EmoLeak: Smartphone Motions Reveal Emotions

Ahmed Tanvir Mahdad*, Cong Shi†, Zhengkun Ye‡, Tianming Zhao§, Yan Wang†,
Yingying Chen¶ and Nitesh Saxena*
*Texas A&M University, College Station, TX 77843, USA,
‡ New Jersey Institute of Technology, Newark, NJ 07102, USA,
‡ Temple University, Philadelphia, PA 19122, USA, § University of Dayton, Dayton, OH 45469, USA,
¶ Rutgers University, New Brunswick, NJ 08901, USA
Email: *{mahdad,nsaxena}@tamu.edu †cong.shi@njit.edu, ‡{zhengkun.ye,y.wang}@temple.edu,
§tzhao1@udayton.edu, ¶yingche@scarletmail.rutgers.edu

2023 IEEE 43rd International Conference on Distributed Computing Systems (ICDCS) | 979-8-3503-3986-4/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICDCS57875.2023.00052

***Abstract*—Emotional state leakage attracts increasing concerns as it reveals rich sensitive information, such as intent, demographic, personality, and health information. Existing emotion recognition techniques rely on vision and audio data, which have limited threat due to the requirements of accessing restricted sensors (e.g., cameras and microphones). In this work, we first investigate the feasibility of detecting the emotional state of people in the vibration domain via zero-permission motion sensors. We find that when voice is being played through a smartphone's loudspeaker or ear speaker, it generates vibration signals on the smartphone surface, which encodes rich emotional information. As the smartphone is the go-to device for almost everyone nowadays, our attack based only on motion sensors raises severe concerns about emotion state leakage. We comprehensively study the relationship between vibration data and human emotion based on several publicly available emotion datasets (e.g., SAVEE, TESS). Time-frequency features and machine learning techniques are developed to determine the emotion of the victim based on speech vibrations. We evaluate our attack on both the ear speakers and loudspeakers on a diverse set of smartphones. The results demonstrate our attack can achieve a high accuracy, with around 95.3% (random guess 14.3%) accuracy for the loudspeaker setting and 60.52% (random guess 14.3%) accuracy for the ear speaker setting.**

I. Introduction

Human emotion has found increasing applications in virtual assistants [1], healthcare [2], education [3], and other emerging applications driven by Artificial Intelligence (AI). The leakage of emotional state causes severe privacy issues similar to those that occur in the vision and audio domains, as it reveals rich sensitive information about individuals, such as intent, demographics, personality, and health information. For example, psychographic profiling based on emotion AI can be leveraged to influence political campaigning [4], engaging in psychological operations rooted in the emotions of voters. Such emotion information can also be used for discrimination against individuals with mental illness during the candidate hiring process and workplace monitoring [5]. Additionally, the literature has discussed several potential harms associated with the leakage of emotional state, such as denial of essential services, risk of harm, and infringement of human rights [6]. It has also been noted that the general public is largely unaware [7] of the implications of AI techniques that can track their emotions or mental state.

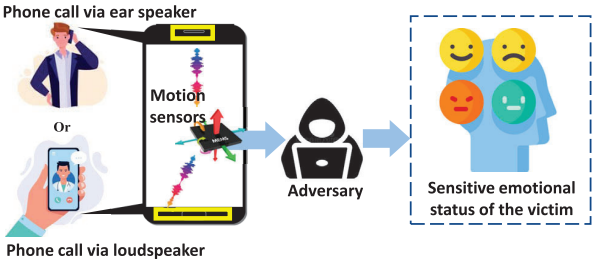


Fig. 1: Overview of emotion eavesdropping from smartphone.

Emotion detection was initially studied in the vision domain [3], [8], where facial landmarks provide rich clues about people's emotions. The investigation of emotion detection is now extending to the audio domain through voice analysis, including the examination of fundamental frequency, formant jitter, shimmer, and spectral energy in speech. Emotion privacy leakage through voice can have more severe consequences compared to vision, as phone calls via smartphones have become the primary means of remote communication. According to Statista [9], 95% of users rely on phone calls for regular communication, making it a lucrative target for adversaries. Recent studies have demonstrated speech eavesdropping using various techniques, such as voice call recording [10], motion sensor data recording [11]–[13], network analysis [14], and external radar systems [15], [16]. However, the potential of deriving emotions through phone calls, especially using zero-permission sensors on the phone, remains unexplored.

In this work, we demonstrate the realistic nature of emotional privacy leakage through phone calls. We propose an eavesdropping attack called *EmoLeak* that can extract the speaker's emotional state from the speech played through either the loudspeaker (e.g., voice calls, recorded audio, YouTube videos) or the ear speaker (e.g., voice calls) by analyzing the zero-permission motion sensors on smartphones. Previous studies have shown the motion sensor's eavesdropping capabilities in speaker and gender identification [17], speech identification [11], [12], speech reconstruction [18], and indoor location identification [19]. However, no research has been conducted on identifying the emotional state of a speaker using vibrations induced by loudspeakers and ear

2575-8411/23/$31.00 ©2023 IEEE
DOI 10.1109/ICDCS57875.2023.00052

speakers. In this work, we address this gap by designing an attack and applying classical machine learning and deep learning approaches to detect the speaker's emotional state. An overview of the attack is illustrated in Figure 1.

We analyzed multiple publicly available speech-emotion datasets, including SAVEE [20], TESS [21], and CREMA-D [22]. Our study focuses on examining how emotional states are encoded into the motion sensor readings by the replayed speech from both the loudspeaker and the ear speaker. To collect data, we recorded accelerometer readings while playing audio files from the dataset through the loudspeaker (utilizing both the top earpiece speaker and the bottom loudspeaker) and the handheld setting (utilizing only the top ear speaker). During our experiment, we analyzed vibration data from both table-top positions (where the phone is placed on a table while collecting data) and handheld positions (resembling the positions and gestures people use during a phone conversation).

By analyzing speech regions and extracting features, our attack is able to infer the speaker's emotional state by applying machine learning and deep learning algorithms to the motion sensor readings. Additionally, we generate spectrograms from the identified speech regions and utilize them to train a CNN-based image classifier for classifying different emotions. The experimental results demonstrate a high level of accuracy (up to 95.3%) in detecting emotions from the vibration data induced by speech, as captured by the accelerometer, using the aforementioned methods. We show that our attack, which is based on motion sensor data, achieves emotion recognition performance comparable to that of approaches based on high-quality audio data. This raises significant privacy concerns, as our attack eliminates the need to collect explicit audio data, thereby bypassing the requirement for permissions.

Our designed attack enables an attacker to discern the emotional state of the caller during a voice call or the speaker in multimedia content. This information can also be correlated with the emotions of the smartphone owner, potentially revealing sensitive information about their content preferences [23].

Our attack extends beyond the loudspeakers to the ear speaker of smartphones, which is used during handheld conversations and produces sounds at much lower volumes. Typically, these speakers generate sounds ranging from 36 dB to 40 dB, and they generally have little to no effect on accelerometer readings. However, we have discovered that certain newer smartphones, such as the OnePlus 7T, have started employing more powerful speakers instead of the smaller ear speakers to deliver higher-quality stereo sound when playing media through the loudspeakers. These speakers generate sound pressure levels ranging from 42 dB to 46 dB, slightly higher than that of other smartphones, yet still comfortable for the human ear during a phone conversation. Our study reveals that speech played from the ear speaker does have some impact on accelerometer data, and it is possible to identify speech regions from it (see Figure 4).

Based on our investigation of vibrations generated by both the loudspeaker and the ear speaker, we propose a novel attack *EmoLeak* for recognizing the emotional state of the speaker. Importantly, we demonstrate that emotions can be classified with reasonable accuracy (95.3% accuracy using loudspeakers and 60.52% accuracy for ear speakers, compared to a random guess rate of 14.27%) using vibrations induced by speakers on the accelerometer data. This finding highlights the real threat of emotional state leakage through motion sensors.

Our Contribution: Our contribution in this research is three-fold:

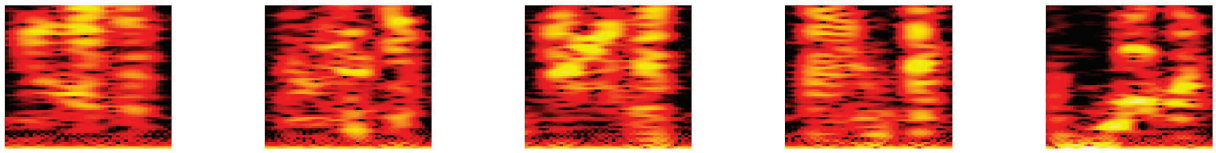
1) ***Eavesdropping Emotion via Smartphone Motion Sensor: A Novel Attack***: We have proposed a new attack for emotion eavesdropping and evaluated its effectiveness by utilizing motion sensor data to identify speech regions, extracting time-frequency domain features, generating spectrograms, and classifying speech emotions using machine learning and deep learning techniques. *To the best of our knowledge, our approach is the first to classify speech emotions from vibrations caused by smartphone speakers.*
2) ***Exploiting Ear Speaker's Vibration on Motion Sensor to Detect Emotion***: We leverage the recent smartphones (having stereo speaker feature) with powerful ear speakers to eavesdrop on the emotional state of the speaker with reasonable accuracy (60.52% accuracy, compared to a random guess rate of 14.27%). *To the best of our knowledge, this is the first research that demonstrates the possibility of detecting emotions using only vibrations produced by the ear speaker.*
3) ***Achieved Reasonable Accuracy in Emotion Detection Datasets Ofhen Comparable to Audio Domain***: Our experiments have shown impressive results, achieving an accuracy of up to 95.3% in classifying emotions using motion sensor data. This level of accuracy is comparable to that achieved in the audio domain, where the same dataset has been known to reach up to 99.5% accuracy [24]–[26]. This suggests that adversaries could potentially use zero-permission motion sensor data to effectively detect the emotional state of a speaker during a phone conversation or audio playback, rather than relying on audio recordings.

II. Context and Prior Works

A. Emotion Detection and Associated Risks

Modern artificial intelligence techniques are capable of identifying the emotional state of a human subject as part of assistive technology for different purposes (e.g., healthcare, market research). However, this emotion AI can be exploited by adversaries who want to gain the private information of a targeted user. There is a growing concern [27], [28] regarding privacy leakage without the user's explicit permission recently due to the potential misuse of this technology.

Researchers have highlighted the lack of awareness among the general public regarding the potential misuse of emotional state leakage to unwanted entities [7]. This technology can detect mental illness in individuals, raising concerns about



(a) Angry (b) Neutral (c) Fear (d) Happy (e) Sad

Fig. 2: Spectrogram generated from accelerometer data during play8ing "Say the word back" through loudspeaker for five different emotion

privacy invasion and discrimination in the workplace during hiring and evaluation processes, as noted in [5]. Regulatory commissions, such as the European Commission, have addressed this issue by drafting regulations [29] and identifying the use of AI in workplace management as high risk. The Information Commissioner's Office of the United Kingdom has also released regulations on AI-related data protection [30] that focus on protecting an individual's rights in organizational settings and address the possibility of bias and discrimination.

In addition, emotional state leakage without the explicit consent of the user can also be misused to create a psychographic profile of a targeted individual. This psychographic profile can be utilized in digital marketing, including political campaigns, which raises concerns about privacy [4]. This psychographic profile has the potential to be used as a prediction tool for a person's health, economic condition, and personal preferences, leading to more extensive privacy breaches. Furthermore, adversaries may exploit sensitive health information, such as a person's mental health state, for illicit purposes.

B. Emotion Detection From Speech Data

Researchers have conducted extensive research on emotion detection using speech data (audio). Generally, popular speech datasets such as CREMA-D, TESS, RAVDESS, and SAVEE are employed for this purpose. Researchers extract various features (e.g., MFCC, spectrogram, statistical features) from these datasets to train powerful machine learning and deep learning models, achieving significant performance on this task. For example, Zeeshan et al. [31] and Pappagari et al. [32] extract MFCC features from the CREMA-D dataset and use them as input to a CNN model to detect six different emotions (Anger, Disgust, Fear, Happy, Neutral, and Sad) with accuracies of 81.5% and 94.99%, respectively. Other research works, such as Zhu et al. [33] and Singh et al. [26], utilize spectrograms of speech data along with CNN models or SVM+RNN models to achieve accuracies of 79.6% and 71.69% on the CREMA-D dataset, respectively. Additionally, Gokilavani et al. [34] leverage multiple features (e.g., MFCC, spectrogram, STFT, and statistical features such as RMS and zero-crossing rate) and a CNN model to achieve an accuracy of 99% on the CREMA-D, TESS, and RAVDESS datasets. These research studies demonstrate that the acoustic signals in human speech contain valuable emotional information that can be effectively captured using appropriate techniques such as feature engineering combined with machine learning or deep learning models.

C. Speech Feature Extraction with Smartphone Motion Sensors

Recent research studies have expanded the concept of speech eavesdropping beyond audio data captured by microphones to include speech inference based on other sensors, such as motion sensors [17], [18], WiFi [35], and mmWave [15], [16]. The underlying idea is to detect conductive vibrations induced by speech playbacks. Notably, Spearphone [17] and AccelEve [36] have demonstrated the feasibility of utilizing the built-in motion sensors of a smartphone to capture vibrations generated by the bottom loudspeaker of the same device. Since the motherboard is shared by both the motion sensor and the loudspeaker, it serves as a conductive medium for propagating vibrations caused by speech playback. These studies have shown promising results in gender detection, speech recognition, and speech reconstruction. However, the potential for extracting the emotional state of the speaker from motion sensor readings remains unexplored. Furthermore, existing attacks require the speech to be played through the bottom loudspeaker of the phone. In contrast, we investigate the possibility of recovering emotional states from speeches played through the top earpiece speaker, which produces sounds that are inaudible to nearby individuals.

III. EmoLeak: Attack Overview and Design

We have designed an attack framework called *EmoLeak* that is capable of eavesdropping on a speaker's emotions using vibrations induced by smartphone speakers. In this section, we will provide an overview of the attacker model, its capabilities, and the design of our attack.

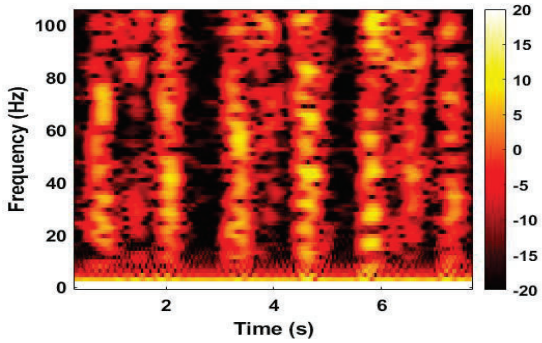
A. Attacker Capability and Threat Model

We assume the attacker can install a benign-looking app on the target user's phone, which runs on the Android operating system, to record and send motion sensor data to remote adversaries. The installed malicious app can record motion sensor data from the background without requiring explicit permission from the user. It can then send the recorded information to the remote adversary for further analysis without leaving any trace. The attacker can capture motion sensor data in the following scenarios: (a) when the target user is engaged in a phone conversation using the ear speaker, (b) when the target user is engaged in a phone conversation using the loudspeaker, and (c) when the target user is playing a multimedia audio file through the loudspeaker. Additionally, the attacker can record multiple conversations or multimedia

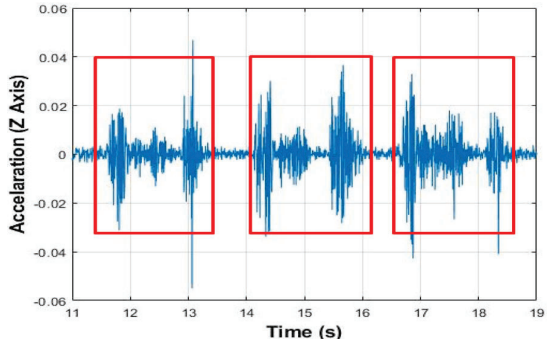
audio files over multiple days to gather more comprehensive training data.

B. Attack Design and Feasibility Analysis

*1) **Efficacy of using Accelerometer Data to Detect Speech Features**:* Prior research [11], [17] has already conducted a comparative frequency response analysis of both the accelerometer and gyroscope and found that the accelerometer is more effective in recognizing speech features compared to the gyroscope, which exhibits a weaker response. Ba et al. [36] have provided a detailed analysis, revealing that the gyroscope demonstrates a lower audio response than the accelerometer. Additionally, gyroscope-based speech recognition schemes (e.g., Gyropohone [13]) primarily rely on the shared surface vibration induced by external speakers, which is not the case in our experiment. Taking into account these observations from prior research, we have decided to utilize the accelerometer in our experiment.



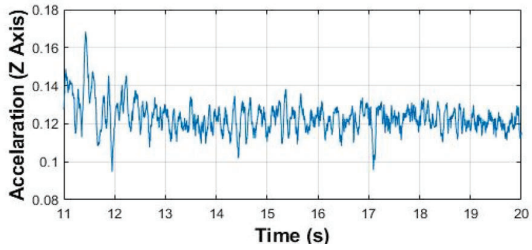
(a) Spectrogram of word regions while playing recording from TESS dataset.



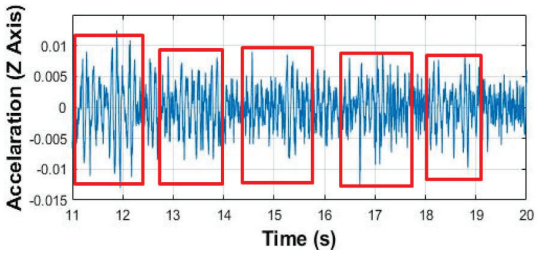
(b) Visual representation of word regions while playing recording from TESS dataset.

Fig. 3: Visual representation of word regions through spectrogram and Acceleration Vs. Time Graph

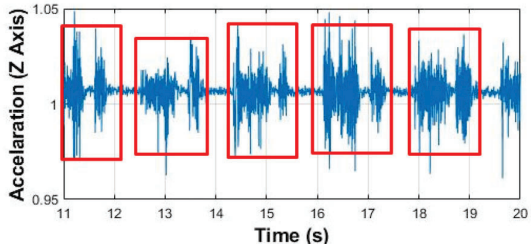
*2) **Speech Region Selection**:* We play recorded audio of actors from publicly available datasets through smartphone speakers, including both the loudspeaker and ear speakers. As mentioned earlier, we collect accelerometer data while the audio is being played and extract the regions where the actual speech is occurring. Our experiment is conducted in both table-top and handheld settings, and we collect speech data from both scenarios.



(a) Raw accelerometer data for earpiece setting (no trace for speech region)



(b) Speech Region detection from accelerotmeter data for earpiece setting after applying 8Hz high-pass filter.



(c) Speech region detection from accelerometer data for loudspeaker setting

Fig. 4: Visual representation of word regions detection in earpiece speaker compared to loudspeaker setting for same speech data.

In the table-top setting, the smartphone accelerometer data should not be impacted by any external vibrations or noises except for the vibration induced by its own speaker. Therefore, we did not use any filter in this case. However, in the handheld setting (where data is collected from the earpiece speaker only), low-frequency noise can be introduced due to hand and body movement during data collection. In this case, adding a filter can effectively reduce the speech features in the accelerometer data, as demonstrated by Zhang et al. in previous research [37].

To further investigate this, we conducted an information gain analysis on both time and frequency domain data with a minimum high pass filter of 1 Hz, as well as without any filter. Our observations reveal that even a 1 Hz high pass filter significantly decreases the information gain. The comparative information gain results are presented in Table I. Based on the insights gained from this information analysis, we made the decision not to apply any filtering during the collection

of time and frequency domain data in order to preserve the speech features from the accelerometer data induced by the ear speaker.

Our speech region extraction program is capable of extracting the speech region from raw accelerometer data for both loudspeakers and table-top settings. The speech region corresponds to the period when a spike in the accelerometer data is observed, indicating vibration induced by the smartphone speaker. Figure 4c provides a visual representation of the extracted speech region from accelerometer data collected in the table-top position. Our detection algorithm achieves a 90% extraction rate for speech regions in the table-top position.

For the handheld and ear speaker-only settings, we applied a high-pass filter of 8Hz solely for the purpose of detecting speech regions. However, we did not use this filter for extracting time-frequency domain data or generating spectrograms later in the process. Figure 4 provides a comparative visual representation of accelerometer data before and after the application of the high-pass filter in the handheld position. Since the earpiece speaker produces lower vibration due to its lower sound pressure, our detection algorithm may not capture all speech regions as effectively as with loudspeaker data. We have observed that our detection algorithm is able to extract a minimum of 45% of word regions from the vibration induced by ear speakers.

TABLE I: Comparison of information gain of some Time-Frequency features with no filters and 1 Hz high-pass filter.

Filter	Information gain of Extracted Features					
	min	mean	max	CV	power	smoothness
no filter	1.31	1.293	1.265	0.994	0.903	0.761
1 Hz	0	0	0	0	0.117	0

*3) **Spectrogram Generation and Time-Frequency Domain Feature Generation from Speech Region**:* We have developed a program that automates the generation of spectrograms for each extracted speech region. Prior to playing the recorded audio from the datasets, we organize the audio segments of the same emotion together and record their respective playback times. For example, in Dataset A, the recording with the "Angry" emotion is played from the 10th second to the 112th second. The program is capable of automatically assigning labels to the spectrograms of each speech region based on the recorded playback times.

Our developed program is also capable of extracting time and frequency domain features from the speech regions that have been extracted. These features can be used as input to machine learning algorithms and deep learning models for emotion detection. Similar to spectrogram generation, we label the time and frequency domain features using the same methodology.

*4) **Efficacy Analysis- Time and Frequency Domain Features**:* The time-frequency features we used are listed in Table II. To assess the effectiveness of these features, we conducted an information gain analysis on the TESS dataset. The analysis demonstrated that all the features listed in Table II exhibit non-zero information gain in both the table-top and handheld settings, suggesting their potential contribution to emotion classification.

TABLE II: Time-Frequency domain features used in this experiment.

Feature	List
Time-domain	Min, Max, Mean, Standard Deviation, Variance, Range, CV, Skewness, Kurtosis, Quantile25, Quantile50, MeanCrossingRate.
Frequency-domain	Energy, Entropy, Frequency Ratio, Irregularity K, Irregularity J, Sharpness, Smoothness, SpecCentroid, SpecStdDev, SpecCrest, SpecSkewness, SpecKurt.

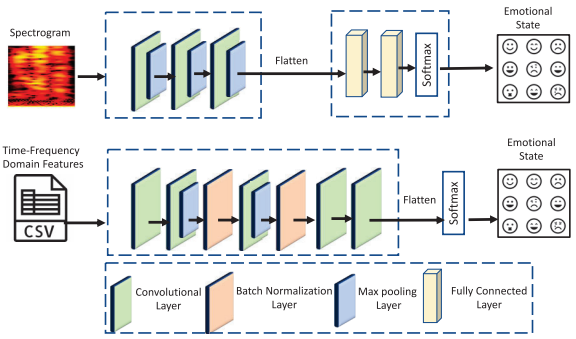


Fig. 5: Visual representation of CNN used in spectrogram classifier (top) and time-frequency domain feature based classifier.

*5) **Efficacy Analysis- Spectrogram**:* We have developed a spectrogram generator that can generate and label spectrograms from the identified speech regions in the accelerometer data. To assess the effectiveness of spectrograms in classifying emotions, we played the same sentence, "Say the word 'Back'", spoken by the same actor but with different emotions. The recorded audio was played using the loudspeaker of the smartphone (OnePlus 7T) in the table-top position. We identified the speech regions and generated a spectrogram for each emotion, as shown in Figure 3. The spectrogram represents a visual representation of the vibration strength captured by the accelerometer over time across various frequencies, and it is expected to exhibit differences for speech with different emotions. These differences can be observed in Figure 2.

IV. Implementation Details

A. Data collection Tools

*1) **Accelerometer Data Collection**:* : We utilize a third-party Android app called "Physics Toolbox Sensor Suite" [38] for collecting accelerometer data. The audio for our experiments is obtained from publicly available emotion datasets, namely, SAVEE, TESS, and CREMA-D. We play the audio files through the smartphone's speakers. As previously mentioned, we collect data in two different positions: table-top (using the loudspeakers) and handheld (using the ear speakers). To exclusively play the audio through the ear speaker, we employ another third-party Android app called "Mobile Ear Speaker Earphone" [39], which redirects all audio output to the ear speaker of the smartphone.

*2) **Speech Region Extraction Tools**:* : We have developed a MATLAB program for analyzing and preprocessing raw accelerometer data, which includes an automatic speech region detection algorithm. In the table-top position and loudspeaker setting, speech region detection is straightforward by analyzing the acceleration data, as shown in Figure 4c. However, in the handheld setting, speech region detection is more challenging. The impact of the ear speaker on the accelerometer data is relatively lower due to its design for phone conversations in the "on-ear" position, resulting in sounds with low sound pressure. Additionally, in the handheld position, low-frequency noises can be introduced due to hand and body movements, unlike in the table-top settings. Thus, to effectively distinguish the speech regions in the handheld setting, we apply an 8 Hz high-pass filter, as illustrated in Figure 4. However, no filter is used during the feature extraction process, as explained in Section III-B.

B. Data Analysis Tools

*1) **Spectrogram Generation**:* We have developed a MATLAB program that can generate spectrograms from each extracted speech region. The program also includes functionality to label each spectrogram image for subsequent classification tasks. As mentioned earlier, in the handheld setting, we collect data in one continuous recording. Therefore, during the audio playback of recordings from the dataset, we group all audio segments of the same emotion together and record their total playback duration (e.g., angry speeches played from the 11th to the 180th second, fear speeches played from the 181st to the 305th second). The spectrogram generation program can automatically assign labels to the spectrograms based on the recorded playback time. Subsequently, we utilize these spectrograms for emotion classification using a Convolutional Neural Network (CNN) classifier.

*2) **Time and Frequency Domain Feature Extraction Tool**:* We have also developed another program using MATLAB to extract time and frequency domain features, as listed in Table II, from each identified speech region. Unlike during speech region detection where we utilize an 8 Hz high-pass filter, we do not apply this filter during the extraction of time and frequency domain features. This is because the filter can significantly remove important speech features, as demonstrated in Table I. The extracted time-frequency domain features are then used for classifying emotions using both classical machine learning algorithms and CNN-based emotion classifiers.

C. Spectrogram-based Emotion Classifier (CNN)

We have used a Convolutional Neural Network (CNN)-based image classifier for the classification of emotions from the spectrogram images. We have implemented this classifier using Python and Keras [40].

*1) **Preprocessing and Data Preparation**:* We divided our labeled generated spectrograms into training and test datasets using an 80/20 train-test split. Subsequently, we generated hdf5 files for the train and test datasets and labeled them accordingly. Furthermore, we resized each spectrogram image from the train and test data to (32x32) dimensions, preparing them to be inputted into the CNN-based image classifier.

*2) **CNN-based Spectrogram Classifier Details**:* We employed a spectrogram-based emotion classifier consisting of three convolutional layers and three fully connected layers. The first convolutional layer, which takes (32x32) images as input, comprises 128 filters and a kernel size of (1,1). The second convolutional layer contains 128 filters, while the third layer has 64 filters. To mitigate overfitting, we incorporated dropout layers with a rate of 0.2 after each of the three convolutional layers. The activation function "ReLU" was utilized in all convolution layers. Additionally, we applied a max pooling layer with a pool size of (2,2) after each convolutional layer to reduce the spatial dimensions of the output.

In the subsequent step, we employ a flatten layer to convert the output into a one-dimensional linear vector, preparing it for the fully connected layers. Following that, the output is propagated through two fully connected layers, each comprising 32 neurons. A dropout layer with a rate of 0.25 is applied to the second fully connected layer. Lastly, the output is passed through a dense layer with the "softmax" activation function, which classifies the spectrogram image into different emotion classes.

D. Emotion Classification by Time-frequency Domain features

We extract time-frequency domain features from each identified speech region, as discussed earlier. These features are used as input for classical machine learning algorithms and CNN classifiers.

*1) **Classical ML-based Emotion Classifier**:* We use Weka [41], a popular data mining and machine learning tool, to preprocess time-frequency feature data and classify emotions. Our time-frequency domain feature extraction program generates time-frequency features and exports them to an output file. We clean the generated data by removing invalid entries such as NaN and blank entries, and prepare the input file with a (.arff) extension for Weka. We use an 80/20 test split and perform 10-fold cross-validation when using ML classifiers.

*2) **CNN-based Emotion Classifier**:* We design and develop a CNN model to classify emotions from time-frequency domain features. We implement the model using Python and Keras, similar to the spectrogram classifier. The CNN model takes time-frequency domain features as input and performs emotion classification based on these features.

As part of preprocessing, we remove invalid data (e.g., NaN, blank entries) and prepare the final (.csv) file for analysis. We apply z-score normalization to transform the data into a standard normal distribution. Our designed model includes five convolutional layers and one fully connected layer in the CNN architecture. The first two convolutional layers have 256 filters each and use the "ReLU" activation function. We incorporate a dropout layer with a rate of 0.25 and a Max Pooling layer with a pool size of 2 after the second convolutional layer. Additionally, we introduce a batch normalization layer after

the third convolutional layer, which consists of 128 filters, to normalize the activation function at each batch. Following this, we include another dropout layer with a rate of 0.25 and a max pooling layer with a pool size of 8. The fourth and fifth convolutional layers consist of 64 filters each and use the "ReLU" activation function. Zero-padding is applied to all inputs in the convolutional layers. A "Flatten" layer is then used after the five convolutional layers to convert the output into a linear vector, preparing it for the fully connected layer. Finally, we employ a fully connected layer with the "softmax" activation function, serving as the output layer for emotion classification.

V. Experiment Outcome and Evaluation

A. Smartphone and Dataset Selection

We have chosen a diverse range of smartphones for our evaluation, including the OnePlus 7T (Android 11.0), OnePlus 9 (Android 13.0), Google Pixel 5 (Android 13.0), Samsung Galaxy S21 (Android 13.0), Samsung Galaxy S21 Ultra (Android 13.0), and Samsung Galaxy S10 (Android 12.0). All the smartphones used in our experiments are equipped with stereo speakers, as claimed by their respective manufacturers. These smartphones feature powerful speakers positioned both at the top (ear speaker) and bottom (loudspeaker) of the device, capable of producing sounds of similar quality and sound pressure.

We have utilized the SAVEE [20] dataset, which consists of 480 speeches from 4 native English male speakers (120 speeches per speaker). This dataset encompasses seven different emotions: anger, disgust, fear, happiness, neutral, sadness, and surprise. We have employed the OnePlus 7T and Google Pixel 5 smartphones to collect data from the audio recordings of the SAVEE dataset.

We have also utilized the TESS (Toronto Emotional Speech Set) dataset [21], which consists of 2800 speeches from two female actors. These speeches encompass seven different emotions: anger, disgust, fear, happiness, pleasant surprise, sadness, and neutral. We have employed five smartphones (OnePlus 7T, Google Pixel 5, Samsung Galaxy S10, Samsung Galaxy S21, and Samsung Galaxy S21 Ultra) to gather and assess the audio data from the TESS dataset.

Another emotion dataset we have utilized is the CREMA-D (CRowd-sourced Emotional Multimodal Actors Dataset) dataset, which comprises 7442 audio-visual clips. For our purposes, we have specifically collected the audio clips, which include speeches from 91 actors representing diverse regions. In contrast to the other two datasets, CREMA-D contains speeches with six emotions: anger, disgust, fear, happy, neutral, and sad. For data collection using this dataset, we have exclusively employed the Samsung Galaxy S10 smartphone.

We have utilized all of the datasets to evaluate attacks in the loudspeaker and table-top settings. However, we have only used the SAVEE and TESS datasets for assessing the ear speaker setting.

TABLE III: Experiment results using SAVEE dataset (Random Guess=14.28%)

Method	**Classifier**	**Accuracy** (OnePlus 7T)	**Accuracy** (Pixel 5)
Time and Frequency Domain Features	Logistic	**53.77%**	44.44%
	multiClassClassifier	51.85%	52.97%
	trees.lmt	51.58%	53.00%
	CNN	46.98%	44.18%
Spectrogram	CNN	39.16%	35.38%

B. Experimental Approach

We collected accelerometer data while playing audio in two different settings. In the table-top setting, we placed the smartphone on a wooden table and set the volume to the maximum level, simulating a phone conversation or media playback through the loudspeaker. On the other hand, in the handheld setting, we collected ear speaker data while holding the smartphone in a typical handheld position, emulating the attack scenario where an attacker can capture accelerometer data induced by the ear speaker during a phone conversation. In the handheld setting, we collected all the data in a continuous manner to minimize the influence of body movement-induced vibrations on different segments of the data.

C. Emotion Recognition in Loudspeaker Settings

TABLE IV: Experiment results using CREMA-D dataset (Random Guess=16.67%)

Method	**Classifier**	**Accuracy** (Samsung Galaxy S10)
Time and Frequency Domain Features	Logistic	58.99%
	multiClassClassifier	58.51%
	trees.lmt	58.99%
	CNN	60.32%
Spectrogram	CNN	53%

For the SAVEE dataset, we used the OnePlus 7T and Google Pixel 5 smartphones to collect accelerometer data. By utilizing extracted time-frequency domain features, we employed classical machine learning algorithms to classify emotions. This approach yielded an accuracy of 53.77% for the OnePlus 7T device. Furthermore, we employed a time-frequency domain-based convolutional neural network (CNN) for emotion classification, achieving an accuracy of 46.98%. Additionally, we generated spectrograms from the identified speech regions and employed a CNN image classifier for emotion classification, resulting in an accuracy of 39.16%. The detailed results of our emotion recognition on the SAVEE dataset in the audio domain are provided in Table III. Moreover, in Table VII, we present a comparison of our results with those of other previous studies that utilize the SAVEE dataset for audio

TABLE V: Experiment results using TESS dataset (Random Guess=14.28%)

Method	**Classifier**	**Accuracy** (OnePlus 7T)	**Accuracy** (Samsung Galaxy S10)	**Accuracy** (Pixel 5)	**Accuracy** (Samsung Galaxy S21)	**Accuracy** (Samsung Galaxy S21 Ultra)
Time and Frequency Domain Features	Logistic	94.52%	78.84%	73.93%	85.79%	82.15%
	multiClassClassifier	91.32%	71.80%	71.75%	84.46%	81.65%
	trees.lmt	94.23%	72.15%	78.48%	87.04%	84.47%
	CNN	**95.3%**	83.2%	82.62%	88.49%	84.38%
Spectrogram	CNN	89.44%	85.37%	80.92%	83.51%	85.74%

domain emotion recognition. While recent works in audio domain emotion detection exhibit significantly better accuracy, our approach utilizing accelerometer vibrations demonstrates reasonable performance, achieving approximately four times better accuracy than random guessing for the SAVEE dataset.

We utilized five smartphones for the TESS dataset in both data collection and analysis. By employing extracted time-frequency domain features in machine learning algorithms, we achieved the highest accuracy of 94.23% using the logistic classifier with the OnePlus 7T device. Additionally, utilizing a CNN as the classifier with time-frequency domain features, we obtained an accuracy of 95.3% (compared to a random guess accuracy of 14.28%). Furthermore, by generating spectrograms for each speech region, our spectrogram-based image classifier achieved an accuracy of 89.44% in classifying emotions from the TESS dataset. Other devices also exhibited decent performance with the TESS dataset. Specifically, the Samsung Galaxy S10, Google Pixel 5, Samsung Galaxy S21, and Samsung Galaxy S21 Ultra devices achieved maximum accuracies of 85.37%, 82.62%, 88.49%, and 84.47%, respectively. A detailed breakdown of the results is provided in Table V, and the confusion matrix for the OnePlus 7T smartphone is shown in Figure 6a. Moreover, we present a comparative analysis with previous works in the audio domain in Table VII, which demonstrates that the performance of the accelerometer in detecting emotions (i.e., 95.3%) is nearly equivalent and comparable to state-of-the-art emotion classification methods.

For the CREMA-D dataset, we exclusively utilized the Samsung Galaxy S10 during data collection and evaluation. Among our evaluation methods, classical machine learning (ML) algorithms employing the time-frequency domain features, achieved an accuracy of 58.99% in classifying emotions using the *logistic* and *tree.lmt* classifiers, as presented in Table IV. Additionally, using a CNN classifier with time-frequency domain features resulted in an accuracy of 60.32%. We also employed a CNN-based image classifier for the generated spectrograms, which yielded a maximum accuracy of 53%. Overall, the performance of emotion classification using vibration-induced accelerometer data demonstrates a reasonable level of accuracy (60.32%, compared to a random guess accuracy of 16.77%) in comparison to recent improvements in audio domain emotion detection using the CREMA-D dataset (as shown in Table VII).

D. Emotion Recognition in Ear Speaker Settings

We use the SAVEE and TESS datasets to evaluate the feasibility of an emotion eavesdropping attack utilizing vibrations induced by the ear speaker in the accelerometer. We collect data in the handheld position during phone conversations, and to mitigate the impact of body and hand movement on accelerometer data, we gather all the data for a particular dataset at one go. For ear speakers, we extract only time and frequency domain features and use both classical machine learning algorithms and CNN to classify them.

We collected accelerometer data from the SAVEE dataset audio played on the ear speaker using OnePlus 7T and OnePlus 9 smartphones. We then evaluated the data using classical machine learning algorithms. The results showed a maximum accuracy of 56.25% for OnePlus 7T and 58.40% for OnePlus 9 smartphones for the SAVEE dataset, with a random guess accuracy of 14.28% (corresponding to the seven emotion classes). By utilizing CNN as the classifier, we observed an improvement in accuracy for the OnePlus 9 smartphone, achieving 60.32% accuracy. For the OnePlus 7T phone, we observed an accuracy of 51.11%. These results demonstrate similar accuracy in classifying emotions compared to the loudspeaker setting.

For the TESS dataset, we used both the OnePlus 7T and OnePlus 9 smartphones to collect and evaluate data. Using classical machine learning algorithms and extracted time-frequency domain features, we achieved a maximum accuracy of 59.57% for emotion classification with the random forest classifier. The corresponding confusion matrix is presented in Figure 6b. When employing a CNN-based classifier, we obtained an accuracy of 54.82% for the OnePlus 7T smartphone. Although the accuracy is lower than that of the loudspeaker setting, it still demonstrates a significant improvement of 4X compared to random guessing (14.28%). The training loss versus validation loss and training accuracy versus validation accuracy graphs are shown in Figure 7c and Figure 7d, respectively.

E. Summary of Results

In this work, we assess an attack scenario where an attacker can collect accelerometer data without requiring any permis-

Angry	Disgust	Fear	Happy	Neutral	Pleasant Surprise	Sad	
72	0	0	0	0	0	0	Angry
0	**73**	0	0	0	0	0	Disgust
0	1	**75**	6	0	0	0	Fear
0	0	5	**59**	0	5	0	Happy
1	0	0	0	**81**	0	0	Neutral
0	0	3	7	0	**73**	0	Pleasant Surprise
0	0	0	0	0	0	**84**	Sad

(a) Confusion matrix generated after classification of time-frequency feature generated from TESS dataset (Loudspeaker Scenario).

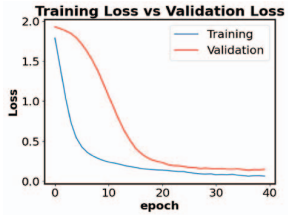
Angry	Disgust	Fear	Happy	Neutral	Pleasant Surprise	Sad	
206	0	0	1	8	4	12	Angry
1	**159**	22	8	15	23	13	Disgust
0	31	**102**	5	16	7	7	Fear
4	20	9	**141**	7	10	16	Happy
12	39	20	5	**71**	19	25	Neutral
6	34	10	6	21	**53**	40	Pleasant Surprise
23	14	15	12	19	25	**132**	Sad

(b) Confusion matrix generated after classification of time-frequency feature generated from TESS dataset for ear speaker scenario (10-fold cross validation is used).

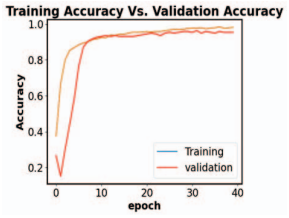
Fig. 6: Confusion matrix generated after emotion classification using time-frequency features from TESS dataset (using OnePlus 7T device).

TABLE VI: Experiment results generated in ear speaker setting (Random Guess=14.28%).

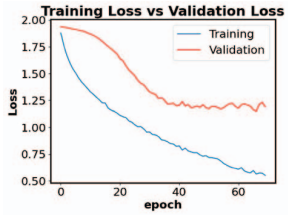
Method	Classifier	SAVEE		TESS
		Accuracy (OnePlus 7T)	**Accuracy** (OnePlus 9)	**Accuracy** (OnePlus 7T)
Time and Frequency Domain Features	Random Forest	53.12%	58.40%	59.67%
	RandomSubspace	56.25%	54.83%	55.45%
	trees.lmt	49.11%	53.76%	53.03%
	CNN	51.11%	60.52%	54.82%



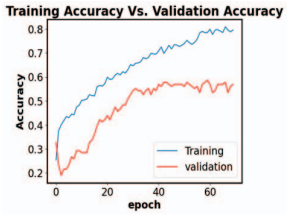
(a) Emotion recognition loudspeaker training loss Vs. validation loss for TESS dataset.



(b) Emotion recognition loudspeaker training accuracy Vs. validation accuracy for TESS dataset.



(c) Emotion recognition ear speaker training loss Vs. validation loss for TESS dataset.



(d) Emotion recognition ear speaker training accuracy Vs. validation accuracy for TESS dataset.

Fig. 7: Emotion Recognition training and validation accuracy graph using CNN using Time-Frequency Domain Features

sion from the smartphone. This data is then utilized to detect the emotional state of the user, either from phone conversations or recorded audio.

In the case of the TESS dataset, our observed accuracy (95.3%) is comparable to the recently reported improved method of emotion detection from the audio domain (highest 99.57%) in the loudspeaker setting. For the SAVEE and CREMA-D datasets, our achieved accuracy in the loudspeaker setting (53.77% and 60.32% respectively) is reasonable compared to recently reported emotion detection from the audio domain (91.7% and 94.99%). Furthermore, our results demonstrate a significant 4X improvement in detecting emotion compared to random guessing rates (14.28% and 16.77% for the SAVEE and CREMA-D datasets respectively). These findings indicate that an attacker can exploit a zero-permission motion sensor to recognize emotions with an acceptable level of accuracy.

For the ear speaker setting, both the SAVEE and TESS datasets exhibit a significant 4X improvement compared to random guessing rates (achieving the highest accuracies of 60.32% and 59.67% respectively). It is worth noting that ear speakers are typically characterized by lower sound pressure and have a negligible impact on accelerometer data. However, recent smartphone trends indicate a growing number of devices incorporating stereo speakers, utilizing both the top ear speaker and bottom speaker for enhanced audio output. This configuration does have a slight impact on the accelerometer, and our speech detection approach can successfully identify more than 45% of speech regions from it. These findings suggest that it is indeed possible to detect emotion with a satisfactory level of accuracy.

VI. Discussion and Future Directions

A. Android limitation on Sampling Rate

To protect potentially sensitive user information, if the app targets Android 12 or higher, the system will limit the data refresh rate for certain motion and position sensors. In light

TABLE VII: Summary Results

Datasets	Vibration Domain (EmoLeak)	Audio Domain (Previous Works)			
SAVEE	53.77%	91.7% [42]	83.20% [43]	81.20% [44]	59.7% [45]
TESS	95.4%	99.52 [26]	99% [34]	96% [25]	94.47% [24]
CREMA-D	60.32%	94.99% [31]	84% [34]	81.5% [32]	79.6% [33]

of this, we have conducted the Android restriction (200 Hz sampling rate) testing in our case. For the TESS dataset and loudspeaker setting, we accomplished an 80.1% accuracy for 200 Hz data. Nevertheless, for the same device with a default sampling rate, we achieved the highest accuracy at 95.3%. The observations indicate that although accuracy was scaled down a bit, it was still a greater than 5X improvement compared to random guessing (14.3%).

B. Risk Mitigation Strategies

The use of smartphone accelerometer data collected through speaker-induced vibrations to eavesdrop on emotional states poses a significant risk to the privacy and security of sensitive health information. To mitigate this risk, it is crucial for smartphone operating system designers to implement a requirement for obtaining users' prior consent before accessing and utilizing motion sensor data. Currently, Android requires user consent for third-party apps to collect motion sensor data with a sampling rate higher than 200 Hz. However, our research suggests that this requirement alone may not be sufficient to fully protect against emotion eavesdropping. As such, we recommend implementing stricter limitations on sensor data collection and requiring explicit user permission for any such data collection to occur. Smartphone designers and manufacturers can consider the hardware design and positioning of the motion sensor as a mitigation approach. This can include relocating the motion sensor away from all speakers of the smartphone, as well as incorporating vibration-absorbing materials around motion sensors to reduce the impact of speaker vibrations on sensor data.

C. Limitation of Emotion Detection using Motion Sensors

Our evaluation approach has certain limitations, one of which is the limited scope of emotions that are assessed. Our selected dataset only covers a restricted range of emotions and does not consider more complex emotions (e.g., ambivalence, envy, nostalgia, guilt, melancholy). Additionally, the distance between the accelerometer and the speaker can vary based on the smartphone's design. This variation could lead to performance variations in detecting emotions, which is also an important factor to consider. Moreover, our approach is susceptible to external noise factors in the environment that may impact the accelerometer data. Particularly in ear speaker scenarios, body movements during conversations can affect the accuracy of emotion detection. Furthermore, smartphone manufacturers may utilize different models of motion sensors, which could have varying levels of sensitivity. This can result in performance discrepancies while detecting emotions.

D. Future Research Directions

Researchers can investigate the accuracy of emotion detection using smartphone motion sensors by expanding the diversity of datasets, involving real-world participants, and incorporating a broader range of complex emotions. Additionally, they can develop more efficient machine learning and deep learning models with enhanced accuracy rates. To gain a more comprehensive understanding of emotion detection, researchers can also test in various environments (i.e., indoor or outdoor) and explore similar vulnerabilities in other smart devices and wearables.

VII. Conclusion

In conclusion, this research explored the potential privacy risks associated with zero-permission motion sensors in determining a person's emotional state through eavesdropping using the motion sensors of their smartphones. By utilizing the in-built motion sensors, an adversary can potentially gather information about a person's emotional state. Our research investigated the feasibility of determining the speaker's emotion from vibration data induced by loudspeakers and earpiece speakers, achieving a reasonable accuracy of up to 95.3% in emotion detection. This research opens up opportunities for further exploration of eavesdropping on non-semantic features of speech using motion sensors and emphasizes the need for additional research on protecting users' privacy.

VIII. Acknowledgement

This work is partially supported by National Science Foundation grants: CNS-2145389, CNS-2120276, CCF-2000480, OAC-2139358, CNS-2201465, CCF-1909963, CCF-2211163, CNS-2120396 and CNS-2152669.

References

[1] P.-S. Chiu, J.-W. Chang, M.-C. Lee, C.-H. Chen, and D.-S. Lee, "Enabling intelligent environment by the design of emotionally aware virtual assistant: A case of smart campus," *IEEE Access*, vol. 8, pp. 62 032–62 041, 2020.

[2] M. Awais, M. Raza, N. Singh, K. Bashir, U. Manzoor, S. U. Islam, and J. J. Rodrigues, "Lstm-based emotion detection using physiological signals: Iot framework for healthcare and distance learning in covid-19," *IEEE Internet of Things Journal*, vol. 8, no. 23, pp. 16 863–16 871, 2020.

[3] M. Mukhopadhyay, S. Pal, A. Nayyar, P. K. D. Pramanik, N. Dasgupta, and P. Choudhury, "Facial emotion detection to assess learner's state of mind in an online learning system," in *Proceedings of the 2020 5th International Conference on Intelligent Information Technology*, 2020, pp. 107–115.

[4] V. Bakir, "Psychological operations in digital political campaigns: Assessing cambridge analytica's psychographic profiling and targeting," *Frontiers in Communication*, vol. 5, p. 67, 2020.

[5] S. Monteith, T. Glenn, J. Geddes, P. C. Whybrow, and M. Bauer, "Commercial use of emotion artificial intelligence (ai): Implications for psychiatry," *Current Psychiatry Reports*, pp. 1–9, 2022.

[6] J. Hernandez, J. Lovejoy, D. McDuff, J. Suh, T. O'Brien, A. Sethumadhavan, G. Greene, R. Picard, and M. Czerwinski, "Guidelines for assessing and minimizing risks of emotion recognition applications," in *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2021, pp. 1–8.

[7] S. Zhang, Y. Feng, L. Bauer, L. F. Cranor, A. Das, and N. Sadeh, ""did you know this camera tracks your mood?": Understanding privacy expectations and preferences in the age of video analytics," *Proceedings on Privacy Enhancing Technologies*, vol. 2021, no. 2, pp. 282–304, 2021.

[8] A. Jaiswal, A. K. Raju, and S. Deb, "Facial emotion detection using deep learning," in *2020 International Conference for Emerging Technology (INCET)*. IEEE, 2020, pp. 1–5.

[9] Statista, "Leading forms of communication among users in the united states as of january 2022," 2022, https://www.statista.com/statistics/1332443/us-users-top-communication-methods/.

[10] Y. Zhou and X. Jiang, "Dissecting android malware: Characterization and evolution," in *2012 IEEE symposium on security and privacy*. IEEE, 2012, pp. 95–109.

[11] W. Su, D. Liu, T. Zhang, and H. Jiang, "Towards device independent eavesdropping on telephone conversations with built-in accelerometer," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 5, no. 4, pp. 1–29, 2021.

[12] M. Gao, Y. Liu, Y. Chen, Y. Li, Z. Ba, X. Xu, J. Han, and K. Ren, "Device-independent smartphone eavesdropping jointly using accelerometer and gyroscope," *IEEE Transactions on Dependable and Secure Computing*, 2022.

[13] Y. Michalevsky, D. Boneh, and G. Nakibly, "Gyrophone: Recognizing speech from gyroscope signals," in *23rd USENIX Security Symposium (USENIX Security 14)*, 2014, pp. 1053–1067.

[14] D. Rupprecht, K. Kohls, T. Holz, and C. Pöpper, "Call me maybe: Eavesdropping encrypted {LTE} calls with {ReVoLTE}," in *29th USENIX security symposium (USENIX security 20)*, 2020, pp. 73–88.

[15] C. Wang, F. Lin, T. Liu, K. Zheng, Z. Wang, Z. Li, M.-C. Huang, W. Xu, and K. Ren, "mmeve: eavesdropping on smartphone's earpiece via cots mmwave device," in *Proceedings of the 28th Annual International Conference on Mobile Computing And Networking*, 2022, pp. 338–351.

[16] S. Basak and M. Gowda, "mmspy: Spying phone calls using mmwave radars," in *2022 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2022, pp. 1211–1228.

[17] S. A. Anand, C. Wang, J. Liu, N. Saxena, and Y. Chen, "Spearphone: A speech privacy exploit via accelerometer-sensed reverberations from smartphone loudspeakers," *arXiv preprint arXiv:1907.05972*, 2019.

[18] P. Hu, H. Zhuang, P. S. Santhalingam, R. Spolaor, P. Pathak, G. Zhang, and X. Cheng, "Accear: Accelerometer acoustic eavesdropping with unconstrained vocabulary," in *2022 IEEE Symposium on Security and Privacy (SP)*. IEEE Computer Society, 2022, pp. 1530–1530.

[19] H. Zheng and H. Hu, "Missile: A system of mobile inertial sensor-based sensitive indoor location eavesdropping," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 3137–3151, 2019.

[20] SAVEE, "Surrey audio-visual expressed emotion (savee)," 2022, https://www.kaggle.com/datasets/ejlok1/surrey-audiovisual-expressed-emotion-savee.

[21] TESS, "Toronto emotional speech set (tess)," 2022, https://www.kaggle.com/datasets/ejlok1/toronto-emotional-speech-set-tess.

[22] CREMA-D, "Crowd sourced emotional multimodal actors dataset (crema-d)," 2022, https://www.kaggle.com/datasets/ejlok1/cremad.

[23] Reddit, "Herded down the rabbit hole — mozilla says youtube is still recommending 'harmful' content," 2021, https://www.biometrica.com/herded-down-the-rabbit-hole-mozilla-says-youtube-is-still-recommending-harmful-content/.

[24] R. Chatterjee, S. Mazumdar, R. S. Sherratt, R. Halder, T. Maitra, and D. Giri, "Real-time speech emotion analysis for smart home assistants," *IEEE Transactions on Consumer Electronics*, vol. 67, no. 1, pp. 68–76, 2021.

[25] N. Patel, S. Patel, and S. H. Mankad, "Impact of autoencoder based compact representation on emotion detection from audio," *Journal of Ambient Intelligence and Humanized Computing*, vol. 13, no. 2, pp. 867–885, 2022.

[26] R. Singh, H. Puri, N. Aggarwal, and V. Gupta, "An efficient language-independent acoustic emotion classification system," *Arabian Journal for Science and Engineering*, vol. 45, no. 4, pp. 3111–3121, 2020.

[27] The Pioneer, "Emotion recognition technology — a new challenge to privacy," 2021, https://www.dailypioneer.com/2021/state-editions/emotion-recognition-technology-----a-new-challenge-to-privacy.html.

[28] Lexology, "Emotion recognition technology - a cause for concern?" 2022, https://www.lexology.com/library/detail.aspx?g=83965e72-5fd1-4887-83ff-03675c9a0438.

[29] European Union, "Proposal for a regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts," 2021, https://www.lexology.com/library/detail.aspx?g=83965e72-5fd1-4887-83ff-03675c9a0438.

[30] Information Commissioner's Office, "Guidance on ai and data protection," 2022, https://ico.org.uk/for-organisations/guide-to-data-protection/key-dp-themes/guidance-on-artificial-intelligence-and-data-protection/.

[31] M. Zeeshan, H. Qayoom, and F. Hassan, "Robust speech emotion recognition system through novel er-cnn and spectral features," in *2021 4th International Symposium on Advanced Electrical and Communication Technologies (ISAECT)*. IEEE, 2021, pp. 01–06.

[32] R. Pappagari, T. Wang, J. Villalba, N. Chen, and N. Dehak, "x-vectors meet emotions: A study on dependencies between emotion and speaker recognition," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7169–7173.

[33] Z. Zhu and Y. Sato, "Reconciliation of multiple corpora for speech emotion recognition by multiple classifiers with an adversarial corpus discriminator." in *INTERSPEECH*, 2020, pp. 2342–2346.

[34] M. Gokilavani, H. Katakam, S. A. Basheer, and P. Srinivas, "Ravdness, crema-d, tess based algorithm for emotion recognition using speech," in *2022 4th International Conference on Smart Systems and Inventive Technology (ICSSIT)*. IEEE, 2022, pp. 1625–1631.

[35] G. Wang, Y. Zou, Z. Zhou, K. Wu, and L. M. Ni, "We can hear you with wi-fi!" in *Proceedings of the 20th annual international conference on Mobile computing and networking*, 2014, pp. 593–604.

[36] Z. Ba, T. Zheng, X. Zhang, Z. Qin, B. Li, X. Liu, and K. Ren, "Learning-based practical smartphone eavesdropping with built-in accelerometer." in *NDSS*, 2020.

[37] L. Zhang, P. H. Pathak, M. Wu, Y. Zhao, and P. Mohapatra, "Accelword: Energy efficient hotword detection through accelerometer," in *Proceedings of the 13th Annual International Conference on Mobile Systems, Applications, and Services*, 2015, pp. 301–315.

[38] V. Software, "Physics toolbox sensor suite," 2022, https://play.google.com/store/apps/details?id=com.chrystianvieyra.physicstoolboxsuite.

[39] M. M. Solutions, "Mobile ear speaker earphone," 2022, https://play.google.com/store/apps/details?id=com.sparkapps.mobileearphone.yp.

[40] Keras, "Keras: The python deep learning api," 2022, https://keras.io/.

[41] Weka, "The data platform for ai," 2022, https://www.weka.io/.

[42] H. A. Abdulmohsin *et al.*, "A new proposed statistical feature extraction method in speech emotion recognition," *Computers & Electrical Engineering*, vol. 93, p. 107172, 2021.

[43] M. Farooq, F. Hussain, N. K. Baloch, F. R. Raja, H. Yu, and Y. B. Zikria, "Impact of feature selection algorithm on speech emotion recognition using deep convolutional neural network," *Sensors*, vol. 20, no. 21, p. 6008, 2020.

[44] N. Hajarolasvadi and H. Demirel, "3d cnn-based speech emotion recognition using k-means clustering and spectrograms," *Entropy*, vol. 21, no. 5, p. 479, 2019.

[45] H. M. Fayek, M. Lech, and L. Cavedon, "Towards real-time speech emotion recognition using deep neural networks," in *2015 9th international conference on signal processing and communication systems (ICSPCS)*. IEEE, 2015, pp. 1–5.