy-to-use security feature [5, 6, 34, 53]. It is also deployed at the application level to provide added security that prompts for user identification via speech at an application launch [39, 51]. Standalone voice assistant systems such as Google Home and Amazon Echo are trained to recognize and execute only an authorized user's voice commands.

Voice authentication uniquely verifies a user by the pitch, tone, and volume of their speech forming the user's voice profile. Once a classification model is trained on this voice profile, a user can authenticate via a *challenge-response* mechanism. For verification, the system proposes a *challenge* to the user requiring a *response* in the user's voice. A user is authorized if the *response* fulfills the *challenge* and the responding voice matches the stored voice signature of a legitimate user. However, it has been shown that the voice authentication systems are prone to attacks that attempt to fool the system into accepting an unauthorized user. A potentially devastating form of such an attack is the *voice synthesis attack* where an attacker tries to impersonate a legitimate user through voice synthesis. These attacks are easy to implement and only require a limited number of short speech samples of a few seconds each, from the victim, for training. Voice synthesis attacks can be launched using a microphone to record the legitimate user's voice samples and use any publicly available speech transformation techniques (CMU Festvox voice converter [19]), or voice modeling technology (e.g., Lyrebird [46]). These attacks are shown to be highly effective, with voice synthesis attacks being successful up to 80%-90% on automated voice recognition systems [49].

We propose a novel voice authentication method based on the unique speech footprints captured by the motion sensors in the *vibration domain*. We make use of the motion sensors' ability (accelerometer, in particular) of picking up minimal vibrations to detect and record the *acoustic vibrations*. These acoustic vibrations, generated from the speakers — ***when an authenticating user's voice response is replayed for a short duration*** (e.g., 1-2 seconds) — produce a unique imprint on the motion sensor recordings. Using this method, we can perform user authentication by utilizing their speech patterns, captured in the vibration domain. This approach can be applied to not only speaker verification for online services but also in the case of personal usage such as smartphone/smart device access and smart voice assistant scenario.

Our method, EchoVib (Figure 1), is based on the premise that vibration features of the user's speech are unique and are hard to imitate or synthesize by a spoofing attack (for example, [49]). This premise is supported by prior research in linguistics [27] which argues that motion-sensing devices like accelerometers can be used for monitoring vocal characteristics such as pitch range, distributions, perturbations, and time domain features such as voice/voiceless ratios, word and syllable duration, speech rate, and pitch change over time. These vocal characteristics can be captured by the accelerometers due to their "airborne-signal-rejecting" capability, and as such are *different* from vocal characteristics captured by a microphone. Sundberg et al. [57] referred to these speech vibration features, in the context of singing, as *phonatory vibrations*. Thus, authentication in EchoVib is derived from the vibrations in the device but crucially the device replaces the human vocal tract as the source of the speech.

Audio playback of the user's response via a device's speakers is crucial to EchoVib because this playback produces on-board phonatory vibrations that can be captured by the motion sensors residing on the device. Das et al. [28] showed that imperfect fabrications during the manufacturing of speakers in the smartphones lead to anomalies during sound production. These anomalies have been found to be unique for each smartphone device making them ideal for device fingerprinting. Moreover, imperfections in accelerometers also lead to device-specific response [20, 29, 33]. This result, when combined with [28], leads to a vibration-based authentication mechanism that is also tied to each individual device. ***Live aerial speech signals are not strong enough to affect these sensors (at least on the typical smartphone devices)*** as shown in [14, 27], thereby highlighting the need for playback for our purpose. An example usage scenario of EchoVib could be: when a user needs to be verified while making a payment during online shopping using a smart device like Alexa or HomePod, the user would need to utter the phrase "Alexa, make payment using XYZ card" and the EchoVib app on the device would replay the user's command for verification (even partial replay for short duration is sufficient).

Voice synthesis attacks exploit spectral features of the victim's voice (as captured in the audio domain by a microphone) by generating audio samples that mimic the victim's voice, thereby fooling most state-of-the-art audio-domain voice authentication systems. EchoVib works on using a live human speech's phonatory vibration response as captured in the vibration domain by the motion sensors. We show that these synthesized voice samples do not contain the phonatory vibration features of the victim's voice samples. Thus, even though these synthesis attacks have succeeded against machine-based speaker verification systems (e.g., Universal Background Modeling in Gaussian Mixture Model (UBM-GMM) [54] and Inter-Session Variability (ISV) [61]) [49] by mimicking the audio frequency response, they are unsuccessful against the proposed EchoVib mechanism (Appendix Figure A.2).

EchoVib is resilient to synthesis attacks in the vibration domain due to the non-linear mapping of vibrations from the audio and the uniqueness of the device's components generating and recording the speech vibrations. The vibrations generated by a speech sample not only depend on the speech sample characteristics but also on the source generating the vibrations in response to the speech signal. In [27], the vibration source was the human vocal tract, while in our proposed EchoVib it is the device that plays backs (echoes) the speech. Thus, the proposed authentication model comprises of user's speech features and the device fingerprint (vibration generating characteristics) that generates the unique vibrations corresponding to the user's speech. Prior works [20, 28, 29, 33] show that subtle differences in the speakers and the accelerometers in the smartphones make each device (even that of the same hardware make and model) unique. This implies that EchoVib relies on both the uniqueness of the speech vibrations (from the speakers) and that of the accelerometer and will thus be resistant to the synthesis attacks. Additionally, the attacker needs to generate a correct response to the EchoVib challenge-response authenticator, while synthesizing a vibration response that could pass the verification. Overall, the attacker would need to match the vibration characteristics and build the correct answer to the posed verification challenge, which is fundamentally challenging to accomplish simultaneously.

We outline the design of our proposed authentication system and its workflow in Figure 1. The *user-facing device* U is the input interface that accepts the user's response to a *challenge-response setup* that needs to be authenticated. An example is a smart device (smartphone, Google Home, Amazon Echo, Apple HomePod, etc.) with a microphone to capture speech. The *EchoVib component* E consists of a speech echoing (via a loudspeaker) and a vibration sensing (via an accelerometer) device. This component can be realized by using the on-board loudspeaker and the accelerometer of a smartphone. The *speaker authentication entity* A is responsible for processing the vibration signature extracted from the user's speech sample output by E. It matches the received vibration signature to a stored model of phonatory vibrations of a legitimate speaker. Finally, the result of the authentication process is sent back to U.

Our Contributions: In this work, we propose an authentication method that aims to improve upon the security of the current voice-based authentication systems. We use an MEMS accelerometer to

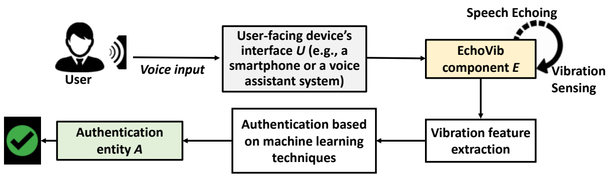

Figure 1: A high-level overview of the proposed EchoVib authentication system. Short playback of user's speech is sufficient to identify the user.

measure the unique vibrational signatures of the user's voice response, replayed via speakers. We apply off-the-shelf hardware and standard machine learning techniques to implement our proposed method and perform speaker authentication, under benign and previously mentioned state-of-the-art voice synthesis attacks.

(1) ***Design & Implementation of a Novel Voice Authentication System*:** We propose a new authentication system (Section 3) that is based on capturing an individual's voice signature in the vibration domain. We develop an instance of our authentication system, using standard machine learning techniques and inexpensive hardware (the embedded accelerometer and the loudspeaker on smartphones). While we apply standard techniques in our voice classification task, we comprehensively analyze different feature sets and classification algorithms that allow unique identification of individuals, from the vibrations of their speech, even if the sampling rate of the voice-vibration sensing device (accelerometer on the smartphones) is low. Our design rationale is further supported by the measurements from a high-resolution *laser doppler vibrometer.*
(2) ***System Performance in Benign Scenario*:** The proposed method is tested on a total of 30 speakers from three datasets (10 speakers each). It is able to authenticate a speaker with high accuracy (true positive ≥ 97%; F-measure ≥ 95%) as shown in Section 5. These results indicate that our method can solely identify a speaker using the vibration pattern of their voice.
(3) ***Resilience against Voice Synthesis Attacks*:** We evaluate our proposed method against state-of-the-art voice conversion attacks (Section 6.2). We show that our authentication method is secure against these attacks due to the hardness of the vibration pattern imitation. Our proposed authentication method is able to distinguish between the converted voice and the original voice with correctly identifying over 85% of the fake samples when trained on an original speaker's voice samples.
We also test our system against the state-of-the-art voice modeling attacks that claim to produce natural sounding human voices, using advanced deep learning algorithms (Section 6.3). We use Lyrebird's [46] voice modeling service to convert the voice samples collected from Amazon Mechanical Turk volunteers into synthesized speech and test them against EchoVib. Our proposed system is able to distinguish and reject over 86% of the generated voice samples when trained to recognize only the original voice samples for each user.

2 BACKGROUND AND RELATED WORK

2.1 Voice Authentication in the Audio Domain

Authentication Features: Current voice authentication systems rely on extracting and distinguishing an individual's voice features. For instance, short-term spectral based features (e.g., MFCC [50] and spectral subband centroids (SSCs) [42]) are usually used to describe a voice's timbre as well as the resonance properties of the supralaryngeal vocal tract, which can be used to distinguish people's voices. Moreover, spectro-temporal features (e.g., modulation frequency [16]) could show the details of a signal's formant/energy transitions and have been demonstrated to contain useful speaker-specific information. Additionally, prosodic features (e.g., fundamental frequency, speaking rate, and energy distribution) refer to non-segmental aspects of speech like syllable stress and intonation patterns, which can also be used for speaker recognition [11, 18].
Voice Impersonation Attacks: Although it has been shown that the aforementioned features can be successfully used for speaker recognition, they are vulnerable to impersonation or spoofing attacks. An adversary could impersonate or synthesize the victim's voice by using the recordings of their daily speeches and using voice synthesize techniques [45] to compromise the system. In addition, the adversary could modify any speaker's voice to sound like the victim to spoof the system by using voice morphing tools [49].
Defenses in the Literature: Recent studies show that advanced speaker models, such as Gaussian mixture model-universal background model (GMM-UBM) [13] and i-vector models [36, 37], could detect voice impersonation. The method proposed in [13] could detect 95.83% of the disguised voices, generated by human voice artists. To defeat the synthetic speech attack, features based on the relative phase shift to classify HMM-based synthetic speech from human speech is proposed [30]. It is able to detect synthetic speech up to 97.5%, while retaining the ability to perform authentication at 97.0%. This method however needs to be trained using the same voice encoding algorithm as the attacker. Wu *et al.* [65] proposed modulation features to capture speech variation cross frames for detecting synthetic speech. They reported an equal error rate of 7.17% (MFCC + magnitude modulation) and 0.89% (Modified group delay cepstral coefficients + phase modulation). The data corpus required for training (4000 human and synthetic speech samples each for training, 3000 each for development) is however significantly bigger than the one used in our work.

Researchers recently proposed to determine the liveness of the sound source by exploiting the physical features of human speech [9, 24, 67, 68]. To defend against the impersonating sound from a loudspeaker, Chen *et al.* [24] utilized the magnetic field emitted from the electro-acoustic transducer. VoiceLive [68] and VoiceGesture[67] exploit time-difference-of-arrival (TDoA) and Doppler shifts to detect the dynamic acoustic characteristics for liveness. However, these liveness detection approaches target to verify the speaker in front of the smartphone, which requires the phone to be held close to the speaker's mouth. Moreover, Feng *et al.* [35] perform user authentication on operating voice assistant (VA) systems through a specialized eyeglasses worn by the user. The eyeglasses are equipped with an accelerometer under high frequency (i.e., $64kHz$) which captures the user's body-surface vibrations used to match with the voice command. This method, however, requires an extra dedicated device like a smart glass to perform authentication in contrast to EchoVib.

Another effort to detect spoofed/fake voice attacks was by Automatic Speaker Verification Spoofing and Countermeasure challenge

(ASVSpoof) [1]. The challenge aims to develop a generalized approach towards detecting spoofed or voice conversion attacks, by developing a classifier that can perform well against both known and unknown spoofing attacks. [64] explored the automatic speech recognition systems' vulnerability against voice spoofing and conversion methods and evaluated the feature-based anti-spoofing countermeasures. They reported a false acceptance rate (FAR) of below 1.0% against known spoofing attacks. The FAR against unknown attack was 12.3% and the proposed method did not perform well against SS-MARY attack [56].

Our proposed mechanism produces a comparable accuracy for detecting synthesized voice, without relying on known spoofing algorithms and a big training dataset. Our reported accuracies ($> 85\%$) are comparable to [64] as we do not rely on previous knowledge of the voice synthesis algorithm. Prior works [30, 65] have reported better accuracies but with significant constraints such as prior knowledge of voice encoder used by the attacker or a large training dataset. Some of these detection techniques are also computationally intensive and often require substantial training data. Additionally, these approaches rely solely on the audio domain, unlike our proposed system EchoVib. Hence, they are vulnerable to an adversary who possesses the knowledge of the system's authentication features and speaker model in the audio domain. Our approach is orthogonal as it works in the vibration domain and can be used in conjunction with the audio domain approaches such as [64].

We have explored the unique effect of the speech played by the phone's built-in speakers on its own motion sensor to distinguish speakers and defend against both voice conversion and synthesis attacks. In particular, by working in the vibration domain, EchoVib can resist the traditional voice synthesis attacks that mimic the features of a given user's voice in the audio domain. Our approach can be seamlessly integrated with traditional voice authentication systems and other defenses that operate in the audio domain, to provide an additional important layer of security.

In [28], both the micro-speakers and the microphones found in smartphones, could be used to fingerprint the device with a high degree of accuracy, even with devices of the same make and model. The work in [20, 29, 33] concluded that hardware imperfections in sensors such as accelerometers can be utilized for uniquely identifying an individual smartphone (device fingerprinting). In addition, [29] also inferred that audio simulation applied to the motion sensors can indeed improve the sensor fingerprinting results. These results justify our premise that the audio simulation of the accelerometer via human speech, produces device-specific phonatory vibrations that can be leveraged as an authentication factor, resilient to the traditional voice synthesis attacks.

2.2 Speech in the Vibration Domain

Motion sensors (i.e., accelerometer and gyroscope) are used in mobile devices and speakers (e.g., HomePod) to meet the demands of various applications. They consist of MEMS structures, which could be easily affected by sound and noises [22, 31, 32]. Due to the effect of sound on motion sensors, WALNUT [60] models the physics of acoustic injection attacks on accelerometers and shows the outputs of sensors are subjected to the acoustic interference. Moreover, an external loudspeaker sound source has been proved to impact motion sensors. For instance, Gyrophone [47] shows that gyroscope could be used in an attack to measure acoustic sound from a loudspeaker sharing the same surface as the sensor/phone and compromise speech privacy (e.g., speech contents) through classification. Speechless [14] further points out, to launch such an attack, there is an essential need for a shared surface between the external loudspeaker and the device containing the motion sensor, which can capture the conductive vibrations.

EchoVib verifies the speech played back by the device's loudspeaker, using its built-in motion sensors. Accelword [66] demonstrated the smartphone's accelerometer capability to extract hotwords (e.g., Okay Google) from human voices. EchoVib further performs user verification (a defensive application, in contrast to the offensive applications explored in Gyrophone) by examining the unique effect on motion sensors when the sound source comes from the device's built-in speakers. Our solution can be implemented by using off-the-shelf hardware and software, easily integrated with most of the mobile devices and voice assistant systems.

3 OUR APPROACH

3.1 Overview

Traditional voice authentication schemes utilize the spectral features of speech for speaker classification. In EchoVib, we use the impact of speech on motion sensors (accelerometer, in particular) such as the ones ubiquitously found in smartphones to defend against *voice synthesis attack*. The speech, when played via a loudspeaker, generates a vibrational response within the device housing the speaker (in our case, a smartphone) or on the device's body and the surrounding surface (in case of an external loudspeaker). This vibrational response is due to the motion of the speaker cone/diaphragm, which is produced due to an underlying coil motion generated from varying magnetic field. Our model assumes that the speech features used in voice authentication, also produce a unique response in the vibration pattern of the speaker diaphragm (henceforth referred to as vibrational features of speech). This response can be recorded by the motion sensors and utilized for speaker recognition.

In this work, we implement EchoVib on smartphones as they have the necessary hardware E: a loudspeaker to replay/echo the user's response (in a *challenge-response setup*) and an on-board accelerometer to record the resulting vibration signature of the user's response, on the same device (Figure 1). In the benign use case scenario, the user trains the system on their verbal responses, similar to the voice authentication systems. The verbal responses of the user are recorded by U and replayed/echoed by E. The resulting vibration signatures are utilized by A to build a training model of the user's phonatory vibrational signature. This training model can then be used by A to authenticate any future user responses.

The attack scenario (Figure A.2) works similar to the benign use scenario, where the attacker provides a response to EchoVib for authentication, presumably imitating a legitimate user's response. EchoVib receives this response via U, relays it via E, and finally verifies it against a stored model of the legitimate user's response by A. For secure authentication, A should be able to reject the attacker's response. In our approach, we believe that the device-specific vibration features of speech cannot be copied by spectral mimicking of the speech sample in the audio domain and hence

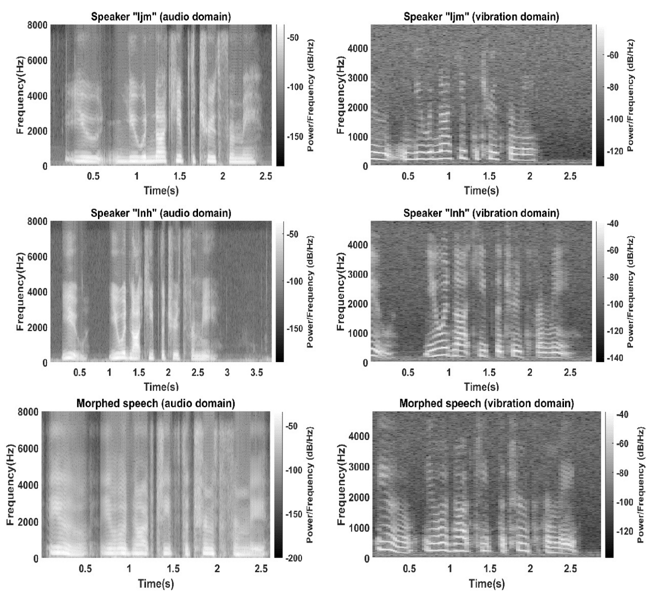

Figure 2: Power spectral density of speech features in audio (left) and vibration domain (right) for speakers "ljm", "lnh" from CMU_ Arctic dataset, and the synthesized speech for "ljm" (bottom) generated using Festvox Transform voice conversion, speaking "You, you would not keep the truth from me."

the morphed samples created by the attacker will not generate the same vibrations as the actual samples in the victim's voice.

3.2 Design Rationale

Benign Setting: The underlying fundamental principle behind EchoVib is the uniqueness of the speech pattern in vibration domain as recorded by the motion sensors. In traditional voice authentication systems, speech features are extracted from a recording of a legitimate user's speech to build a speech profile for that user. We term this process as *speech pattern in the audio domain* as the speech features are extracted from the audio recording of the speech. In contrast, our proposed system EchoVib uses the effect of speech on motion sensors and extracts features from this recording of motion sensor readings in the presence of speech to build a user's profile, hence termed as *speech pattern in the vibration domain* or *phonetic vibrations* as explained in Section 1.

To observe differences in the individual speakers' voice in the vibration domain, we used a laser doppler vibrometer (PDV-100) [4] to measure the phonatory vibrations in the body of the smartphone during speech replay. The laser doppler vibrometer points a laser beam at the vibrating surface (smartphone's body in our experiment, as shown by our experimental set-up in Appendix Figure A.1), and uses the doppler effect on the reflected beam to measure the vibrations. The sampling rate for the vibrometer, in our experiments was 10kHz therefore the vibration effects can be captured with a similar resolution as the audio samples.

Figure 2 shows the effect of audio on accelerometer when the system plays back a speaker's voice sample while recording accelerometer readings. We tested two audio samples, one recorded for speaker "ljm" and other for speaker "lnh" taken from the CMU_Arctic dataset [41] speaking the same sentence "You, you would not keep the truth from me". In the audio domain, we can see noticeable differences between the frequency distribution in the speech pattern of the two speakers. This difference in the effect of speech of individual speakers is also reflected in the vibration domain (captured by the vibrometer) while also indicating that the frequency distribution in the vibration domain for a particular speaker is different from the frequency distribution in the audio domain. Thus, we believe that the uniqueness of speech pattern among different speakers is retained in the vibration domain as well.

Threat Model and Attack Setting: Audio domain voice authentication has been shown to be susceptible to multiple attack vectors that aim to gain unauthorized access by presenting the authentication system, a speech sample closely resembling a legitimate user's voice response. In this work, we consider two such types of attacks namely voice conversion attack and voice modeling attack.

Our attack model can be defined by three phases: voice sample collection, voice synthesis model generation, and attacking voice authentication systems using synthesized voice samples. In *voice sample collection*, the attacker aims to collect sufficient speech samples in a targeted victim's voice. The attacker can eavesdrop and record the victim's voice unknown to them using a hidden voice recorder, or use publicly available speech samples in the victim's voice (through videos hosted on popular social media platforms). In *voice synthesis model generation*, the attacker attempts to build a voice synthesizing system, trained on the collected voice samples of the victim in the first phase. The model takes in a speech sample (not in the victim's voice) and generates a speech sample that closely resembles the speech pattern of the victim.

Once the attacker has trained the voice synthesis system on an intended victim's speech samples, it can try to gain unauthorized access to sensitive information or resources and are secured via a voice authentication mechanism. For this purpose, we also assume that the attacker has temporary or permanent access to the victim's device, that is normally used by the victim, and holds the sensitive information or resources. The attacker can then play the synthesized voice sample to the device as a response to the *challenge-response setup* and tries to get authenticated by the voice authentication system. In another scenario, the attacker can try to fool a remote voice authentication system (such as online banking) by playing responses crafted by the voice synthesizing system in the victim's voice and attempts to exploit the victim by illegally gaining access to the victim's confidential data and resources (for example, bank account details, passwords, PINs, etc.).

Figure 2 shows the original and attacker-generated speech for a speaker in the audio and vibration domain (as captured by the laser doppler vibrometer). In the audio domain, the power spectrum for both speech samples shows a similar pattern, though the morphed speech has more diffused power spread compared to the original sample. However, the effect in vibration domain is different, especially in the frequency distribution pattern. Thus, we believe that the vibration domain is helpful in capturing spectral features that cannot be imitated by the morphing model of the attacker.

4 SYSTEM DESIGN

As detailed in Section 2.2, the phonatory vibrations from an audio source have the capability to travel through the shared medium and affect the motion sensors on a smartphone. While the effect may be limited due to the low sampling rate of the motion sensors (restricted by the operating system), it can still be unique and utilized for speaker authentication. In our authentication model, we propose to utilize the speech characteristics captured by the smartphone motion sensors, for speaker identification. While the sampling rate of MEMS motion sensors is around 8-10kHz, they are limited to a much lower sampling rate as per the on-board Operating Systems, such as Android on the smartphone (for example, approximately 200Hz or lower by Android). The fundamental frequency of an adult male is between 85-180Hz and for an adult female, it ranges from 165-255Hz [17, 58]. As per the Nyquist sampling theorem, a device with a sampling rate of 200Hz can only capture frequencies up to 100Hz from a signal. Thus, it may not be able to capture the full range of frequencies present in the speech though due to harmonic effect, still it may be possible, as our design shows, to capture some part of the missing frequencies due to aliasing.

4.1 System Setup

In a challenge-response authentication protocol, the challenge consists of a pre-determined passphrase asked by the verifier (interface U) and the valid audio response must have matching spectral features to a verified user's speech characteristics. In contrast to a speaker verification system where the verification is performed on the response from the user (an audio of the user's voice), EchoVib measures the response of motion sensors by playing back the user's response. In our setup, we use the smartphone's microphone as the interface U accepting the user's response (*challenge-response setup*) and the component E replaying/echoing it is constituted by the smartphone's on-board loudspeaker and its accelerometer.

4.2 Workflow

The workflow of the proposed authentication model is a multi-step procedure as described below:

- *Initialization/Training Phase:* During the initialization phase, the user is required to utter multiple phrases/sentences for training the authentication model. EchoVib replays the user's voice sample and records the corresponding accelerometer readings to measure the impact of phonatory vibrations, generated during the playback. EchoVib then extracts speech features from the vibration domain (accelerometer readings) and trains a machine learning classifier on the extracted feature set. The details of the speech feature set and the machine learning classifier are described in Section 4.3 and Section 4.3.1.
- *User Credential Input Phase:* Here, the user provides a voice response to U for authentication. The user's response is replayed/echoed while the accelerometer is used to record its phonatory vibrations by E. EchoVib uses a smartphone as both U and E since it has the microphone to accept the user's response, the speakers for echoing the received response, and an on-board accelerometer to record its phonatory vibrational signature. Thus, it requires less hardware and more mobility at the cost of limited sampling rate of the accelerometer.
- *Feature Extraction and Verification Phase:* The feature extraction phase consists of extracting the required speech features from the recorded accelerometer readings during speech replay. Once extraction is complete, EchoVib (A, in particular) attempts to recognize the speaker based on the speech features fed to the classifier algorithm. If the speech is categorized as belonging to one of the trusted users, the authentication is complete, and the user is verified otherwise the user is rejected.

4.3 Feature Set & Classification Models

We now discuss the speech feature sets that can be used to identify a user based on the provided motion sensor readings. In Section 5, we provide a comparison of the two features sets for our classification.

Mel-frequency Cepstrum Coefficients: Mel-frequency Cepstrum Coefficients (MFCC) are used in speaker identification where the speech signal is expressed on a mel-frequency scale, consisting of a logarithmically spaced filter modeled on the human ear reception of the speech. MFCC features are not affected by the changes in the variation of the vocal chords during the speech generation hence present a stable speaker characterization basis.

Time and Frequency Domain Feature Set: The time and frequency domain features (Appendix Table A.1) consist of several statistical measurements of the signal over x, y, and z axis. Some features are calculated individually per axis (for example, minimum, maximum, quartile, inter quartile range) while some are calculated over all the three axes (total absolute area of the signal and total signal magnitude averaged over time). Using both types of features helps us identify the temporal characteristics of the signal in addition to its spectral features (insensitive to phase variation) making it easier to analyze signal properties in a specific frequency band.

Feasibility Analysis: We now examine how the speech features could identify different speakers and distinguish the attacker-generated speech from the original speech. Figure 5 presents the distribution of speech features derived from two speakers' speech samples in vibration domain, when they speak 100 different sentences (CMU_ Arctic dataset). We observe that many speech features in vibration domain such as second quartiles, third quartiles, mean-cross rate and the ratio of standard deviation to mean derived from Z accelerations and X accelerations show diverse distributions for the two speakers. These speech features capture the unique speech footprints of the speakers and can be utilized to identify the speakers. Figure 6 compares distributions of the derived speech features from the attacker-generated speech and the original speech of speaker "ljm" in the vibration domain, when she speaks 60 different sentences (CMU_ Arctic dataset). We find that the unique features such as variance, standard derivation, second/third quartiles, ratio of the standard deviation to mean, kurtosis, energy from the X, Y and Z accelerations show different distributions for the morphed and the original speech. Thus, these speech features can effectively recognize the morphed speech from the speaker's original speech.

Figure 3 and Figure 4 further illustrate the use of speech features in our scheme. In particular, Z accelerations, second quartiles, third quartiles and mean-cross rate, classify the two speakers' speech into two well-separated clusters, with the same speaker's speech in the same cluster as shown in Figure 3. Figure 4 illustrates that the unique features like the variance of Z accelerations, the ratio of standard derivation of X acceleration to its mean and the entropy of Y accelerations distinguish the morphed speech from the original speech of speaker "ljm". By including more speech features, different speakers can be better classified, and their morphed speech can be better distinguished, in the higher dimensional feature space.

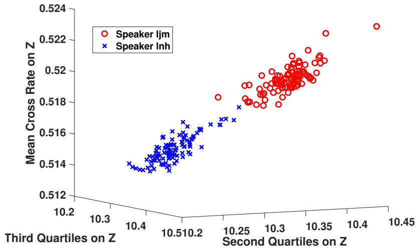

Figure 3: Illustration differences of using the speech features to distinguish two speakers "ljm" and "lnh" (CMU_ Arctic dataset) in vibration domain.

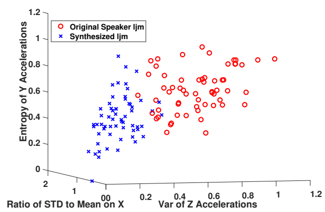

Figure 4: Illustration of using the speech features to distinguish the original speaker "ljm" speech from the synthesized speech generated using Festvox Transform voice conversion (CMU_ Arctic dataset) in the vibration domain.

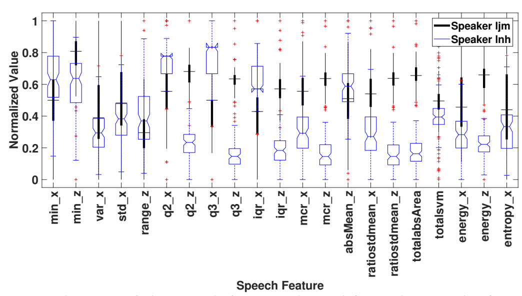

Figure 5: Distributions of the speech features derived from the speech of two speakers "ljm" and "lnh" (CMU_ Arctic dataset) in vibration domain.

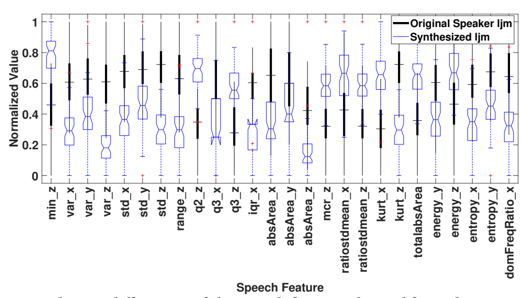

Figure 6: Distribution differences of the speech features derived from the original speaker "ljm" speech and the synthesized speech generated using Festvox Transform voice conversion (CMU_ Arctic dataset) in vibration domain.

4.3.1 Classification Algorithms. We test the following classification algorithms on the two feature sets and choose the best performing pair of classification algorithm and feature set (Section 5).

Support Vector Machine: It is a binary classifier that uses a hyperplane to divide the input variable space into two categories. The hyperplane is determined during the training phase by using optimization techniques to maximize the binary separation of input variables. For a multi-class scenario, several binary support vector machines can be used together to categorize the input variables.

Regression based Classifiers: A logistic regression classifier predicts the outcome of a dependent variable based on one or more independent variables using a logistic function.

Decision Tree based Classifiers: Random tree and Random forest are decision tree based classifiers that tries to predict the outcome of an output variable when supplied with multiple input variables.

5 IMPLEMENTATION AND EVALUATION

5.1 Setup and Preliminaries

Voice Datasets: We used Voxforge [8], CMU_Arctic [41], and a speech dataset built using Amazon Mechanical Turk workers. For Voxforge and CMU_Arctic datasets, we selected ten speakers reading CMU_Arctic sentences. We use CMU_Arctic dataset as it provides over 100 samples per speaker while Voxforge dataset was chosen especially for testing voice conversion attacks as it has been shown susceptible to these attacks [49]. "Amazon Mechanical Turk dataset" was created for testing Lyrebird speech modeling attack that uses Lyrebird web-based audio recording tool. To build this dataset, we published an IRB approved Amazon Mechanical Turk HIT and asked ten workers (American English accent) to speak and record sentences provided by Lyrebird web-based tool. Participation in the survey was voluntary and the participants could withdraw any time. We converted all samples to WAV file format at 16 kHz sampling rate. The spoken sentences in all of our datasets were short in length averaging about 8 words per sentence.

Equipment: While EchoVib is designed to be used with any speaker (loudspeaker or inbuilt device speaker) and motion sensors (external motion sensors or on-board motion sensor), we have used smartphone to implement EchoVib. A smartphone houses both a speaker and on-board motion sensors on the same device, providing a convenient setup for EchoVib. We used five different smartphones: Samsung Galaxy S8, Samsung Note 5, Samsung Galaxy S6, Samsung Note 4, and LG G3 to act as devices to implement EchoVib.

Samsung Galaxy S6 and LG G3 have Invensense MPU-6500 as the embedded motion sensor chip housing a 6-axis gyroscope and accelerometer. Samsung Galaxy Note 4 has Invensense MPU-6515 as the motion sensor chip that has a similar capability as MPU-6500. Samsung Galaxy S8 has the LSM6DSL motion sesnor chip from STMicroelectronics while Samsung Note 5 has K6DS3TR accelerometer sensor from the same manufacturer. These motion sensor chips have similar output data rates to the motion sensors on smartphones available in the market, restricted by Android platform to be around 200Hz. In addition, they share similar physical characteristics such as mechanical resonant frequency and precision with other popular phones in the market. Hence, we believe our choice of phones represents a fair variety of motion sensor capabilities.

On each smartphone, we logged the motion sensor readings while a user's speech response was being played through the phone's speakers. The phone was kept flat on a wooden surface with no external motion being present in the surrounding environment. Another possible scenario could be keeping the phone hand-held where it may be in motion due to the user's body movement. In case of a user walking while holding the phone, it has been observed [66] that the effect of such motion can be eliminated by applying a high pass filter (> 2Hz) on the recorded signal. We used 10 fold cross validation method for speaker classification. The sample space

Table 1: 10 speaker classification on Voxforge dataset (10-fold cross validation). Random Forest yields highest accuracies.

Time frequency features	Samsung Galaxy S6			Samsung Note 4			LG G3		
Classifier	**Precision**	**Recall**	**F-measure**	**Precision**	**Recall**	**F-measure**	**Precision**	**Recall**	**F-measure**
Simple Logistic	0.87	0.84	0.85	0.88	0.88	0.88	0.94	0.94	0.94
Support Vector Machine (SMO)	0.89	0.88	0.88	0.89	0.89	0.89	0.95	0.94	0.94
Random Forest	0.96	0.96	0.96	0.91	0.91	0.91	0.94	0.94	0.94
Random Tree	0.87	0.87	0.88	0.79	0.79	0.79	0.92	0.92	0.92
J48 Tree	0.89	0.88	0.88	0.83	0.83	0.82	0.92	0.92	0.92
MFCC features									
Simple Logistic	0.80	0.74	0.78	0.73	0.73	0.73	0.85	0.84	0.84
Support Vector Machine (SMO)	0.87	0.87	0.87	0.69	0.68	0.68	0.80	0.80	0.80
Random Forest	0.91	0.91	0.91	0.75	0.75	0.75	0.85	0.84	0.85
Random Tree	0.77	0.77	0.77	0.58	0.58	0.58	0.68	0.69	0.69
J48 Tree	0.84	0.84	0.84	0.73	0.73	0.73	0.82	0.82	0.82

is divided into ten disjoint subspaces, of equal size, in a random manner. Nine subspaces are used for training while the remaining one subspace is used for testing. We show the configuration of each classifier in Weka in Appendix Table A.1.

Signal Processing and Machine Learning Tools: Once the sensor readings were recorded by our application, we transferred the log file containing the readings, to an offline system for processing. In the real world, EchoVib can send the sensor readings to a trusted cloud server that processes them for verification and returns the result to EchoVib. EchoVib could also be implemented on the device itself, provided the device contains enough processing resources without any power constraint. For processing the sensor readings, we used Matlab's signal processing functions to extract the relevant feature set from each recorded sample. The classification model was trained and tested on Weka toolkit [15] that provides a workbench for implementing machine learning algorithms.

Classification Metrics: For measuring the effectiveness of EchoVib, we use *True Positive* and *False Positive* as metrics to determine what percentage of provided responses have been correctly identified as belonging to a legitimate user (true positive) and what percentage of provided responses have been incorrectly identified for the legitimate user (false positive) and average the results over the number of tested speakers. An effective authentication model should have a high value of true positive and a very low value of false positives from a security perspective. In addition, we also use *precision*, *recall*, and *F-measure* to measure the efficiency of EchoVib in classifying speakers. *Precision* is defined as the ratio of correct class predictions to the total number of classes. *Recall* is defined as the ratio of correct class predictions to total number of correct class instances. *F-measure* is the harmonic mean of *precision* and *recall* attaching equal weights to both precision and recall values.

5.2 Comparing Feature Sets and Classifiers

While both MFCC and time-frequency domain features can be used to classify a speaker's voice in *audio domain*, our implementation requires a feature set that works well in the *vibration domain*. We compared the accuracy of both the feature sets using the classifiers (Section 4.3.1) and the metrics (Section 5.1). We tested three smartphones, multiple datasets, and signal processing tools (Section 5.1). The results from our feature set comparison study on the Voxforge dataset are detailed in Table 1. Using the metrics in Section 5.1, we observe that time-frequency domain features performed better than MFCC feature set with an F-measure value > 0.90 for all three phone models compared to MFCC feature set that showed the best F-measure value of 0.75 for Samsung Note 4 and 0.84 for LG G3. For classification, Random Forest classifier outperformed other classifiers with an F-measure value > 0.90 for all three phone models compared to other classifiers (F-measure value slightly below 0.90). Hence, in the rest of our evaluation, we use Random Forest along with the time-frequency domain feature set.

Moreover, MFCC features are more suitable for audio signals, sampled at a higher frequency than a motion sensor's sampling rate (restricted to approximately 200Hz by the Android platform in our setup). Thus, we believe time-frequency domain features capture more information about the speech vibration than the MFCC features, for a better classification accuracy. Therefore, we choose time-frequency domain feature set and Random Forest classifier. This result also affirms the premise of EchoVib to rely on effect of speech impact on motion sensors in vibration domain. Motion sensors, due to limited sampling rate, may not be comparable to the microphones in capturing audio features. However, the measured vibrations by the motion sensors can be unique as shown in our results and on the power spectrum in Figure 2.

5.3 Speaker Classification and Verification

We used 58 voice samples each from the ten speakers belonging to the Voxforge dataset, 100 samples each from the ten speakers belonging to the CMU_Arctic dataset, and 50 samples from each of the ten speakers in Amazon Mechanical Turk dataset. Since the samples in these datasets consist of complete sentences, we use the entirety of the recorded sensor reading for the voice sample to extract relevant speech features from the sensor readings. We used Galaxy S6, Note 4 and LG G3 for Voxforge and CMU_Arctic datasets, and Note 4 and LG G3 for Amazon Mechanical Turk dataset.

Device Fingerprint Verification: EchoVib theorizes that the speech vibrations captured in the vibration domain are unique per device as each smartphone's speaker and accelerometer behave in a distinctive manner [20, 28, 29, 33]. We empirically verified the unique fingerprints on mobile devices in terms of their vibration response. We collected the "training" vibration samples on an LG G3 phone using the Amazon Mechanical Turk dataset (58 single words spoken by 8 speakers) through the phone's loudspeaker for building the EchoVib speaker classifier. We also collected a set of vibration samples on another device with the same G3 phone model for "testing". The speaker classification accuracy, using 10-fold classification, was 0.114, which is lower than random guessing accuracy (0.125) showing that each authentication model is inherently tied to the device on which the training has been performed.

10-Speaker Classification: In 10-speaker classification, a multi-class classifier labels the test samples as one of the ten predefined

Table 2: Speaker classification using 10-fold cross validation and Random Forest classifier (Voxforge Dataset)

	10-Speaker Classification		
Devices	**Precision**	**Recall**	**F-measure**
Samsung Galaxy S8	0.97	0.97	0.97
Samsung Note 5	0.98	0.98	0.98
Samsung Galaxy S6	0.96	0.96	0.96
Samsung Note 4	0.91	0.91	0.91
LG G3	0.91	0.91	0.91

Table 3: Speaker verification using 10-fold cross validation and Random Forest classifier (Voxforge Dataset)

Devices	True Positive	False Positive	Precision	Recall	F-measure
Samsung Galaxy S8	0.99	0.03	0.99	0.99	0.99
Samsung Note 5	0.99	0.01	0.99	0.99	0.99
Samsung Galaxy S6	0.99	0.10	0.99	0.99	0.98
Samsung Note 4	0.98	0.13	0.98	0.98	0.98
LG G3	0.97	0.13	0.98	0.98	0.98

Table 4: Speaker classification using 10-fold cross validation and Random Forest classifier (CMU_ Arctic Dataset)

	10-Speaker Classification		
Devices	**Precision**	**Recall**	**F-measure**
Samsung Galaxy S8	0.91	0.91	0.91
Samsung Note 5	0.98	0.98	0.98
Samsung Galaxy S6	0.91	0.91	0.91
Samsung Note 4	0.91	0.91	0.91
LG G3	0.91	0.91	0.91

Table 5: Speaker verification using 10-fold cross validation and Random Forest classifier (CMU_ Arctic Dataset)

Devices	True Positive	False Positive	Precision	Recall	F-measure
Samsung Galaxy S8	0.99	0.05	0.99	0.99	0.99
Samsung Note 5	0.99	0.00	0.99	0.99	0.99
Samsung Galaxy S6	0.99	0.08	0.99	0.99	0.99
Samsung Note 4	0.98	0.11	0.98	0.98	0.98
LG G3	0.97	0.14	0.98	0.98	0.98

speakers, from our datasets. An example scenario could be a smart home device used by a family, where the inbuilt smart voice assistant's response is tailored as per family members' stored profiles. For each dataset, we consolidated the speech samples from all the speakers and performed speaker classification on the combined dataset. Our results from 10-speaker classification using 10-fold cross-validation method and Random Forest classifier are shown in Table 2 (Voxforge dataset), Table 4 (CMU_Arctic dataset), and Table 6 ("Amazon Mechanical Turk dataset"). For Voxforge dataset, EchoVib showed speaker classification accuracy with an F-measure > 0.99. CMU_Arctic dataset also provided a highly efficient F-measure score (averaging 0.91) for EchoVib implementation with Samsung Note 5 slightly outperforming other phone models.

Speaker Verification: Speaker verification uses a binary classifier to assign a test sample to a particular speaker. Speaker verification is performed regularly on more personal, non-shared devices such as smartphones, geared to respond only to voice commands of a single authorized entity. We use *true positive* and *false positive* metrics in addition to *precision*, *recall* and *F-measure* for speaker verification performance assessment. For the speaker verification task, in each dataset, we trained the classifier on speech samples for one speaker. We then tested the classifier on rest of the dataset that contains speech samples from the same speaker and that from the rest of the speakers in the dataset. This task is performed for each speaker in each dataset. From Table 3 and 5, we observe that EchoVib on Samsung Note 5 has better accuracy than other tested phone models with a high true positive value of 0.99 and a low false positive of

Table 6: Speaker classification using 10-fold cross validation and Random Forest classifier (Amazon Mechanical Turk Dataset)

	10-Speaker Classification		
Devices	**Precision**	**Recall**	**F-measure**
Samsung Note 4	0.95	0.95	0.95
LG G3	0.99	0.99	0.99

Table 7: Speaker verification using 10-fold cross validation and Random Forest classifier (Amazon Mechanical Turk Dataset)

Devices	True Positive	False Positive	Precision	Recall	F-measure
Samsung Note 4	0.98	0.08	0.99	0.99	0.99
LG G3	0.99	0.02	0.99	0.99	0.99

0.01. For CMU_Arctic dataset, Samsung Note 5 implementation of EchoVib outperformed with a true positive value of 0.99 and a false positive value of 0.00. Similar high values of true positives and low values of false positives are acquired by the other tested phones in the Amazon Mechanical Turk dataset as shown in Table 7.

6 ROBUSTNESS TO VOICE SYNTHESIS

Voice synthesis attack (Section 3.2), involves converting a *source speaker*'s voice sample into a voice sample closely resembling a *target speaker*'s voice. In our scenario, any entity that desires to use the system is presented with a *challenge-response setup* where a correct response is rewarded with authorized access or a *voice command setup* where a voice command is executed only when it is deemed to have originated from a verified speaker, following a *challenge-response setup*. The *challenge-response* game can be text-dependent, text-independent or text-prompted. In a text-dependent setting, a pre-determined passphrase needs to be spoken in an authorized user's voice, while in a text-independent scenario, a natural conversation is established that has to be in the authorized user's voice. Text-prompted setting requires the user to read aloud a given phrase, randomly generated by the system, that should match the voiceprint of a legitimate user.

Voice conversion attacks use statistical modeling techniques to build a model of the *target speaker*'s voice pattern that is then used to map *source speaker*'s voice pattern and regenerate speech in *target speaker*'s voice [19]. This method requires only 10-20 sentences in a *target speaker*'s voice, to convert any given voice sample into the *target speaker*'s voice, making these attacks very practical and easy to launch. In voice modeling attack, deep learning is used to build an acoustic model of *target speaker*'s voice, using a few samples containing only a minute of audio and generate any desired sentence in *target speaker*'s voice. Mukhopadhyay et al. [49] showed that voice conversion attacks have a very high success rate against many existing voice authentication services, hence we test EchoVib to see if it is able to distinguish between an actual legitimate user's voice and a fake legitimate user's voice that was generated using voice conversion or voice modeling.

6.1 Attack Setup

Festvox Voice Conversion for the CMU_Arctic and the Voxforge Datasets: We used Festvox voice conversion technique to create attacked samples for CMU_Arctic and Voxforge datasets. To synthesize the victim's voice, we used Festvox voice conversion technique [59], which produces the synthesized samples by mapping features of a source speaker's voice (i.e., the attacker) to a

Table 8: Evaluation using Random Forest classification against voice conversion attack. Over 90% morphed sample detection rate indicates EchoVib resilience against such attacks.

Devices	VoxForge Dataset		CMU Arctic Dataset	
	True Positive	False Positive	True Positive	False Positive
Samsung Galaxy S8	0.99	0.01	0.99	0.00
Samsung Note 5	0.98	0.02	0.99	0.01
Samsung Galaxy S6	1.00	0.00	0.99	0.01
Samsung Note 4	0.91	0.19	0.93	0.09
LG G3	0.95	0.07	0.85	0.15

target speaker's voice (i.e., the victim). Festvox employs acoustic-to-articulatory inversion mapping based on Gaussian Mixture Model followed by a spectral conversion between speakers for transforming the source speaker's voice to a target speaker's voice. This conversion is independent of the phonetic information of the speech. Festvox voice conversion had been used in the past to create voice samples [49] that have been used on Voxforge dataset to fool common voice authentication systems. To train the Festvox voice conversion tool, we used 50 CMU Arctic sentences spoken by both the source and the target speaker. We selected a male speaker from the CMU Arctic dataset as the attacker (bdl speaker), and as the victim, we selected five female and five male speakers from the CMU Arctic dataset, and ten male speakers from the Voxforge dataset. After training the system, we created 100 samples of the target speaker by feeding the source speaker's sample (bdl) to the trained system.

Lyrebird Voice Modeling for Amazon Mechanical Turk Dataset: Voice modeling technology by Lyrebird [46] aims to generate speech samples that closely resemble human speech and claim to sound more natural. This technology uses raw audio characteristics to build an acoustic model of a person by training on a set of voice samples from that person. Advanced machine learning algorithms, such as deep learning, are utilized to train the acoustic model and this model can produce speech samples mimicking a person's voice and transcribing to any desired phrase or sentence [25]. To create the synthesized voice for the Amazon Mechanical Turk dataset, we used the Lyrebird voice synthesis tool. Lyrebird relies on deep learning techniques to extract features of the speaker's voice from only a few minutes of the speech and produce naturally sounding samples in the speaker's voice.

Since Lyrebird has not officially published their API, we created the samples using Lyrebird web-based voice synthesizers. We published a HIT on Amazon Mechanical Turk and asked the participants to record their voice on Lyrebird website. The participants were compensated $3 for their effort. The participants were informed that the purpose of this task was to evaluate the performance of a speech synthesis tool and the samples would solely be used for the research. Note that Amazon Mechanical Turk IDs do not identify the users and therefore the collected audio does not carry any identification parameter. We created ten accounts on Lyrebird and asked each of the 10 Turk users to log in to one of these accounts and record their voice as instructed by Lyrebird web-based tool speaking 50 sentences displayed on the website. These recorded samples are used to train the Lyrebird speech synthesizer to create a model of the speaker's voice. Using this model, we generated 100 samples of the speaker's voice using Lyrebird text to speech tool.

User Movement and Noise Interference: Our experiments were performed in a lab environment with the phone placed on a flat surface or handheld. As mentioned in Section 5.1, user body movement such as walking can be accounted for by applying a suitable high pass frequency filter to the accelerometer output [66]. Noise present in the surroundings could affect the quality of the recorded speech, which in turn could affect the vibrations during the speech replay. A suitable white/brown noise filter could be applied to the recorded speech to remove such ambient sounds. Our Amazon Mechanical Turk voice dataset was recorded by Amazon Mechanical Turk workers in their own surroundings (home or office), different from our quiet lab environment. We were still able to achieve high speaker classification (over 90.00%) and acceptable false positive rate (as low as 1%) for voice conversion attacks.

Table 9: EchoVib Evaluation using Random Forest classification against voice modeling attack using Lyrebird. Morphed sample detection rate was over 85% indicating the resilience of EchoVib against these attacks.

Devices	Amazon Mechanical Turk Dataset	
	True Positive	False Positive
Samsung Note 4	0.87	0.13
LG G3	0.99	0.00

6.2 EchoVib vs. Voice Conversion Attack

We will now describe the performance of EchoVib based on its ability to distinguish between a legitimate user's voice and the fake voice sample, generated in the legitimate user's voice, by an attacker. We train our Random Forest classifier on legitimate voice samples of a speaker and evaluate the true positive and false positive rate of this classifier, when presented with fake voice samples of the targeted speaker. A low false positive rate would indicate that EchoVib was able to identify most of the fake voice samples as not belonging to the speaker and thus rejects them. This is a desired trait for our speaker authentication system. As in Section 5.3, we use the same datasets (58 samples per speaker for Voxforge, 100 samples per speaker for CMU_Arctic) for training our EchoVib verification model with Random Forest. We show the true positive and false positive rate for detecting the fake samples in Table 8. It shows that EchoVib can detect fake samples with a high true positive rate (> 0.90) on all phone models for Voxforge dataset and CMU_Arctic dataset. Since the fake samples were generated using a technique that is successful against voice authentication system [49], our results seem to promise an improved success rate in detecting the fake voice samples.

6.3 EchoVib vs. Voice Modeling Attack

We tested the Amazon Mechanical Turk dataset, created for generating morphed voice samples using Lyrebird deep learning technology, against EchoVib. Similar to the voice conversion attacks, we train a Random Forest classifier to identify a legitimate user's voice and then test the morphed voice samples against the classifier. A high true positive rate for morphed samples would indicate that EchoVib was successful in identifying (and thus rejecting) morphed samples, not matching with any of the stored users' phonatory vibration patterns. The results (Table 9) indicate that morphed samples were identified correctly with a high true positive rate (≥ 0.85) for all the implementations of EchoVib. Since, voice modeling techniques aim to reproduce any speech sample from a limited set of voice samples of a user, in a natural voice, our results show that these tools cannot be exploited to fool EchoVib into accepting an artificially generated voice sample as a legitimate user.

Table 10: Speaker classification using 10-fold cross validation and Random Forest classifier (Voxforge Dataset) with half-a-second and one-second echoed speech

	10-Speaker Classification		
Devices	**Precision**	**Recall**	**F-measure**
Samsung Note 4 (Half-a-second)	0.71	0.72	0.71
LG G3 (Half-a-second)	0.76	0.76	0.76
Samsung Note 4 (One-second)	**0.83**	**0.83**	**0.83**
LG G3 (One-second)	0.82	0.82	0.82

6.4 Partial Speech Replay for Speaker Classification

To evaluate how many words of the user's speech to be played/echoed back by EchoVib, we pick up phonatory vibrations from partially replayed speech. We played back half-a-second (around one word) and one-second recorded speech (around two words) using Voxforge dataset. The 10-speaker classification results using 10-fold cross-validation and Random Forest classifier are shown in Table 10. Although the performance degrades with shorter echoed speech, our system can still obtain acceptable classification accuracy with only one or two echoed words.

7 ENTROPY ESTIMATION

We employ three metrics, min-entropy, β-success rate, and α-guesswork [21], to measure the strength of the voice signature in EchoVib. Min-entropy estimates the worst case security whereby an attacker attempts to guess the most likely vibration signature before giving up, while β-success rate measures the expected success for the attacker limited to β guesses. α-guesswork estimates the security against the attacker that aims to compromise a certain proportion $\alpha \leq 1$ of signatures in EchoVib before quitting. Specifically, we calculated the min-entropy (H_∞), β-success rate (λ_β), and α-guesswork (G_α) of each axis using the accelerometer readings of the echoed speech samples for each speaker. We used the three phones (Galaxy S6, Note 4 and LG G3), and β and α are set to 3/30 and 0.1/0.5, respectively. For each accelerometer axis, we report the min-entropy, β-success rate, and α-guesswork estimates (in bits) in Table 11. We note that the estimated α-guesswork values only have a slight change across different α values. We believe this might be caused by the distribution bias inherently existent within our dataset, along with the small-scale data volume collected in our work to demonstrate the feasibility of the approach. Similar to [62], we also report the security estimate in bits, by combining them across the three axes (Table 11). As the three axes may have relative independent distributions, the total min-entropy, β-success rate, or α-guesswork will be additive across the three axes.

We can compare our estimates with those of the other authentication factors. In terms of password authentication, Wang et al. [63] studied the strength of human-generated passwords and found that the human-chosen 4-digit PINs offered an equivalent of 6.62 bits of security against an online guessing attacker limited to 30 guesses (β-success rate), while 6-digit PINs offered 7.24 bits of security. Against an offline guessing attacker looking for 50% success rate (α-guesswork), the offered security is 8.41 bits (4-digit PINs) and 13.21 bits (6-digit PINs). Vibration signatures in EchoVib offer around 15-30 bits of security against attackers limited in guessing attempts (online attacks) on different phone models. Against an offline attacker aiming for 50% success rate, the vibration signatures from the accelerometer offer 16-35 bits of security. This indicates that the raw speech vibration signatures are at least as secure as 6-digit PINs, and the security could be increased by combining it over different devices (leveraging device fingerprinting, given the entropy measures across different devices are different).

Using other traditional biometric authentication factors, Adler et al. [12] reported that the facial image biometrics has 47 bits of security. Inthavisas et al. [40] argued that in a compromised scenario where the attacker has access to the biometrics template, the estimated entropy of their proposed speech cryptographic key regeneration scheme is between 18-30 bits. The security of spoken passwords and fingerprints is reported to be 46 bits [48] and 69 bits [26], respectively. In addition, Sadeghi et al. [55] showed that the maximum achievable security strength of five state-of-the-art electroencephalogram (EEG) based authentication systems is 83 bits using Naive Bayes Classifier and 36 bits using SVM classifier.

Although some of the aforementioned biometric studies report higher entropies than our estimates, this comparison may not be accurate due to the differences in the evaluation settings, different subjects and, most importantly, the size of the datasets. An accurate entropy estimation requires large scale datasets from thousands of users in biometric systems. Our dataset is limited due to the exploratory nature of our work introducing a new authentication paradigm. To further improve the security of EchoVib against guessing attacks, we could also collect and combine the user's speech samples from different devices. This scheme would allow us to capture the unique device fingerprint (background replay scenario in Appendix Figure A.3). Future work should explore more accurate entropy estimations using larger datasets. We emphasize that our approach can also be seamlessly integrated with the traditional voice biometrics systems to improve their security against guessing attacks. We have demonstrated that our approach can effectively defeat targeted voice synthesis attacks in the audio domain against which voice biometrics are vulnerable.

8 DISCUSSION AND FUTURE WORK

Possible Implementation using Current Hardware: Since EchoVib requires a replay/playback of the user's response in a *challenge-response setup* we envision two settings in which EchoVib could be implemented (Appendix Figure A.3). Foreground Replay is the setup that we use in our experiments where the user interface already has an on-board loudspeaker and accelerometer (e.g., a user-facing smartphone) and the capturing of phonatory vibrations is performed by replaying the user's response in the foreground (i.e., on the user's device itself) via the on-board loudspeaker while its on-board accelerometer records the resulting vibrations. Such a replay mechanism is suitable for the voice authentication applications that secure a smartphone or an app on the smartphone.

The echoing speech functionality to capture speech vibrations can also be transparently implemented on a supplementary entity that has a loudspeaker to replay the user's speech sample and an accelerometer to capture the resulting vibrations. Such a background replay mechanism can be feasible in practice, especially given that many voice authentication and audio processing systems already outsource their detection tasks to a cloud system that can acquire

Table 11: Min-entropy (H_∞), β-success rate (λ_β), and α-guesswork (G_α) for the recorded sensor readings from different phone models

	Samsung Galaxy S6					**Samsung Note 4**					**LG G3**				
	H_∞	λ_3	λ_{30}	$G_{0.1}$	$G_{0.5}$	H_∞	λ_3	λ_{30}	$G_{0.1}$	$G_{0.5}$	H_∞	λ_3	λ_{30}	$G_{0.1}$	$G_{0.5}$
X axis	4.652	4.665	5.132	5.032	5.033	3.577	3.606	4.941	4.076	4.078	9.561	9.638	10.139	11.642	11.741
Y axis	4.684	4.659	5.120	5.003	5.005	3.439	3.469	4.913	3.829	3.830	9.380	9.461	9.991	11.380	11.486
Z axis	5.523	5.533	5.708	6.134	6.135	4.440	4.466	5.196	5.236	5.237	10.536	10.620	11.009	12.397	12.567
Sum across axes	14.859	14.857	**15.960**	16.169	**16.173**	11.456	11.541	**15.050**	13.141	**13.145**	29.477	29.719	**31.139**	35.419	**35.794**

the necessary resources for EchoVib. However, this implementation may introduce a time-lag undesirable from a user's perspective.

User Experience: In the partial speech replay experiment, EchoVib produces acceptable performance on partial speech containing only one or two words. This removes the tedious training requirement from an ease of use perspective, when the user does not need to speak long sentences repeatedly. The EchoVib design allows authentication in the foreground (in the user's vicinity) or in the background (off-site dedicated system) depending on the use-case (Section 8). With the feasibility of short sentences and even partial speech replays along with the ability to perform authentication in the background, EchoVib limits the loudness issue that may arise when trying to get maximum response from the accelerometer, during the speech playback in the authentication phase.

Potential Sophisticated Vibration Morphing Attack: Phonatory vibrations measured in the vibration domain form the backbone of the EchoVib model. A possible attack vector could be created by attempting to generate vibrations that mimic the phonatory vibrations of a legitimate user. However, the phonatory vibrations captured by the accelerometer are mapped non-linearly from the speech signal captured by a microphone. In particular, these vibrations, referred to as the *phonetic vibrations* in [57], are indicative of the amplitude of the fundamental frequency of the voice and also correlated to the pitch of the voice (except for basses).

To mimic the vibration patterns, the attack would have to measure the amplitude of the fundamental frequency of the original voice and adjust the amplitude of the synthesized voice sample's fundamental frequency. In addition, it would have to imitate the frequency features of the targeted voice. Since the fundamental frequency's amplitude also causes the vibrations, in addition to the low frequency components of the speech sample, just enhancing the low frequency components in the synthesized voice sample may not be enough to fool EchoVib. Also, the vibration source affects the vibration signal so the attacker also needs to closely copy the device's loudspeaker characteristics and the accelerometer fingerprint during the targeted voice playback.

We also argue against an attacker that may attempt to mimic the vibrations by obtaining several voice samples of a potential victim and trying to generate a new voice sample containing same vibration features as the victim's samples. We believe that it may be a hard task to replicate those vibration features either manually or automatically. We do not consider a replay attack where the attacker replays a prior eavesdropped sample of the victim's voice. The attacker can extract vibration features from the victim's voice sample and use a conversion technique (FestVox Transform [19] in the vibration domain) to generate a new vibration signature having similar features as the victim's voice's vibration signature.

Nonetheless, there still remains the task of mapping the converted vibration signature back to the audio domain (as the authentication system still expects an audio input) that we believe to be a difficult, if not impossible, task, given the vibration features are not the same as audio features that forms the audio signal. A future investigation into the possibility and accuracy of such sophisticated attacks might confirm our insights. Nevertheless, our proposed system serves its desired purpose of considerably raising the bar against the most powerful audio domain attacks that are effective against audio domain voice authentication systems.

Adversarial Machine Learning on Speaker Verification: Recently, adversarial machine learning attacks on speech and speaker verification systems have drawn much attention [10, 23, 43]. By adding perturbations into the users' speech, the crafted speech can be falsely accepted as some adversary-desired speaker. Thus, a machine learning classifier trained only in the audio domain can be fooled into falsely accepting an unauthorized speaker. Some of the perturbation generation techniques involve time-domain inversion, random phase generation, high frequency addition or time scaling, or a gradient estimation technique to construct a spoofed voice sample. These features are different from the features associated with vibration features such as the pitch range and distribution, low range frequencies, and pitch change over time.

EchoVib avoids the vulnerabilities exploited by these attacks, as perturbations applied only in the audio domain would not completely translate to the vibration domain, unless the perturbations are crafted to be effective in both audio and vibration domains. Further, most of these attacks seem to require a large number of queries to succeed, which is impractical as any burst query attacks can be easily blocked by throttling (web services consistently thwart online password guesses by this mechanism). These attacks could also be considered less powerful than the voice synthesis/morphing attack that we evaluated EchoVib against which need to be trained on a particular speaker's speech pattern for morphing.

Limited Sampling Rate and Sensor Fusion: Limited sampling rate of the motion sensors on smartphones was a challenge but even in this limited capability, speech footprints recorded in vibration domain have unique features that allow for an accurate speaker classification. We expect the performance of EchoVib to improve in a more natural environment if more sophisticated motion sensors are utilized for authentication in vibration domain. In addition, multiple sensors can possibly be used to artificially increase the combined sampling rate, overcoming the imposed restriction on the sampling rate from the Operating System.

Hardware Characteristics: The smartphone devices used in our experiments were from multiple models, having different MEMS motion sensor chips. However, the accelerometer specifications of these chips were similar, having a full-scale acceleration range of ±2/±4/±8/±16g. The smartphone body for the newer devices like Galaxy S8, Note 5, and S6 have a glass back and front and an aluminum frame. The older devices like LG G3 and Note 4, however have a plastic back and a glass front. LG G3 and Note 4 have loudspeaker placed on the lower back of the device while

Galaxy S6, S8, and Note 5 have bottom firing loudspeakers. While it is possible that the loudspeakers firing on the back may produce stronger vibrations (as the phones were placed on their back), the newer phones have better speakers with more metallic/glass body. Overall, we did not notice any major difference between the results due to hardware differences.

Acoustic Interference from External Sources: In our work, we have considered human voice transmission from the user to the recording device (the smartphone) via aerial routes. The speech could also be transmitted through walls, roofs or ground, and potentially interfere with the accelerometer. However due to attenuation, the energy contained in the acoustic wave is dissipated quickly. Attenuation is directly proportional to the frequency and the distance traveled within the propagation medium. Given that speech transmission via air is the shortest distance to the sensors, the least amount of attenuation would occur in this scenario. Speech transmission through walls, roofs or ground, involves travel through the air followed by the solid medium. Anand et al. [14] showed that human speech does not contain enough energy to travel by the ground to a smartphone placed on a table and significantly impact the accelerometer sensor.

9 CONCLUSION

In this paper, we proposed a novel voice-based authentication system EchoVib, showing that vibrations generated from a person's speech and captured via the accelerometer on a smartphone are unique and can be used for identifying thereby rejecting *voice synthesis attack*. The user authentication accuracies, combined with the *voice synthesis attack* identification accuracies lead us to believe that EchoVib can improve on the security of current voice authentication systems against voice synthesis attacks without degrading the authentication accuracy. Since we implemented the proposed scheme on a smartphone, our proposed authentication system shows a potential to be implemented with minimum hardware requirements (a loudspeaker and an accelerometer).

10 ACKNOWLEDGMENTS

We would like to thank our shepherd, Dr. Ding Wang, and the anonymous reviewers for their insightful comments and constructive feedback on the paper. This work was partially supported by the following NSF grants: CNS-1714807, CNS-1526524, CNS-1547350, CNS-2030501, CCF-2028876, CNS-1820624, CNS-1814590 and the ARO grant: W911NF-19-1-0405.

REFERENCES

[1] [n.d.]. Automatic Speaker Verification Spoofing and Countermeasures Challenge. http://www.asvspoof.org/.

[2] [n.d.]. Barclays rolls out voice biometrics for phone banking. https://www.finextra.com/newsarticle/29245/barclays-rolls-out-voice-biometrics-for-phone-banking Accessed: 06/25/2018.

[3] [n.d.]. HSBC rolls out voice and touch ID security for bank customers. https://www.theguardian.com/business/2016/feb/19/hsbc-rolls-out-voice-touch-id-security-bank-customers Accessed: 06/25/2018.

[4] [n.d.]. Pdv-100 portable digital vibrometer: Vibration sensor for mobile use. https://www.polytec.com/us/vibrometry/products/single-point-vibrometers/pdv-100-portable-digital-vibrometer/. Accessed: 12/13/2018.

[5] [n.d.]. Set your device to automatically unlock. https://support.google.com/nexus/answer/6093922?hl=en Accessed: 06/25/2018.

[6] [n.d.]. Unlocking Bixby with a Voice Password. https://us.community.samsung.com/t5/How-To-s/Unlocking-Bixby-with-a-Voice-Password/td-p/138591 Accessed: 06/25/2018.

[7] [n.d.]. Voice Verification. https://www.wellsfargo.com/privacy-security/voice-verification/ Accessed: 06/25/2018.

[8] [n.d.]. Voxforge. http://www.voxforge.org/.

[9] 2020. Void: A fast and light voice liveness detection system. In *USENIX*.

[10] Hadi Abdullah, Washington Garcia, Christian Peeters, Patrick Traynor, Kevin R. B. Butler, and Joseph Wilson. 2019. Practical Hidden Voice Attacks against Speech and Speaker Recognition Systems. *CoRR* abs/1904.05734 (2019).

[11] Andre G Adami, Radu Mihaescu, Douglas A Reynolds, and John J Godfrey. 2003. Modeling prosodic dynamics for speaker recognition. In *ICASSP*, Vol. 4. IEEE.

[12] A. Adler, R. Youmaran, and S. Loyka. 2006. Towards a Measure of Biometric Information. In *2006 Canadian Conference on Electrical and Computer Engineering*. 210–213.

[13] Talal B Amin, James S German, and Pina Marziliano. 2013. Detecting voice disguise from speech variability: Analysis of three glottal and vocal tract measures. In *Proceedings of Meetings on Acoustics 166ASA*, Vol. 20. ASA.

[14] S Abhishek Anand and Nitesh Saxena. 2018. Speechless: Analyzing the Threat to Speech Privacy from Smartphone Motion Sensors. In *IEEE S&P*.

[15] Machine Learning Group at the University of Waikato. 2017. Weka 3: Data Mining Software in Java. https://www.cs.waikato.ac.nz/ml/weka/index.html.

[16] Les Atlas and Shihab A Shamma. 2003. Joint acoustic and modulation frequency. *EURASIP Journal on Applied Signal Processing* 2003 (2003).

[17] R. J. Baken. 1987. *Clinical Measurement of Speech and Voice*. 177.

[18] Katarina Bartkova, David Le Gac, Delphine Charlet, and Denis Jouvet. 2002. Prosodic parameter for speaker identification. In *ICSLP*.

[19] Alan W Black and Arthur Toth. 2005. TRANSFORM: flexible voice synthesis through articulatory voice transformation. http://www.festvox.org/transform/

[20] Hristo Bojinov, Yan Michalevsky, Gabi Nakibly, and Dan Boneh. 2014. Mobile Device Identification via Sensor Fingerprinting. *ArXiv* abs/1408.1416 (2014).

[21] Joseph Bonneau. 2012. The science of guessing: analyzing an anonymized corpus of 70 million passwords. In *2012 IEEE Symposium on Security and Privacy*. IEEE, 538–552.

[22] Simon Castro, Robert Dean, Grant Roth, George T Flowers, and Brian Grantham. 2007. Influence of acoustic noise on the dynamic performance of MEMS gyroscopes. In *ASME*.

[23] Guangke Chen, Sen Chen, Lingling Fan, Xiaoning Du, Zhe Zhao, Fu Song, and Yang Liu. 2019. Who is Real Bob? Adversarial Attacks on Speaker Recognition Systems. https://arxiv.org/abs/1911.01840

[24] Si Chen, Kui Ren, Sixu Piao, Cong Wang, Qian Wang, Jian Weng, Lu Su, and Aziz Mohaisen. 2017. You can hear but you cannot steal: Defending against voice impersonation attacks on smartphones. In *ICDCS*.

[25] Thomas Claburn. 2017. Lyrebird steals your voice to make you say things you didn't – and we hate this future. https://www.theregister.co.uk/2017/04/24/voice_stealing_lyrebird/

[26] T. Charles Clancy, Negar Kiyavash, and Dennis J. Lin. 2003. Secure Smartcard-based Fingerprint Authentication *(WBMA '03)*.

[27] Robert F. Coleman. 1988. Comparison of microphone and neck-mounted accelerometer monitoring of the performing voice. *Journal of Voice* 2, 3 (1988).

[28] Anupam Das, Nikita Borisov, and Matthew Caesar. 2014. Do you hear what I hear?: Fingerprinting smart devices through embedded acoustic components. In *CCS*. ACM.

[29] Anupam Das, Nikita Borisov, and Matthew Caesar. 2016. Tracking Mobile Web Users Through Motion Sensors: Attacks and Defenses. In *NDSS*.

[30] Phillip L De Leon, Michael Pucher, Junichi Yamagishi, Inma Hernaez, and Ibon Saratxaga. 2012. Evaluation of speaker verification security and detection of HMM-based synthetic speech. *IEEE TASLP* 20, 8 (2012).

[31] Robert Neal Dean, Simon Thomas Castro, George T Flowers, Grant Roth, Anwar Ahmed, Alan Scottedward Hodel, Brian Eugene Grantham, David Allen Bittle, and James P Brunsch. 2011. A characterization of the performance of a MEMS gyroscope in acoustically harsh environments. *IEEE TIE* 58, 7 (2011).

[32] Robert N Dean, George T Flowers, A Scotte Hodel, Grant Roth, Simon Castro, Ran Zhou, Alfonso Moreira, Anwar Ahmed, Rifki Rifki, Brian E Grantham, et al. 2007. On the degradation of MEMS gyroscope performance in the presence of high power acoustic noise. In *ISIE 2007*. IEEE.

[33] Sanorita Dey, Nirupam Roy, Wenyuan Xu, Romit Choudhury, and Srihari Nelakuditi. 2014. AccelPrint: Imperfections of Accelerometers Make Smartphones Trackable. In *NDSS*.

[34] Megan Ellis. 2018. How to Lock/Unlock an Android Phone With Your Voice Using Google Assistant. https://www.makeuseof.com/tag/lock-unlock-android-phone-voice-google-assistant/

[35] Huan Feng, Kassem Fawaz, and Kang G Shin. 2017. Continuous authentication for voice assistants. In *MobiCom*.

[36] Rosa González Hautamäki, Tomi Kinnunen, Ville Hautamäki, and Anne-Maria Laukkanen. 2015. Automatic versus human speaker verification: The case of voice mimicry. *Speech Communication* 72 (2015).

[37] Rosa González Hautamäki, Tomi Kinnunen, Ville Hautamäki, Timo Leino, and Anne-Maria Laukkanen. 2013. I-vectors meet imitators: on vulnerability of speaker verification systems against voice mimicry.. In *Interspeech*. Citeseer.

[38] Pindrop Security Inc. 2018. BUSINESS USE OF VOICE TECHNOLOGY TO TRIPLE IN NEXT 12 MONTHS ACCORDING TO PINDROP'S LATEST SURVEY. https://www.pindrop.com/press-release/business-use-voice-technology-triple-next-12-months-according-pindrops-latest-survey/

[39] Sensory Inc. 2016. AppLock from Sensory Keeps Apps Safe with Face and Voice Biometrics. http://www.sensory.com/applock-sensory-keeps-apps-safe-face-voice-biometrics/

[40] K. Inthavisas and D. Lopresti. 2012. Secure speech biometric templates for user authentication. *IET Biometrics* 1, 1 (2012), 46–54.

[41] Alan W Black John Kominek. 2003. CMU ARCTIC Databases for Speech Synthesis. http://festvox.org/cmu_arctic/cmu_arctic_report.pdf

[42] Tomi Kinnunen, Bingjun Zhang, Jia Zhu, and Ye Wang. 2007. Speaker verification with adaptive spectral subband centroids. In *International Conference on Biometrics*.

[43] Felix Kreuk, Yossi Adi, Moustapha Cissé, and Joseph Keshet. 2018. Fooling End-to-end Speaker Verification by Adversarial Examples. *CoRR* abs/1801.03339 (2018).

[44] Jeanne Lee. 2016. More Banks Turn to Biometrics to Keep an Eye on Security. https://www.nerdwallet.com/blog/banking/biometrics-when-your-bank-scans-your-voice-face-or-eyes/

[45] Johan Lindberg and Mats Blomberg. 1999. Vulnerability in speaker verification-a study of technical impostor techniques. In *EUROSPEECH*.

[46] Lyrebird. [n.d.]. https://lyrebird.ai/

[47] Yan Michalevsky, Dan Boneh, and Gabi Nakibly. 2014. Gyrophone: Recognizing Speech from Gyroscope Signals.. In *USENIX*.

[48] F. Monrose, M. K. Reiter, Qi Li, and S. Wetzel. 2001. Cryptographic key generation from voice. In *Proceedings 2001 IEEE Symposium on Security and Privacy. S P 2001.*

[49] Dibya Mukhopadhyay, Maliheh Shirvanian, and Nitesh Saxena. 2015. All your voices are belong to us: Stealing voices to fool humans and machines. In *ESORICS*.

[50] K Sri Rama Murty and Bayya Yegnanarayana. 2006. Combining evidence from residual phase and MFCC features for speaker recognition. *IEEE signal processing letters* 13, 1 (2006).

[51] Linda Musthaler. 2013. Authentify xFA provides simple, secure primary authentication using digital certificates and voice biometrics. https://www.networkworld.com/article/2168093/security/authentify-xfa-provides-simple-secure-primary-authentication-using-digital-certificates-an.html

[52] Kenneth Olmstead. 2017. Nearly half of Americans use digital voice assistants, mostly on their smartphones. http://www.pewresearch.org/fact-tank/2017/12/12/nearly-half-of-americans-use-digital-voice-assistants-mostly-on-their-smartphones/

[53] Jack Purcher. 2015. Apple Invents a Simple Voice Command to unlock your iPhone. http://www.patentlyapple.com/patently-apple/2015/04/apple-invents-a-simple-voice-command-to-unlock-your-iphone.html

[54] Douglas A. Reynolds, Thomas F. Quatieri, and Robert B. Dunn. 2000. Speaker Verification Using Adapted Gaussian Mixture Models. *Digit. Signal Process.* 10, 1 (Jan. 2000).

[55] K. Sadeghi, A. Banerjee, J. Sohankar, and S. K. S. Gupta. 2017. Geometrical Analysis of Machine Learning Security in Biometric Authentication Systems. In *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*. 309–314.

[56] Schröder, Marc and Trouvain, Jürgen. 2003. The German Text-to-Speech Synthesis System MARY: A Tool for Research, Development and Teaching. *International Journal of Speech Technology* 6 (2003).

[57] Johan Sundberg. 1992. Phonatory Vibrations in Singers: A Critical Review. *Music Perception: An Interdisciplinary Journal* 9, 3 (1992).

[58] I. R. Titze. 1994. *Principles of Voice Production.* Prentice Hall (NCVS.org).

[59] Tomoki Toda, Alan W Black, and Keiichi Tokuda. 2004. Acoustic-to-articulatory inversion mapping with gaussian mixture model.. In *INTERSPEECH*.

[60] Timothy Trippel, Ofir Weisse, Wenyuan Xu, Peter Honeyman, and Kevin Fu. 2017. WALNUT: Waging doubt on the integrity of mems accelerometers with acoustic injection attacks. In *EuroS&P*. IEEE.

[61] Robbie Vogt and Sridha Sridharan. 2008. Explicit Modelling of Session Variability for Speaker Verification. *Comput. Speech Lang.* 22, 1 (Jan. 2008).

[62] K. Wallace, K. Moran, E. Novak, G. Zhou, and K. Sun. 2016. Toward Sensor-Based Random Number Generation for Mobile and IoT Devices. *IEEE Internet of Things Journal* 3, 6 (2016), 1189–1201.

[63] Ding Wang, Qianchen Gu, Xinyi Huang, and Ping Wang. 2017. Understanding Human-Chosen PINs: Characteristics, Distribution and Security *(ASIA CCS '17)*.

[64] Z. Wu, P. L. De Leon, C. Demiroglu, A. Khodabakhsh, S. King, Z. Ling, D. Saito, B. Stewart, T. Toda, M. Wester, and J. Yamagishi. 2016. Anti-Spoofing for Text-Independent Speaker Verification: An Initial Database, Comparison of Countermeasures, and Human Performance. *TASLP* 24, 4 (2016).

[65] Zhizheng Wu, Xiong Xiao, Eng Siong Chng, and Haizhou Li. 2013. Synthetic speech detection using temporal modulation feature. In *ICASSP*. IEEE.

[66] Li Zhang, Parth H Pathak, Muchen Wu, Yixin Zhao, and Prasant Mohapatra. 2015. Accelword: Energy efficient hotword detection through accelerometer. In *MobiSys*. ACM.

[67] Linghan Zhang, Sheng Tan, and Jie Yang. 2017. Hearing Your Voice is Not Enough: An Articulatory Gesture Based Liveness Detection for Voice Authentication. In *ACM SIGSAC CCS*.

[68] Linghan Zhang, Sheng Tan, Jie Yang, and Yingying Chen. 2016. Voicelive: A phoneme localization based liveness detection for voice authentication on smartphones. In *ACM SIGSAC CCS*.

A. APPENDIX

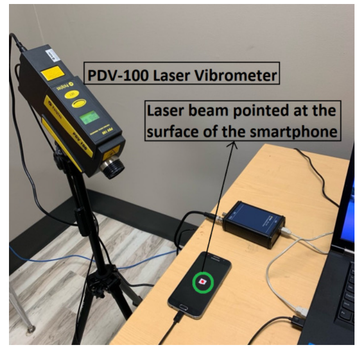

Figure A.1: The setup for surface vibration measurement using laser vibrometer.

Table A.1: Time-frequency features calculated from accelerometer readings on X, Y and Z axis

Time Domain
Minimum; Maximum; Median; Variance; Standard deviation; Range
CV: ratio of standard deviation and mean times 100
Skewness (3rd moment); Kurtosis (4th moment)
Q1, Q2, Q3: first, second and third quartiles
Inter Quartile Range: difference between the Q3 and Q1
Mean Crossing Rate: measures the number of times the signal crosses the mean value
Absolute Area: the area under the absolute values of accelerometer signal
Total Absolute Area: sum of Absolute Area of all three axis
Total Strength: the signal magnitude of all accelerometer signal of three axis averaged of all three axes
Frequency Domain
Energy
Power Spectral Entropy
Frequency Ratio: ratio of highest magnitude FFT coefficient to sum of magnitude of all FFT coefficients

Table A.2: Configuration details for Weka workbench Classification Algorithms

Classification Algorithm	Configuration
Simple Logistic	Number of boosting iterations = 0 Max. number of boosting iterations = 500 Heuristic stop = 50 Output data precision = 2 decimal places
Support Vector Machine using Sequential minimal optimization (SMO)	Complexity parameter = 1.0 Tolerance parameter = 0.001 Random seed = 1 Kernal= PolyKernal with cache size = 250007 and exponent = 1 Calibrator = Logistic with the log-likelihood ridge value = 1.0e-8, precision = 4 Round-off error epsilon = 1.0e-12 Output data precision = 2 decimal places
Random Forest	Max depth = unlimited Number of trees in the forest = 100 Bag size as percent = 100 Random number seed = 1 Number of threads used = 1 Output data precision = 2 decimal places
Random Tree	Max depth = unlimited Min. total weight of leaf instance = 1.0 Min. variance proportion = 0.001 Random number seed = 1 Output data precision = 2 decimal places
J48	Min. leaf instances = 2 Confidence factor = 0.25 Random number seed = 1 Output data precision = 2 decimal points

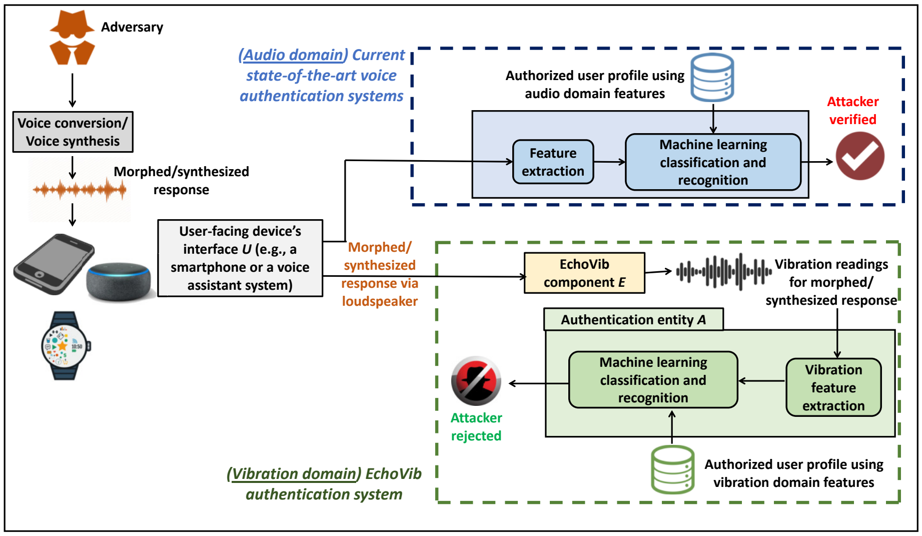

Figure A.2: EchoVib authentication model in an adversarial setup (voice synthesis attacks) where EchoVib uses vibration domain to correctly reject the morphed voice samples while current state-of-the-art voice authentication systems incorrectly accept and authorize the morphed speech (as shown in [49])

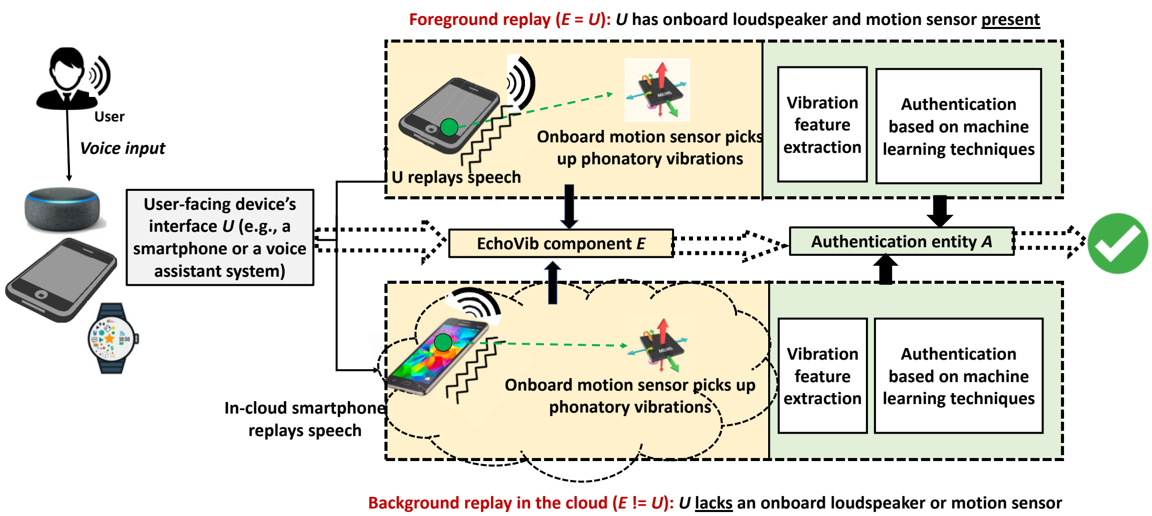

Figure A.3: EchoVib high-level overview of the benign setting depicting different use cases with foreground and background replay. Our implementation reported in this paper uses smartphone with its on-board loudspeaker and embedded motion sensor (accelerometer), which can be used in either foreground or background replay settings.