Deep Learning enabled Keystroke Eavesdropping Attack over Videoconferencing Platforms

Xueyi Wang, Yifan Liu, Shancang Li

***Abstract*—The COVID-19 pandemic has significantly impacted people by driving people to work from home using communication tools such as Zoom, Teams, Slack, *etc*. The users of these communication services have exponentially increased in the past two years, e.g., Teams annual users reached 270 million in 2022 and Zoom averaged 300 million daily active users in videoconferencing platforms. However, using edging artificial intelligence techniques, new cyber attacking tools expose these services to eavesdropping or disruption. This work investigates keystroke eavesdropping attacks on physical keyboards using deep learning techniques to analyze the acoustic emanation of keystroke audios to identify victims' keystrokes. An accurate context-free inferring algorithm was developed that can automatically predict keystrokes during inputs. The experimental results demonstrated that the accuracy of keystroke inference approaches is around 90% over normal laptop keyboards.**

***Index Terms*—Acoustic Emanations, Microsoft Teams, Acoustic Keyboard Eavesdropping, CNN, Random Forest.**

I. Introduction

The COVID-19 pandemic and working from home have led to significant changes in the way people work and videoconferencing platforms (such as *Teams, Zoom, Slack, etc*.) have become the preferred model in many scenarios, which likely will continue after the pandemic ends. Videoconferencing plays an increasingly important role in various industry sectors, including education, business, government, *etc*.

The videoconferencing services offer the ability to have remote virtual meetings with a large number of participants from all over the world. However, the dramatic increase in videoconferencing applications also raised security and privacy concerns, which include: (1) protecting privacy and videoconferencing data; (2) security vulnerabilities in videoconferencing platforms; (3) external cyber attacks; (4) breaches of intellectual property, sensitive data, *etc*. Specifically, as pervasive platforms, the widely used videoconferencing platforms have increased the attack surface exploited by malicious actors [1].

In the past few years, tremendous research efforts have been conducted to enhance the security of videoconferencing applications. Specifically, research efforts on cyber threats, *such as malware attack, malicious link, password-protection, suspicious activity identification, etc.*, show great potential for enhancing the security of these platforms. However, side-channel vulnerabilities are still challenging for videoconferencing platforms, including keystroke eavesdropping attacks on the physical keyboard that use acoustic emanation of keystroke audios to identify specific keystrokes; touch screen input attacks, *etc*. Most works used keyboard acoustic emanations audio collected from additional one or more professional recording microphones to provide sound details and some of them use dictionary-based methods to polish the predicted results [1]. Because random text like inputting passwords does not necessarily under the word spelling, frequency, and grammar limit [2]. Compared to traditional acoustic localization methods, our model is more resilient and doesn't require additional recording devices or harsh experimental premises.

This work investigates a side-channel attack on videoconferencing application, *acoustic keyboard eavesdropping*, employing a deep learning model to automatically predict and record keystrokes. The main contributions include:

- Developed a Bisection-based method to conduct auto-segmenting keystroke events from long recording audio;
- A feature extraction model is developed using Mel spectrogram to extract keystroke features;
- Experimental results demonstrate the proposed scheme can achieve high accuracy for predicting keystrokes via keyboard acoustic emanations on the Teams platform.

II. Keystrokes Audio Processes

This work investigates the keystroke inferring over Teams platform, which includes the following four phases: 1) Collecting raw audio data from Teams platform; 2) Segmenting the raw audio into individual keystroke audios; 3) Extracting the main features from the audios; and 4) Predicting keystrokes via ML/DL algorithms. Figure 1 details the procedures.

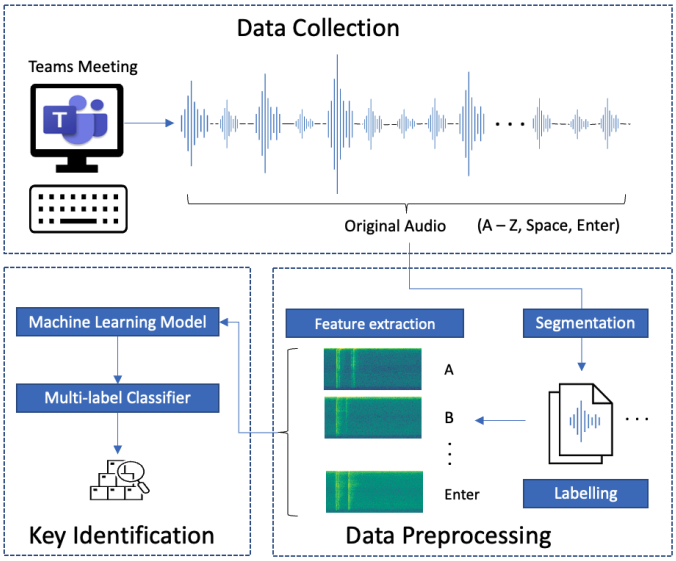

Fig. 1. Keystrokes Prediction using Deep Learning

A. Keystroke Audios Collection

In this work, two laptops, L_a (MacBook Pro 13, 2019) and L_b *ThinkPad P50*, were used to collect the raw keystroke audios over Teams (*Version 1.5.00.3356*) recording. The raw keystroke audios were collected in *mp4* format with *16000*Hz sample rate. Total 50 groups of 28-key typing audio data were collected on both laptop models from Teams meeting with no Noise Suppression (from keys *A - Z, Space*, and *Enter*)[1].

B. Keystroke Audios Segmentation

Each regular keystroke event waveform is indicated as a segment highlighted in blue in Figure 2, which includes two distinct peak values corresponding to two special events: finger pressing the key cap (*Key Press Point, KPP*) and finger releasing the key cap (*Key Release Point, KRP*). Because of the various typing habits like typing gestures, strength, and speed, the KRPs are not always recognizable. As a result, this work uses the amplitude around the KPP value to identify the start of a keystroke event.

In this work, the audio segmentation is based on the following assumptions as also illustrated in Figure 2:

1) Every key pressing peak amplitudes are greater than the background noise amplitude;
2) A keystroke event is in the range of $-0.18s$ and $0.82s$ around the KPP that lasts for 1 second;
3) The least interval between two keystrokes was set as $0.625s$.

The initial detecting threshold was set a bit lower than the highest amplitude value from the whole waveform and then stored all the time frames whose related amplitudes are all greater than the current threshold. If the interval between the stored two time frames is greater than $0.625s$, the latter one will be identified as a new KPP to represent a new keystroke event. The Bisection method was used to auto-adjusting the threshold until finding all the keystroke events.

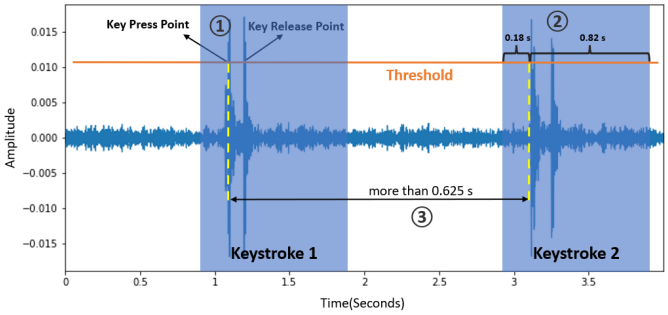

Fig. 2. Keystroke Event Segmentation

C. Keystroke Audios Feature Extraction

For the feature extraction part, this work uses the Mel-spectrogram feature, which has been proven that has an excellent performance in Acoustic Scene Classification scenarios [3]. Mel-spectrogram is a spectrogram generated from raw sound signals using Fourier Transforms and then converted to Mel scale, which mirrors human perception.

[1]The collected dataset is available on https://github.com/kumisaki/ Context-Free-Keystroke-Eavesdropping-Attack-over-Teams-Platform/

III. Keystrokes Prediction over Teams

Based on the extracted features, different ML/DL models, including *CNN, MLP, XGBoost, etc.* were trained to predict keystrokes. This work uses 80%/20% train/test split on the dataset and conducts a comparison study between neural network algorithms (CNN, MLP) and tree structure algorithms (XGBoost, Random Forest) in this multi-class audio classification task. For each algorithm, the accuracy was obtained in the average of 10 times model fitting under both L_a and L_b of the 50 groups 28-key dataset. Table I presents the predicting accuracy against different deep learning models used.

TABLE I
KEYSTROKES PREDICTING ACCURACY WITH DIFFERENT ML/DL MODELS OVER TEAMS

	CNN	MLP	XGBoost	RF	SVM	KNN
L_a	0.891	0.870	0.846	0.841	0.824	0.817
L_b	0.898	0.885	0.883	0.921	0.839	0.779

CNN combined with the Mel-spectrogram model shows a high and stable prediction accuracy on both L_a (89.1%) and L_b (89.8%). The random forest (RF) shows a 92.1% accuracy on L_b on average.

IV. Conclusion and Discussion

This work attempts to investigate the deep learning enabled keystroke eavesdropping attack targeting commonly used videoconferencing platforms, such as *Teams, Zoom, etc.*, inferring keystrokes in real environments. The developed prediction models show capable performance in identifying typed characters from the keystroke sound signals. Because of its high reproducibility, wide range, stealthy, and huge damage impact, the keystroke eavesdropping attack is considered a high-risk level attack in the DREAD model.

To prevent keystroke eavesdropping attacks over Teams and other platforms from gaining sensitive information and credentials, the following ways can be helpful: 1) Use multi-factor authentication and CAPTCHA to prevent sniffing of credentials; 2) Use high noise suppression over videoconferencing platforms to stop raw keystroke sounds collection; 3) Based on the proposed solution, use generative adversarial networks (GANs) generate extra noise to confuse the trained models.

References

[1] J. Yu and *et al.*, "An indirect eavesdropping attack of keystrokes on touch screen through acoustic sensing," *IEEE Transactions on Mobile Computing*, vol. 20, no. 2, pp. 337–351, 2021.

[2] L. Lu, J. Yu, Y. Chen, Y. Zhu, X. Xu, G. Xue, and M. Li, "Keylistener: Inferring keystrokes on qwerty keyboard of touch screen through acoustic signals," in *IEEE INFOCOM 2019-IEEE Conference on Computer Communications*. IEEE, 2019, pp. 775–783.

[3] T. Zhang, G. Feng, J. Liang, and T. An, "Acoustic scene classification based on mel spectrogram decomposition and model merging," *Applied Acoustics*, vol. 182, p. 108258, 2021.