CIS 5603

Project Report

Zhengkun Ye

05/03/2022

Abstract - ML/DL method-based User Identification using touches on Smartwatches

User identification is a fundamental and pervasive aspect of modern mobile device usage, both as a means of maintaining security and personalized services. Verifying oneself is necessary to gain access to smartphones, bank accounts, and customized news feeds; information and resources which must be available on demand. As such, repeated acts of authentication can grow tedious and consume unnecessarily long portions of daily routines involving mobile devices. Studies on cellphone addiction suggest that user identification procedures encompass up to 9% of daily usage time, with related inquiries showing strong interest in more convenient practices. Many existing Identification/Authentication approaches require active user input, specialized sensing hardware, or personally identifiable information such as fingerprints or face scans. In this AI course project, I propose a low-effort identification scheme that validates the user by sensing finger press via commodity microphones and speakers.

1. Introduction

User identification is integral to many everyday activities. Many services aim to make the identification process as quick and easy as possible. The project idea was coming from a paper named EchoLock which published in ASIA CCS 2020. EchoLock can serves as a novel technique leveraging acoustic sensing of structure-borne sound to measure biometric characteristics and can identify users with over 94%accuracy, without requiring any active user input. The proposed method in this report is to utilize acoustic signal and vibration data collected by Android Smartwatches to build up a cross-domain system capable of identifying different users. Noise and vibration measurements are impacted by the presence of obstacles, such as people. Distinct individuals may produce unique reflections on sound waves and can potentially be identified based on these properties. There will be no specialized hardware requirements. Potential applications of this proposed system include unlocking edge devices, securing messages, reducing dependency on tethered smartphones for user interaction, and providing more convenient interactions (e.g., vs. typing on small watch screens).

2. Related Work

2.1 Audio-Domain Voice Authentication. The conventional user authentication method invented for the voice access system is mainly to train a model and then extract the voice characteristics in the audio domain to identify the user [6, 8, 9, 14, 17, 18]. Voice authentication systems distinguish an individual's voice feature, for instance, Mel-Frequency Cepstral Coefficients (MFCCs) [13] and Spectral Subband Centroids (SSCs) [10] describe a voice's timbre and vocal-tract resonances and are used extensively as unique short-term spectral based features to differentiate people's voices. Likewise, in order to extract details of the user's voice for speaker recognition, researchers have proposed Prosodic parameter to be used in a speaker identification system [4], prosodic features such as intonation, stress, and rhythm can be utilized to formalize the context and meaning of words. In addition, the modulation frequency [3] can present practical information of a signal's energy transitions for user identification process. Nevertheless, only relying upon the audio-domain features is endangered to acoustic-based attacks, adversaries can synthesize voice [11] to trick edge devices into executing commands to leak secrets or to modify critical information. For the particular case of voice assistants, they all are vulnerable to simple replay attacks.

2.2 Vibration Domain Speech Recognition. Motion sensor are able to capture acoustic sounds. Accelerometer and gyroscope are widely used in edge devices and smart speakers (e.g.,Amazon Echo, Google Nest, HomePod) to meet the needs of various applications. They consist of microelectromechanical systems (MEMS) structures [1], which is easy to be affected by sound and noises. Moreover, the WALNUT [16] simulated the physical attack of sound injection on the accelerometer and showed that the output of the sensor was affected by acoustic interference. In addition, the sound from an external speaker has been proven to impact the motion sensor. By way of illustration, Gyrophone [12] show that gyroscopes can be used in attacks to measure the sound of speakers that share the same surface as the sensor/smart phone, and compromise voice privacy (such as voice content) through classification. Accelword [19] demonstrated the smartphone's accelerometer capability to extract hotwords (e.g., Okay Google) from an individual's voice. EchoVib [2] verifies the speech played back by the device's loudspeaker and further performs user verification by examining unique effect on built-in motion sensors.

2.3 Voice Authentication Using Second Factors. Voice authentication is the process of identifying the user based on their voice, which is also a promising solution that can aid tedious traditional authentication systems with remarkable accuracies. Authentication with second factor, combining the use of a password and a token device. A traditional two-factor authentication (TFA) scheme requires the user to enter a password and copy a short, random, and one-time verification code from the token over to the authentication process. This improves security because the attacker needs not only the user's password but also the current verification code to hack into the user's account. Two Microphone Authentication (2MA) [5] systems take advantage of the presence of multiple microphones being present in an ecosystem to authenticate the source of a command. 2MA authentication framework demonstrate that such a construction works using independent devices (e.g., a mobile phone and a voice assistant) increase the effort required by an attacker to inject such commands successfully. Moreover, Listening-Watch [15] has been proposed as a low-effort two-factor authentication system using speech signals based on a wearable device and active sounds that is resistant to co-located and remote attack. Furthermore, VAuth[7], a system that requires the user to wear an additional device that is in continuous contact with user's body provides continuous authentication for voice assistants with high accuracy and very low false positive rate.

Most traditional voice authentication schemes involve training on the user's voice features for speaker classification. The performance may depend on the user's vocal features or cultural background and requires rigorous training to obtain an inch-perfect outcome. There is no theoretical guarantee that they provide good security in general. Meanwhile, a dedicated training process with motion sensor data can result in higher communication overhead. Many detailed sensing methods may require additional hardware or modifications to the smartwatch. In addition, most user identification credentials are behavioral based, not physiological. For the approach of the proposed system, we leverage acoustic sensing and motion sensing to detect behavioral and physiological user information using only built-in hardware.

3. Challenges

3.1 Human Variability

- Differences in human wrists, hands, and fingers are very subtle and difficult to quantify using limited sensors
- Users may have inconsistencies in how they interact with smartwatches (e.g., touching, position on wrist, tightness, movement)

3.2 Hardware Constraints

- Compared to most other mobile devices, smartwatches are much less powerful
- Less battery and computing power
- Smaller touch screen area
- Lower quality speaker and microphones

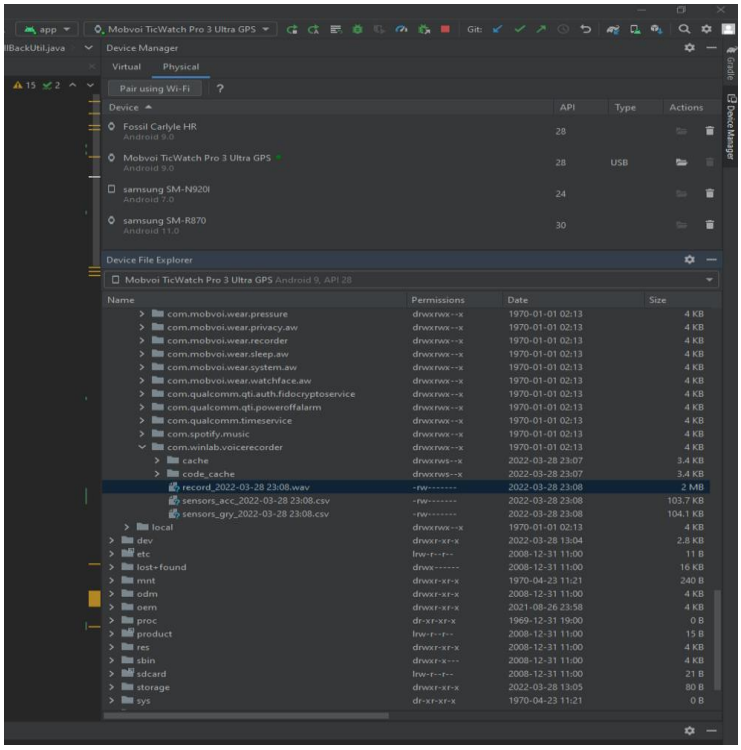
Fewer built-in sensors

3.3 Signal Design

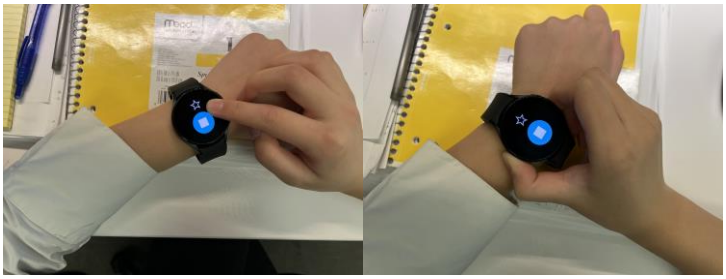
- Frequency response cannot be assumed to be the same as other mobile devices
- Higher sampling rates less feasible on smartwatch hardware
- Sensing area is much smaller, sensors are packed closely together
- Generalizability, signal must be deployable on different hardware models

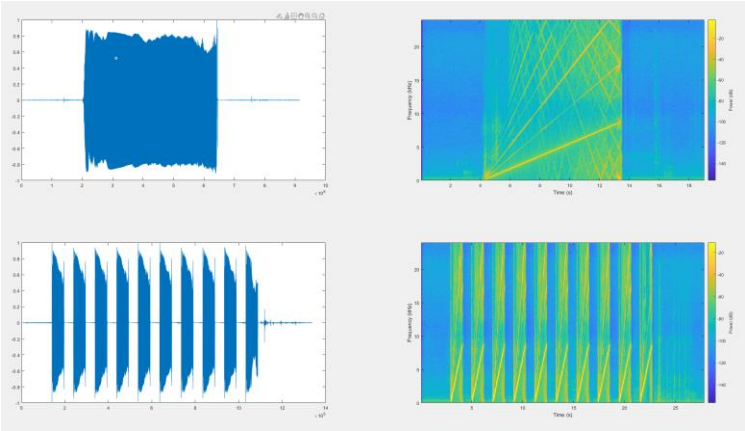
4. Feasibility Study

- First step was to build up an App tailored to my project that can capture the audio signal and vibration at the same time.
    i. The output format of audios was .WAV with 44.1K HZ sample rate (Sample Rate can be adjusted with one line code)
    ii. The output format of vibration datasets was .CSV files
        1. The vibration datasets were collected from Accelerometer + Gyroscope sensors
        2. App Available at: https://github.com/yzhengk/Accelerometer_Gyroscope

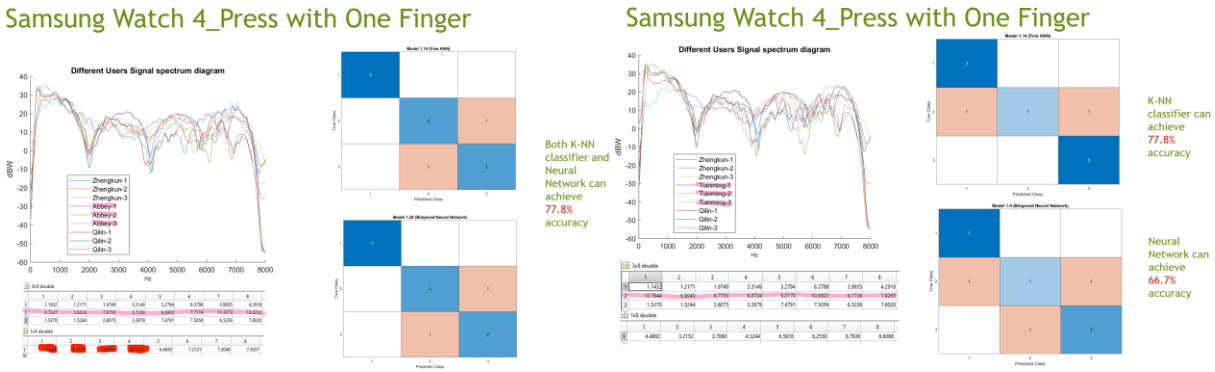


- **Equipment:** Samsung Watch 4/ Fossil Gen 5 (Android Smartwatches)
- **Signal:** 10 times repeated Chirp Signal, Frequency range: 50Hz-8kHz; Sampling rate: 16kHz
- **Two gestures:**
  - Press with one finger on the surface of the smartwatch
  - Press with two sides (one finger covers the speaker)



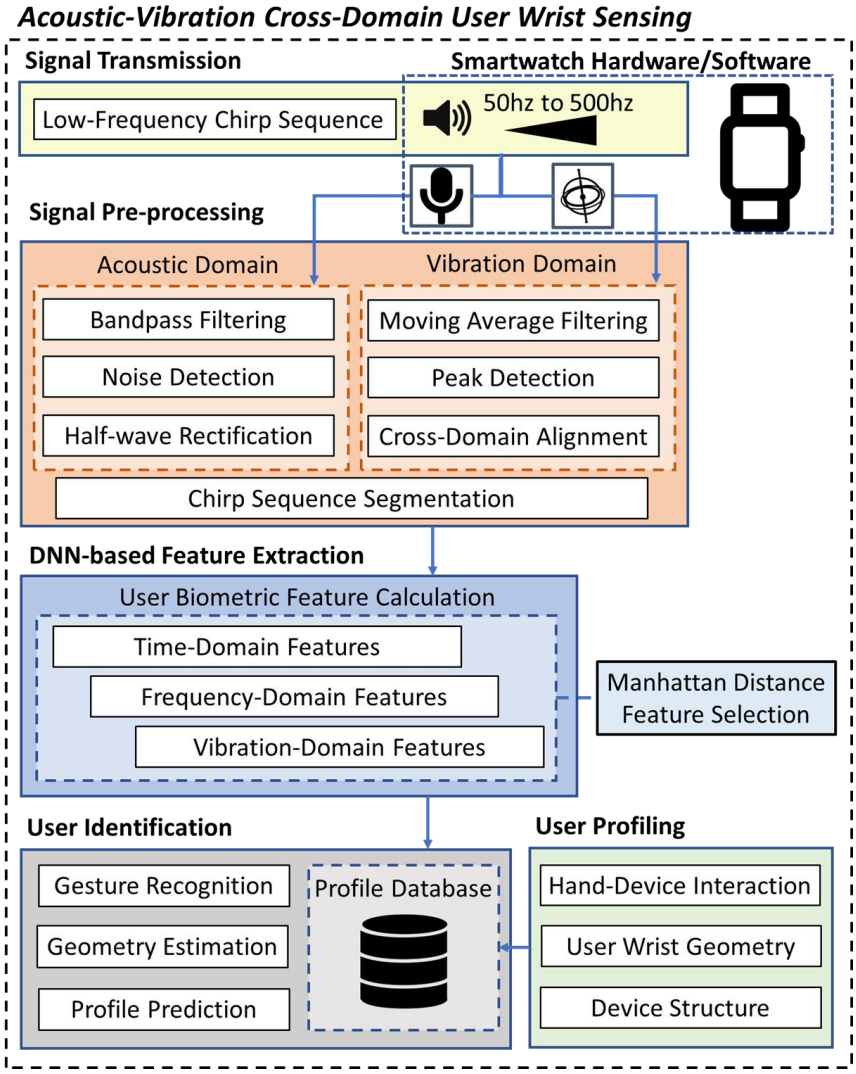


Method: Explore various candidate machine learning-based classifiers including Bagged Decision Trees (BDT), Linear Discriminant Analysis (LDA), K-Nearest Neighbor (KNN), Support Vector Machines (SVM) and Neural Networks.



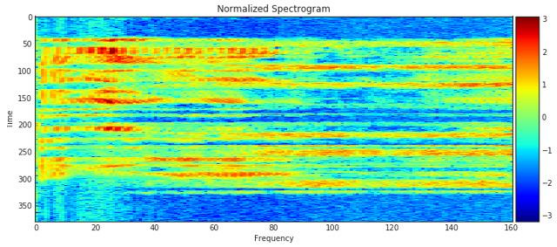
- Hardware differences can result in highly different microphone recordings despite transmitting the same signal in identical conditions

5. System Architecture

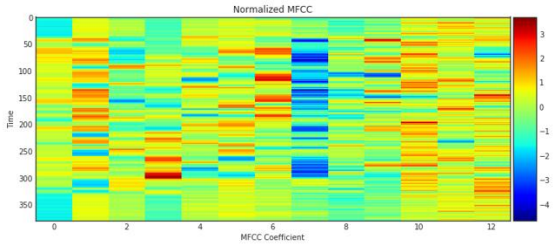


6. Feature Extraction

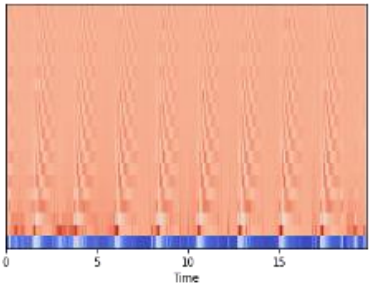
In this project, for acoustic signals I obtained, two features were captured. One is time domain feature, specifically, spectrogram. Another one is Mel-frequency cepstral coefficients, which is also usually called as MFCC in short. In sound processing, the mel-frequency cepstrum is a representation of the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear mel scale of frequency. Mel-frequency cepstral coefficients are coefficients that collectively make up an MFC. The difference between the cepstrum and the mel-frequency cepstrum (MFC) is that in the MFC, the frequency bands are equally spaced on the mel scale, which approximates the human auditory system's response more closely than the linearly spaced frequency bands used in the normal spectrum.



Feature 1: Spectrogram



Feature 2: MFCCs



The first dimension (20) is the number of MFCC coefficients, and the second dimensions (1141) is the number of time frames.

The number of MFCC is specified by n_mfcc, and the number of time frames is given by the length of the audio (in samples) divided by the hop_length.
- Output length = (seconds) * (sample rate) / (hop_length)

MFCCs are computed over a window, i.e., number of samples. To compute MFCC, fast Fourier transform (FFT) is used and that exactly requires that length of a window is provided.

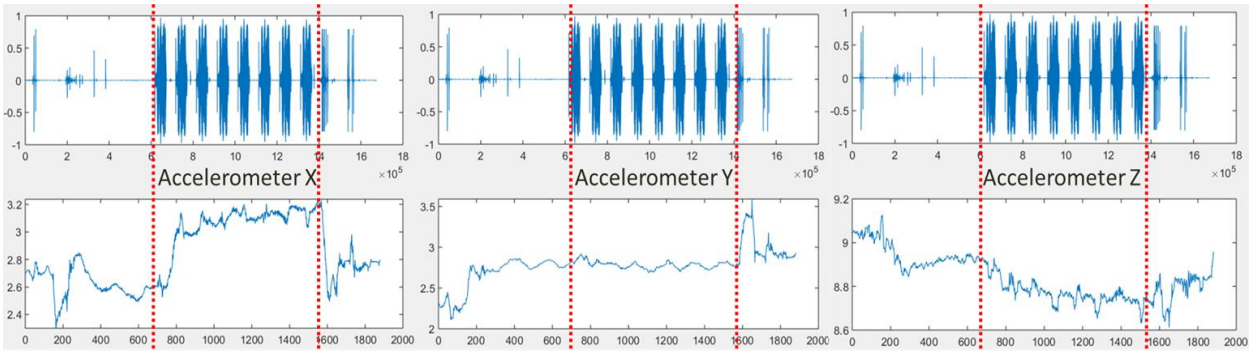
The steps it takes to compute them:
- Spectrograms, using the Short-Time-Fourier-Transform (STFT)
- The Mel spectrogram, from applying Mel scale filterbanks to the STFT
- Mel Frequency Cepstral Coefficients, from applying the DCT transform on the mel-spectrogram.

By default, Python package librosa has sample rate 22050 and window equals 512 (hop).

7. Evaluation

I collected **13** people's data from Temple University and implemented all for evaluation. From what I observed, smartwatches have significant response to 50-8 kHz frequency band. However, the higher frequency band (more than 1kHz) signal makes a loud and jarring sound to human. I also try to use the lower frequency band 50-500Hz. Whereas the lower frequency signal can be easily affected by other sound (white noise/ human talking/ sound from machines). In the experiments, I also conducted the signal pre-processing. The purpose of pre-processing is to suppress noise to prevent from interfering chirp signals extraction. The main idea is that I want to detect and distinguish noise frames to build an estimator and clean the recorded acoustic sound by subtracting estimated noise components and applying half-wave rectification. I believe that, in this case, the advantages include: 1. Simple yet efficient. 2. Smoothly mitigate noise without destroying characteristics of chirp signals.

For vibration data, I utilized moving average filtering since the dataset was more volatile over short period of time. I also applied Savizky-Golay filter; frame length was set to 15 samples for low latency computing. For peak detection, the peak in vibration domain was normally correlated to major events in acoustic domain, e.g., start or end of signal, individual chirps. Three iterations of local maxima were computed over frame length 15 samples. About the cross-domain alignment, acoustic and vibration domains use different sampling rates, therefore we downsample the acoustic domain to match the vibration domain in time, then upsample back to the original acoustic sampling rate to find the matching index. Due to the limited time, this part was not fully evaluated and shall be quite valuable for the future exploration.



For feature comparison between the different users, the simple method was to calculate the difference between MFCCs. I used the city block distance (also referred to as Manhattan distance) to measure the distance between MFCC features. I conducted the new three attempts on three different dates, for Tianming's (my lab mate) three attempts, each time he took off the smartwatch and then wear it on wrist again, we both used right thumb to cover the speaker (left side), right index finger attached to the right side. I calculate the Manhattan distance between MFCC values of each two chirp signals, and then normalize the distance by dividing 1000 to make the number (ratio) small to read. The number is smaller, MFCCs are more similar.

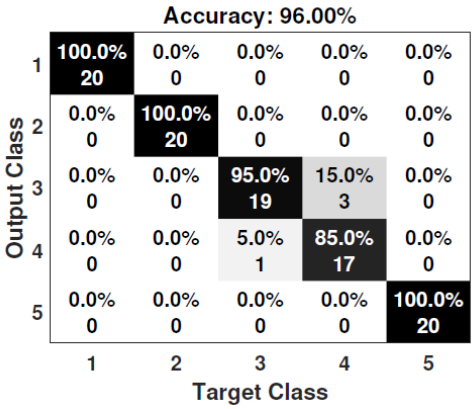
For MFCCs, Pressing Two Sides (one finger covering the speaker of the smartwatch) would cause more significant difference than pressing with one finger on the surface of the smartwatch. For gesture "Press One Finger," we may still see clear difference between different users. For MFCC feature, the two gestures can be utilized. Shape of MFCC: (X, Y), at each time window, the MFCC feature yields a feature vector that characterizes the sound within the window. Note that the MFCC feature is much lower-dimensional than the spectrogram feature, which could help an acoustic model to avoid overfitting to the training dataset.

For DNN-Based Feature Extraction. One-dimensional CNN model with three hidden convolutional layers followed by a fully connected layer and an output classification layer. The input of DNN model is Spectrum with MFCCs. The different type of input has different dimension (Spectrum 8000*1). I trained this different type data separately (They can be combined through data reduction). Model details: The convolution layers have 64, 128, and 256 filter maps with 1×7, 1×5, and 1×3 filter sizes respectively After each convolution layer, system will use the Mean-pooling layer to reduce parameters to prevent overfitting, then compute the categorical-cross entropy of the SVM classification as metric to judge the effectiveness of features extracted by CNN model.

For building the user profile, the CNN model was used to get the feature (An array of vectors). When the participant presses the smartwatch, the data will be collected and get the features through the CNN model. I compute the Manhattan distance between the feature extracted from the input and profiles. I can get the deviation of each data of each feature and get the average deviation of each feature. If the average deviation is less than the threshold value (set by user), the system will accept the user.

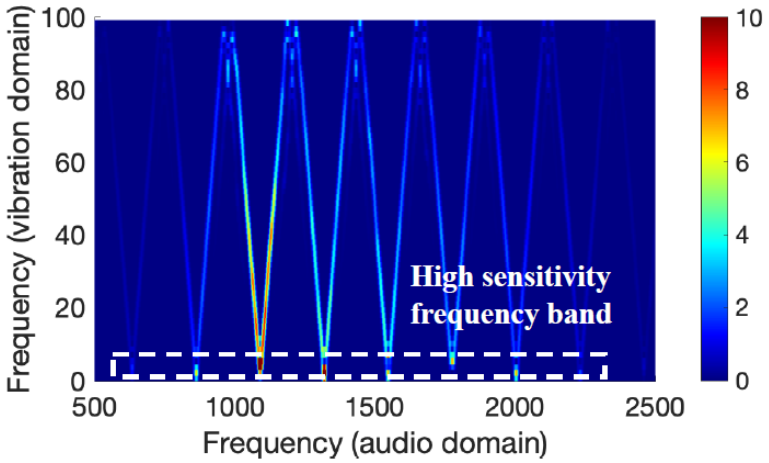
8. Conclusion

For this final project, I utilized acoustic sensing on smartwatches to capture biometric information, both behavioral and physiological. Unlike other works that rely solely on vibration domain information or external sensors, we combine acoustic domain and vibration domain information using only built-in hardware. I studied frequency responses on multiple smartwatch models to determine generalizable acoustic strengths and weaknesses and design an optimal, unintrusive signal deployable on Android (WearOS) hardware. The proposed system with machine learning/deep learning identification technique was quick to conduct, low effort to use, and demonstrated accuracy over 95%.



9. Future

Federated transfer learning would be a great direction to dig in. Data sparsity is a praiseworthy problem. I would like to collect more datasets and then explore more about time-frequency representation derivation, accelerometer artifact mitigation and improve the overall accuracy of the system. Accelerometers are sensitive to low-frequency vibrations due to their design purpose of capturing low-frequency body movements. I am also planning to study different attacking scenarios with various impact factors for the proposed identification system. To derive meaningful representations of voice sounds in the vibration domain for attack detection, I may apply short-time Fourier transformation (STFT) to the vibration signal to derive time frequency representations. Particularly, applying FFT on the vibration signal within a sliding window to obtain frequency representations. I empirically determined the window size and the number of FFT points to be 64. I can further compute the square of FFT magnitudes to obtain the power representations. By sliding the window across the time-series vibration signals and repeating the process, I can obtain the spectrogram representing the vibrations in time and frequency dimensions. Ideally, I look forward to demonstrating the robustness of the designed system against various attack scenarios and adverse conditions. The system is expected to provide reliable user identification simultaneously.



REFERENCES

[1] 2007. Influence of Acoustic Noise on the Dynamic Performance of MEMS Gyroscopes. ASME International Mechanical Engineering Congress and Exposition, Vol. Volume 9: Mechanical Systems and Control, Parts A, B, and C.

[2] 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 (Virtual Event, Hong Kong) (ASIA CCS '21). Association for Computing Machinery, New York, NY, USA, 67–81. https://doi.org/10.1145/3433210.3437518

[3] Les Atlas and Shihab A Shamma. 2003. Joint acoustic and modulation frequency. EURASIP Journal on Applied Signal Processing 2003 (2003), 668–675.

[4] Katarina Bartkova, David Le Gac, Delphine Charlet, and Denis Jouvet. 2002. Prosodic parameter for speaker identification. In Proc. 7th International Conference on Spoken Language Processing (ICSLP 2002). 1197–1200.

[5] Logan Blue, Hadi Abdullah, Luis Vargas, and Patrick Traynor. 2018. 2MA: Verifying Voice Commands via Two Microphone Authentication. In Proceedings of the 2018 on Asia Conference on Computer and Communications Security (Incheon, Republic of Korea) (ASIACCS '18). Association for Computing Machinery, New York, NY, USA, 89–100. https://doi.org/10.1145/3196494.3196545

[6] Joseph P Campbell. 1997. Speaker recognition: A tutorial. Proc. IEEE 85, 9 (1997), 1437–1462.

[7] Huan Feng, Kassem Fawaz, and Kang G. Shin. 2017. Continuous Authentication for Voice Assistants. In Proceedings of the 23rd Annual International Conference on Mobile Computing and Networking (Snowbird, Utah, USA) (MobiCom '17). Association for Computing Machinery, New York, NY, USA, 343–355. https://doi.org/10.1145/3117811.3117823

[8] Matthieu Hébert. 2008. Text-dependent speaker recognition. In Springer handbook of speech processing. Springer, 743–762.

[9] Apple IOS. 2019. Siri. https://www.apple.com/ios/siri/.

[10] Tomi Kinnunen, Bingjun Zhang, Jia Zhu, and Ye Wang. 2007. Speaker verification with adaptive spectral subband centroids. In International Conference on Biometrics. Springer, 58–66.

[11] Johan Lindberg and Mats Blomberg. 1999. Vulnerability in speaker verification-a study of technical impostor techniques. In Sixth European Conference on Speech Communication and Technology.

[12] Yan Michalevsky, Dan Boneh, and Gabi Nakibly. 2014. Gyrophone: Recognizing Speech from Gyroscope Signals. In USENIX Security Symposium. 1053–1067.

[13] 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), 52–55.

[14] Douglas A Reynolds and Richard C Rose. 1995. Robust text-independent speaker identification using Gaussian mixture speaker models. IEEE transactions on speech and audio processing 3, 1 (1995), 72–83.

[15] Prakash Shrestha and Nitesh Saxena. 2018. Listening Watch: Wearable Two-Factor Authentication Using Speech Signals Resilient to Near-Far Attacks. In Proceedings of the 11th ACM Conference on Security Privacy in Wireless and Mobile Networks (Stockholm, Sweden) (WiSec '18). Association for Computing Machinery, New York, NY, USA, 99–110. https://doi.org/10.1145/3212480.3212501

[16] 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 Security and Privacy (EuroS&P), 2017 IEEE European Symposium on. IEEE, 3–18.

[17] Ehsan Variani, Xin Lei, Erik McDermott, Ignacio Lopez Moreno, and Javier Gonzalez-Dominguez. 2014. Deep neural networks for small footprint text-dependent speaker verification. In 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 4052–4056.

[18] WeChat. 2017. Voiceprint. https://thenextweb.com/apps/2015/03/25/wechat-on-ios-now-lets-you-log-in-using-just-your-voice/.

[19] Li Zhang, Parth H. Pathak, Muchen Wu, Yixin Zhao, and Prasant Mohapatra. 2015. AccelWord: Energy Efficient Hotword Detection through Accelerometer. In Proceedings of the 13th Annual International Conference on Mobile Systems, Applications, and Services (Florence, Italy) (MobiSys '15). Association for Computing Machinery, New York, NY, USA, 301–315. https://doi.org/10.1145/2742647.2742658