2nd Year ENSICAEN Project:
GREYC Forensic Project

Keystroke Recognition by Sound

Belghomari Abdelmalek[a], Willaume Cedric[b], Bedda Mohamed-Abderrahmane[c],Sriha Haykel[d],Kamtchueng Winnie[e]

Keywords: Artificial Intelligence; Machine Learning; Keyboard keystroke recognition; State of the art; User authentication; Cybersecurity; Forensic

Abstract: This paper presents a state-of-the-art analysis of our ongoing second-year academic project, focused on keyboard keystroke recognition. Throughout the academic year, we explore evolving techniques and methodologies, such as machine learning algorithms and biometrics, to accurately identify typed text and commands. Our review discusses recent milestones, challenges, and potential applications in fields like cybersecurity and user authentication. As our project progresses, this analysis serves as a foundation for our research and development efforts, aiming to contribute to the advancement of keyboard keystroke recognition. Copyright © 2023 ENSICAEN

[a] **abdelmalek.belghomari@ecole.ensicaen.fr**
[b] **cedric.willaume@ecole.ensicaen.fr**
[c] **mohamed-abderrahmane.bedda@ecole.ensicaen.fr**
[d] **haykel.sriha@ecole.ensicaen.fr**
[e] **winnie.kamtchueng-fodjo-kouam@ecole.ensicaen.fr**

[Version: v1.00]

Contents

1. Introduction

The study of keystroke recognition has emerged as a significant interdisciplinary field, intertwining elements of computer science, cybersecurity, and human-computer interaction. It centers on analyzing typing patterns to identify unique characteristics in an individual's typing behavior. This analysis has diverse applications, from enhancing security protocols to user authentication and even behavioral studies.

Recent technological advancements have significantly influenced the development of this field. The integration of advanced computational methods and data analysis techniques has enabled researchers to extract and interpret complex patterns from keystroke data. These developments have not only improved the accuracy and reliability of keystroke recognition systems but have also broadened their potential applications.

In this context, our research project uniquely contributes by focusing on data captured through audio. This approach allows us to explore the rich acoustic information associated with keystrokes, providing a novel perspective in understanding and analysing typing patterns.

Despite these advancements, the field of keystroke recognition is continually evolving, with ongoing research seeking to refine methodologies and overcome existing limitations. Challenges such as dealing with variability in typing patterns, adapting to different keyboard layouts, and ensuring user privacy remain key areas of focus.

This State of the Art review aims to provide a comprehensive overview of the field of keystroke recognition. It will examine the fundamental concepts, trace the evolution of key methodologies, and discuss the current challenges and potential future directions of this research area. The review will serve as a foundation for understanding the broader context and significance of keystroke recognition, setting the stage for further discussion and exploration in subsequent sections.

2. State Of The Art

2.1. A Practical Deep Learning-Based Acoustic Side Channel

Firstly, we will delve into one of the articles we reviewed to gain more knowledge about this recognition field: *A Practical Deep Learning-Based Acoustic Side Channel*[1].

In the "Data Collection" phase of the study, the research team carefully selected equipment, opting for an iPhone 13 mini and Zoom software to record keystrokes. The choice of the iPhone 13 mini was driven by its microphone capabilities and widespread availability, ensuring practicality and relevance. Despite the inherent noise reduction features of Zoom software, the team innovatively developed "Algorithm 1: Zoom keystroke threshold setting" to effectively isolate keyboard keystrokes from audio recordings.

Emphasizing the significance of feature extraction and data processing, the study employed mel-spectrograms to visually represent sound waves. This approach not only captured the distinct acoustic characteristics of keystrokes but also facilitated seamless integration with deep learning models. By converting audio data into a visual format, the researchers bridged the gap between audio processing and image classification, showcasing the sophistication of their methodology. It's noteworthy that while Mel-frequency cepstral coefficients (MFCC) were considered, they were deemed less suitable due to the risk of losing relevant data in a context where human speech was not the target. A similar approach, avoiding the use of MFCC, will be adopted in our study for similar reasons.

In the Model Selection and Implementation section (3.2), the study elucidated the deployment of the CoAtNet

model, recognized for its prowess in image classification and efficiency in training. CoAtNet incorporates depth-wise convolutional layers and global relative attention layers, synergizing convolution and self-attention methods for rapid data processing and pattern recognition. The implementation was carried out using PyTorch.

Hyperparameter tuning played a pivotal role in optimizing model performance. In our MacBook keystroke classifiers' implementation, we adhered to specific hyperparameters. Models underwent 1100 epochs with a batch size of 16, adopting cross-entropy loss and the Adam optimizer. A maximum learning rate of 5e-4 was used, employing a linear annealing schedule. Data preprocessing involved time shifts at a rate of 40%, accompanied by masking up to 10%, with two masks per axis. Analysis encompassed 64 mel frequency bands, executed using a fast Fourier transform (FFT) window size of 1024 and a hop length of 225. Data was randomized and normalized during preprocessing.

In the results, the study achieved remarkable accuracy rates of 95% for phone-recorded keystrokes and 93% for Zoom-recorded keystrokes, marking a substantial advancement in the realm of acoustic side-channel attacks. These outcomes bear significant implications, underscoring the need for heightened cybersecurity awareness and proactive measures to safeguard sensitive data. The study's findings not only advance theoretical insights into acoustic side-channel attacks but also underscore practical applications. They shed light on alternative recording methods' potential and affirm the efficacy of mel-spectrograms and self-attention transformers in keystroke classification. Furthermore, the observation that false-classifications tend to cluster around the correct key on the keyboard hints at intriguing possibilities for real-world Acoustic Side-Channel Attacks (ASCA).

2.2. Don't Skype & Type! Acoustic Eavesdropping in Voice-Over-IP

Therefore, the *Don't Skype & Type! Acoustic Eavesdropping in Voice-Over-IP* [2] presents a keyboard acoustic eavesdropping attack that involves Voice-over-IP (VoIP) called Skype & Type. This is because, nowadays, people are engaged in many other activities while on call, and typing is one of these activities. It is robust to limited available bandwidth that degrades call quality, as well as human speech.

For the scenario, we have two people; it could be more, but let's consider just two people, one of them being an attacker and the other the victim. They are on a call on their computer, using Skype. The victim, during the call, types a message, and thanks to VoIP, the attacker can manage to retrieve the plain text.

The acoustic information can be degraded by VoIP software, but some VoIP software converts audio into a single signal, which solves the problem, such as Skype and Google Hangouts. Furthermore, Skype is one of the more popular VoIP services used on the planet. By focusing on it, they could target a large number of individuals, and the good news is the particularity of the other VoIP services doesn't significantly affect the results.

For the study, we considered three scenarios:

1. Complete profiling: Here, we have some of the victim's keyboard acoustic emanations. 2. User profiling: We don't have labelled data from the victim on the target device. But we can collect training data of the victim while the victim is using the same type of device (accomplice, for example). 3. Model profiling: There, we know nothing about the victim, and we shall try first to profile what laptop the victim is using. We should have a database of sounds from previous attacks (we know the type of laptop).

Target text is short, around 10 characters, which could be random (a dictionary is a particular case of random text). As features, they chose MFCC because, out of all the statistical properties of a sound spectrum, the MFCC resulted in 90.61

To perform the classification of the target device, they used a 10-NN classifier that outperformed other classifiers such as Random Forest and Logistic Regression (LR). Nevertheless, for key classification, it was LR that outperformed others (such as NN, LDA, RF) by scoring 90% for top-1 accuracy and 100% from 5 guesses, but it is very close to SVM.

The Skype & Type attack (S & T attack) can extract plain text from a short audio. As a result, for a waveform that lasts 70ms, we obtain 100

It was trained with three devices widely used: Lenovo, Mac Book, Toshiba. We had the worst results for Lenovo (60%) and (97% for Mac Book and 95% for Toshiba). This could be explained by the type of material; Lenovo laptop models are made of cheap plastic materials.

The voice, which overlaps the audio, could damage our experience by 20%, only up to 5dB, but for a normal call, it could not happen.

Aggressive downsampling and communication errors can greatly hinder the accuracy of the attacker, and samples could be lost (under 40Kbit/s), but we could assume that we work in a quality between 70-50Kbit/s, where there is no urgent damage.

Nevertheless, the attack is not infallible due to a duckling effect, and also if we perform short random transformations to the sound whenever a keystroke is detected. Even simpler, we could just mute the call while typing or not typing at all!

2.3. Keyboard Acoustic Emanations Revisited

In the article *Keyboard Acoustic Emanations Revisited*[3] the analysis is initially conducted on the same keyboard with the same user. Researchers based the identification of the keyboard-emitted sound on segmenting the recorded audio file into shorter audio files of individual characters. To achieve this segmentation, they use the Fast Fourier Transform (FFT), allowing them to identify the beginning of each keystroke on the keyboard. After this segmentation step, researchers calculate Mel-Frequency Cepstral Coefficients (MFCC) for a more precise characterization of each keystroke. This methodological choice stems from a previous study conducted by pioneers in the analysis of keyboard-emitted sounds, namely Asonov and Agrawal. In that study, they demonstrated that the emitted sounds yield better results when analyzed with MFCC rather than FFT. This process enables researchers to obtain vectors representing each keystroke.

In the second step, the clustering step, researchers use the Expectation-Maximization (EM) algorithm on Gaussian mixtures to group the vectors obtained earlier into K acoustic classes. For the choice of the value of K, the author mentions testing different values from 40 to 55 and found that K = 50 yielded the best results in their context. However, a problem arises: keystrokes of the same key are sometimes placed in different acoustic classes, and conversely, keystrokes of different keys can be in the same acoustic class. To provide more precision, they assign to each audio file of a keystroke a certain probability of being in an acoustic class instead of assigning a fixed class.

Once the conditional distribution step of the acoustic classes is completed, there are improvement steps. Researchers use a Hidden Markov Model (HMM) based on the English language, which allows predicting possible hidden states (letters that could follow the typed letters) based on observable states (the letters already typed). The model uses the correlation between consecutively typed keys. For example, if the current key can be either "h" or "j" (e.g., because they are physically close on the keyboard) and we know the previous key is "t," then the current key is more likely to be "h" because "th" is more common than "tj."

Another way to improve the results is by using the knowledge of the English language, using a spell checker called "Aspell." However, despite notable improvements, standard spell checkers are limited in the types of spelling errors they can handle, being constrained to at most two letters wrong in a word. Additionally, they are designed to handle common errors made by human typists, not the types of errors made by acoustic emanation classifiers.

Nevertheless, the spell checker will accept expressions such as "fur example" as correct spelling because "fur" is a dictionary word, even if the original phrase is probably "for example." Therefore, they added an n-gram language model, which takes into account the relative frequency of words and grammar issues. For example, some words are more common than others, and there are rules for forming phrases and expressions.

They applied this method to four different data sets. The first two data sets were conducted without background noise, with respective durations of 12m17 (containing 2514 keystrokes) and 26m56 (containing 5476 keystrokes). The latter two were conducted with surrounding noise, with respective durations of 21m49 (containing 4188 keystrokes) and 23m54 (containing 4300 keystrokes).

For the first data set, 35% of the words with 76% of the characters are correctly identified before improvement. After processing, there are 75% of correctly identified words and 87% of identified characters.

For the second data set, 38% of the words and 80% of the characters are correctly identified. After processing, there are 71% of correctly identified words and 87% of identified characters.

For the third data set, 32% of the words and 73% of the characters are identified. After processing, there are 57% of identified words and 80% of identified characters.

Finally, for the last data set, 23% of the words and 68% of the characters are correctly identified. After processing, there are 51% of correctly identified words and 75% of identified characters.

2.4. Reconnaissance de saisie sur clavier par analyse acoustique

In the article *Reconnaissance de saisie sur clavier par analyse acoustique*[4] The study looked at three distinct approaches to identifying keyboard keys from recorded sounds:

Temporal analysis relies on cross-correlation between signals to assess their similarity. By experimenting with various weightings and normalizations of the extracted signatures, the study showed that optimizing the weights of temporal signatures reduces recognition errors.

On the other hand, frequency analysis focuses on the spectral similarity of signals via the Discrete Fourier Transform (DFT). By normalizing energy signatures for a fair comparison of spectra, this approach uses the Euclidean distances between spectra to assign a key to each signal.

Finally, a combined method was developed by merging the DFT-based temporal and frequency approaches. Adjustments were made to balance the influence of these approaches, paying particular attention to a correction to balance exponent values. Manual tests identified optimal parameters for this combination, offering a better balance between the two approaches. The error rate was 0.3% on average, indicating a very small margin of error.

2.5. Analyse de la dynamique de frappe au clavier sonore pour l'identification

In the article *Analyse de la dynamique de frappe au clavier sonore pour l'identification, le profilage et l'extraction du texte saisi* [5] The project has two main focuses. Firstly, it aims to identify individuals from audio files by

segmenting the files into 10-second and 5-second segments. The approach is based on the use of pyAudioAnalysis in Python to create "SVM" files for each pair of individuals in a database. Initial test results showed around 90% success in identifying individuals from 40 10-second audio files per person, and around 96% success with 5-second audio files. However, these tests were carried out on a restricted database, justifying the need to extend the tests to larger databases to fully evaluate the robustness and performance of the algorithm, particularly when using 5-second audio files.

The second part of this project concerns the detection of keyboarded text from audio data. For this, a variety of audio data is collected, ranging from keyboard noises to human voice samples, as well as various ambient sounds. This data is separated into training and test sets. The preparation phase involves the use of audio files containing the sounds emitted by the keyboard during the input of a corpus of words, with associated text files indicating the precise moment of each keystroke (keylogger). These files are used to segment the audio files into letters, then into words, using the average duration of a keystroke to create new audio files corresponding to each letter. At the same time, audio files containing various noises are segmented and their characteristics, such as MFCC (Mel-Frequency Cepstral Coefficients), are extracted to differentiate keystrokes from other ambient sounds. Recognition of the typed text is performed using MFCC-based algorithms, which classify the audio segments and reconstruct the typed words from the detected key noises, comparing their similarity with a preparatory database to propose a list of the most likely words corresponding to the typed mystery word.

2.6. Keyboard Acoustic Emanations: An Evaluation of strong passwords and typing styles

In the article *Keyboard Acoustic Emanations: An Evaluation of strong passwords and typing styles*[6] The document explores the impact of typing style, type of input data, and detection techniques on the success of these eavesdropping attacks.

The study shows that typing style highly influences the acoustic signatures of keystrokes. The two primary styles, hunt and peck and touch typing, were analyzed. The study found that touch typing changes the sound of keystrokes to the point that it reduces the similarity between audio sounds of keys, thus lowering the risk of attacks.

The type of data being used plays an important role in the detectability of keystrokes. Strong passwords present a considerable challenge for acoustic detection due to their randomness and lack of contextual patterns. On the other hand, English text or weak passwords may yield higher detection rates due to language models and dictionary tools that can be used in conjunction with raw acoustic signals.

The document compares several detection techniques, including Dynamic Time Warping, signal time correlation, and a time-frequency classification approach. The study suggest that while changes in typing style affect the audio signal's amplitude, the emanated signals retain certain detectable similarities in both frequency and time domains, which can be used for improved detection accuracy.

The performance of password detection techniques was measured by their ability to accurately detect keystrokes and reconstruct strong passwords. Two techniques, cross-correlation (X-Corr) and time-frequency classification (Tim-Frq), were found to be most effective. The study also discusses the usefulness of using exhaustive search strategies for password recovery, highlighting the potential for such attacks to significantly reduce the time required for brute-force password cracking but still posing a limited threat under realistic conditions.

In conclusion, the document acknowledges the vulnerability of keyboard typing to audio emanations, influenced by typing style, input data, and detection technique. It suggests that acoustic eavesdropping poses a limited threat under realistic typing conditions and with strong passwords. However, there is still room for concern, as the order of password spaces can be reduced, thus expediting potential attacks. The document calls for future work on extending research to other character types and input scenarios, including the acoustic emanations of laptop keyboards.

3. Summary Table

Paper	Year	Principle	Accuracy(percentage)
Don't skype & type	2017	MFCC	91%
Don't skype & type	2017	LF	100%
A Practical Deep Learning-Based Acoustic Side Channel attack on keyboards	2023	CoAtNet	93%
Keyboard Acoustic Emanations Revisited	2009	MFCC/HMM	87% without any noise
Analyse de la dynamique de frappe au clavier sonore pour l'identification, le profilage et l'extraction du texte saisi	2022	SVM/MFCC	96%
Reconnaissance de saisie sur clavier par analyse acoustique	2011	Intercorrelation/DFT	99%
Keyboard Acoustic Emanations: An Evaluation of strong passwords and typing styles	Unknown	Tim-Frq	82.69%

Table 1. Comparing Accuracy of Different Keystroke Recognition Models and Approaches

4. Taxonomy

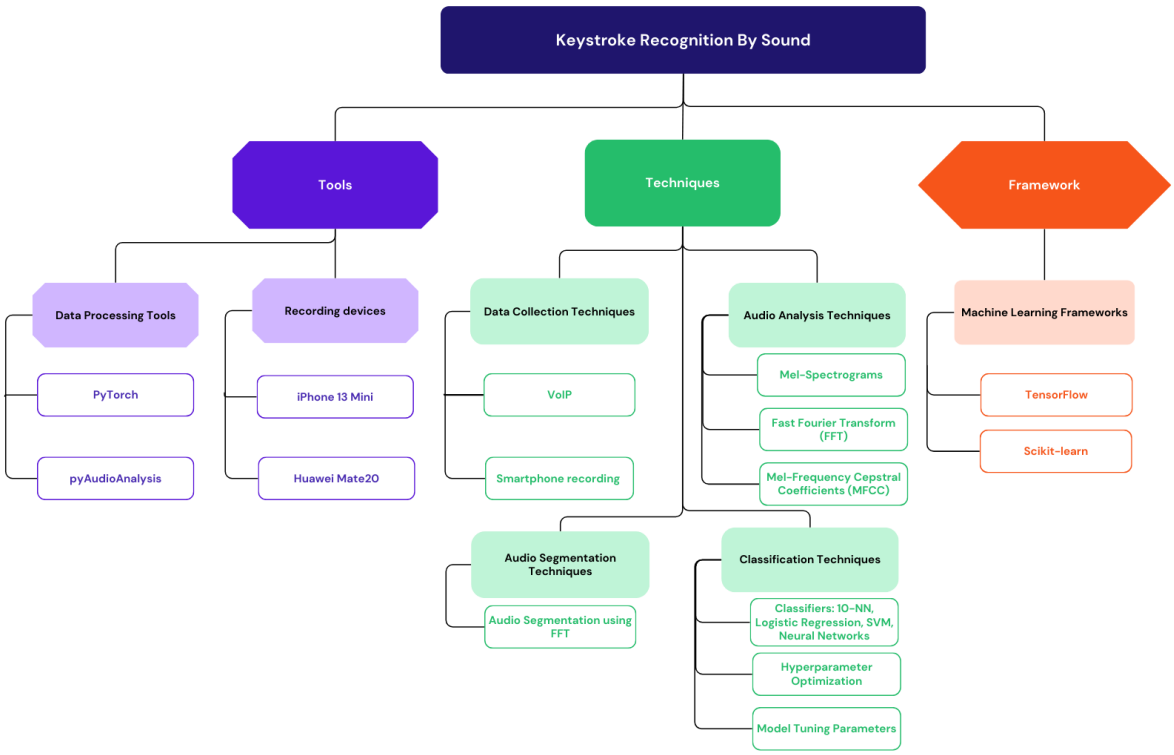

Figure 1. Keystroke Recognition by Sound Taxonomy

5. Conclusion

In a nutshell, our research strategy draws inspiration from the valuable insights gleaned from the studies we've examined. We have determined that utilizing smartphone microphones for audio recording serves as an optimal choice, primarily due to their widespread availability and the exceptional audio quality they can provide.

Building on the methodologies explored in the articles, we will initiate our investigation with a focused approach, beginning with word-by-word recognition. This initial step is designed to simplify the complexity of the task at hand, allowing us to establish a strong foundation for more intricate analysis at the keystroke level.

Our decision to incorporate the widely-used MFCC (Mel-Frequency Cepstral Coefficients) model aligns with the successful approaches observed in the studies we reviewed. By leveraging this model and the superior audio recording capabilities of smartphones, we aim to achieve a high degree of accuracy in our acoustic analysis.

Ultimately, our research endeavor is driven by the pursuit of a comprehensive understanding of acoustic side-channel attacks and their practical implications. We emphasize the importance of proactively addressing cybersecurity concerns in an ever-evolving digital landscape. Our study may unveil alternative recording methodologies and affirm the effectiveness of techniques such as MFCC and self-attention transformers in keystroke classification.

Furthermore, the potential for real-world Acoustic Side-Channel Attacks (ASCA) presents a captivating avenue for future exploration.

References

[1] Joshua Harrison, Ehsan Toreini & Maryam Mehrnezhad *A Practical Deep Learning-Based Acoustic Side Channel Attack on Keyboards* (Durham University, UK, 2023).

[2] Alberto Compagno, Mauro Conti, Daniele Lain & Gene Tsudik *Don't Skype & Type! Acoustic Eavesdropping in Voice-Over-IP* (ACM Asia Conference, 2017).

[3] Li Zhuang, Feng Zhou & J. D. Tygar *Keyboard acoustic emanations revisited* (University of California, Berkeley, year).

[4] Hala Mahmoud & Victor Malherbe *Reconnaissance de saisie sur clavier par analyse acoustique* (Internship raport, Centrale Supelec, 2011).

[5] Jarossay Max *Analyse de la dynamique de frappe au clavier sonore pour l'identification, le profilage et l'extraction du texte saisi* (Internship report, Greyc, 2022).

[6] Tzipora Halevi & Nitesh Saxena *Keyboard Acoustic Emanations: An Evaluation of strong passwords and typing styles* (press article, Polytechnic Institute of New York University, year).

