Biometric Authentication via Keystroke Sound

Joseph Roth   Xiaoming Liu   Arun Ross
Department of Computer Science and Engineering
Michigan State University, East Lansing MI 48824
{rothjos1,liuxm,rossarun}@msu.edu

Dimitris Metaxas
Department of Computer Science
Rutgers University, Piscataway, NJ 08854
dnm@cs.rutgers.edu

Abstract

Unlike conventional "one shot" biometric authentication schemes, continuous authentication has a number of advantages, such as longer time for sensing, ability to rectify authentication decisions, and persistent verification of a user's identity, which are critical in applications demanding enhanced security. However, traditional modalities such as face, fingerprint and keystroke dynamics, have various drawbacks in continuous authentication scenarios. In light of this, this paper proposes a novel non-intrusive and privacy-aware biometric modality that utilizes keystroke sound. Given the keystroke sound recorded by a low-cost microphone, our system extracts discriminative features and performs matching between a gallery and a probe sound stream. Motivated by the concept of digraphs used in modeling keystroke dynamics, we learn a virtual alphabet from keystroke sound segments, from which the digraph latency within pairs of virtual letters as well as other statistical features are used to generate match scores. The resultant multiple scores are indicative of the similarities between two sound streams, and are fused to make a final authentication decision. We collect a first-of-its-kind keystroke sound database of 45 *subjects typing on a keyboard. Experiments on static text-based authentication, demonstrate the potential as well as limitations of this biometric modality.*

1. Introduction

Biometric authentication is the process of verifying the identity of individuals by their physical traits, such as face and fingerprint, or behavioral traits, such as gait and keystroke dynamics. Most biometric authentication research has focused on *one shot* authentication where a subject's identity is verified only *once* prior to granting access to a resource. However, one shot authentication has a number of drawbacks in many application scenarios. These include short sensing time, inability to rectify decisions, and enabled access for potentially unlimited periods of time [17]. In contrast, *continuous* authentication aims to continuously verify the identity of a subject over an extended period of time thereby ensuring that the integrity of an application is not compromised. It can not only address the aforementioned problems in one shot authentication, but can also substantially enhance the security of systems, such as computer user authentication in security-sensitive facilities.

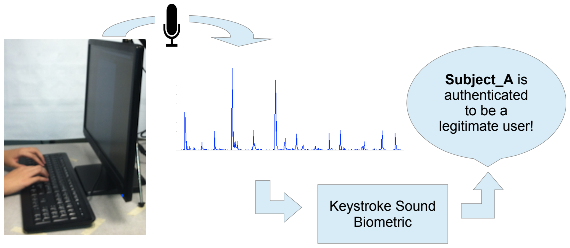

Subject_A is authenticated to be a legitimate user!
Keystroke Sound Biometric

Figure 1. The concept of *keystroke sound* as a biometric: The sound of a user typing on the keyboard is captured by a simple microphone attached to the PC and is input to our proposed system, which authenticates the user by verifying if the characteristic of the acoustic signals is similar to that of the claimed identity.

Research on continuous authentication is relatively sparse, and researchers have explored a limited number of biometric modalities, such as face, fingerprint, keystroke dynamics, and mouse movement, for this purpose. However, each of these modalities has its own shortcomings. For example, face authentication demands uninterrupted sensing and monitoring of faces - an intrusive process from a user's perspective. Similarly, the fingerprint sensor embedded on the mouse requires user collaboration in terms of precision in holding the mouse [15]. Although it is nonintrusive, keystroke dynamics utilizes key-logging to record typed texts and thus poses significant privacy risk in the case of surreptitious usage [13, 2].

In this paper, we consider another potential biometric modality for continuous authentication based on keystroke acoustics. Nowadays, microphone has become a standard peripheral device that is embedded in PCs, monitors, and webcams. As shown in Figure 1, a microphone can be used

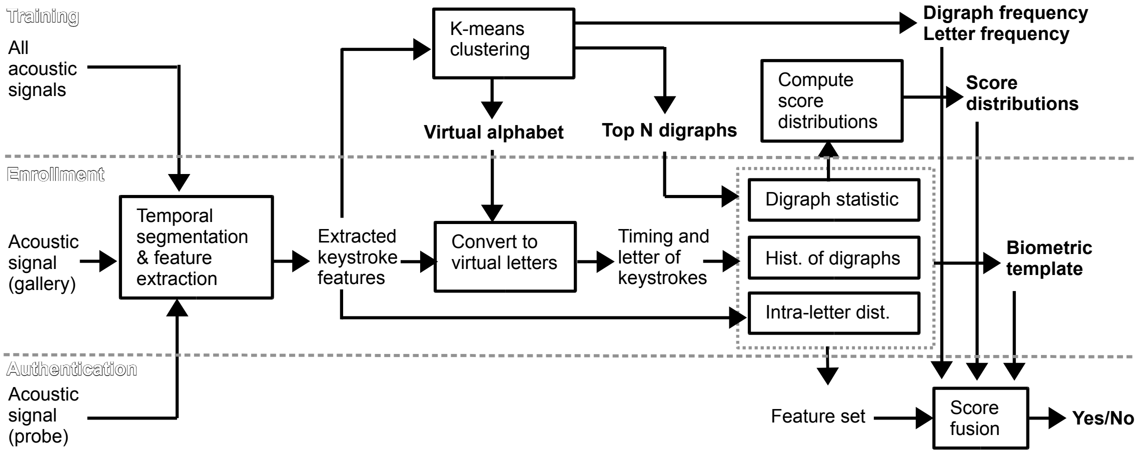


Figure 2. Architecture of the biometric authentication via keystroke sound, where the boldface indicates the output of each stage.

to record the sound of a user typing on the keyboard, which is then input to the proposed system for feature extraction and matching. Compared to the other biometrics modalities, it has a number of advantages. It utilizes a readily available sensor without requiring additional hardware. Unlike face or fingerprint, it is less intrusive and does not interfere with the normal computing operation of a user. Compared to keystroke dynamics, it protects the user's privacy by avoiding direct logging of any keyboard input. Keystroke acoustics does have the disadvantage of dealing with environmental noise, but proper audio filtering or a directed microphone should help in a noisy environment.

Our technical approach to employ this novel biometric is largely motivated by prior work in keystroke dynamics. One of the most popular features used to model keystroke dynamics is digraph latency [9, 8, 3], which calculates the time difference between pressing the keys of two adjacent letters. It has been shown that the word-specific digraph is much more discriminative than the generic digraph, which is computed without regard to what letters were typed [14]. Assuming that the acoustic signal from keystroke does not explicitly carry the information of what letter is typed, we propose a novel approach to employ the digraph feature by constructing a *virtual alphabet*. Given the acoustic signals from all training subjects, we first detect segments of keystrokes, whose mel-frequency cepstral coefficients (MFCC) [4] are fed into a K-means clustering routine. The resultant set of cluster centroids is considered as a virtual alphabet, which enables us to compute the most frequent digraphs (a pair of cluster centroids) and their latencies for each subject. Based upon the virtual alphabet, we also consider a number of other feature representation and scoring schemes. Eventually a score level fusion is employed to determine whether a probe stream matches with the gallery stream. In summary, this paper has three main contributions:

◇ We propose a novel keystroke sound-based biometric modality that offers non-intrusive, privacy-friendly, and potentially continuous authentication for computer users.

◇ We collect a first-of-its-kind sound database of users typing on a keyboard. The database and the experimental results are made publicly available so as to facilitate future research and comparison on this research topic.

◇ We propose a novel virtual alphabet-based approach to learn various score functions from acoustic signals, and a score-fusion approach to match two sound streams.

2. Prior Work

To the best of our knowledge, there is no prior work in exploring keystroke sound for the purpose of biometric authentication. As a result, we focus our literature survey on various biometric modalities for continuous authentication, and other applications of keystroke sound.

Face is one of the most popular modalities suggested for continuous user authentication, with the benefit of using existing cameras embedded in the monitor [15, 11]. However, continuously capturing face images creates an intrusive and unfavorable computing environment for the user. Similarly, fingerprint has been suggested for continuous authentication by embedding a fingerprint sensor on a specific area of the mouse [15]. This can be intrusive since it substantially constrains the way a user operates a mouse. Mouse dynamics has been used as a biometric modality due to the distinctive characteristics exhibited in its movement when operated by a user [12]. However, as indicated in a recent paper [18], more than 3 minutes of mouse movement on average is required to make a reliable authentication decision, which can be a bottleneck when the user does not continuously use the mouse for an extended period of time.

Keystroke dynamics utilizes the habitual patterns and rhythms a user exhibits while typing on a keyboard. Although it has a long history dating back to the use of telegraph in the 19th century and Morse Code in World War II,

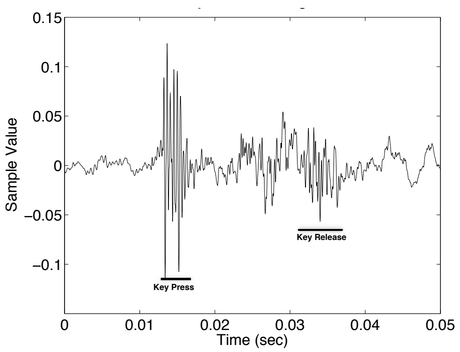


Figure 3. A raw acoustic signal containing one keystroke.

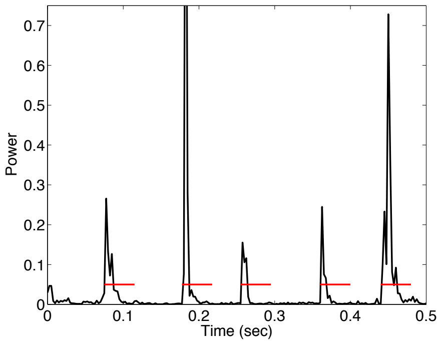


Figure 4. Temporal segmentation by thresholding the FFT power.

most of the prior work still focuses on static text [7, 19, 2], i.e., all users type the same text. Only a few efforts have addressed the issue of free text (i.e., a user can type arbitrary text) which is necessary for continuous authentication [10, 16]. Nevertheless, one major drawback of keystroke dynamics is the fact that it poses significant privacy risk to the user because all typed texts can be potentially recorded via key logging.

As far as recorded keystroke sound is concerned, there has been a series of work on acoustic cryptanalysis in the computer security community. The focus has been on designing advanced signal processing and machine learning algorithms to recover the typed letters from the keystroke sound [1, 20, 6]. One of their main assumptions is that when pressed, each key/letter will emit a slightly different acoustic signal, which motivates us to learn a *virtual* alphabet for our biometrics authentication application. Assuming the success of this line of work in the future, we may leverage it to combine the best of both worlds: keystroke dynamics based authentication using recovered keys and enhanced discrimination due to additional 1D acoustic signals over simple key logging signals.

3. Our Algorithm

As a pattern matching scheme, our algorithm seeks to calculate a similarity score between a probe sound stream recorded during the authentication stage and a gallery sound stream, whose features have been pre-computed and saved as a biometric template during the enrollment stage. Both sound streams are recorded when a user is typing on a keyboard with minimal background noise, and they are described in detail in Section 4.1. In this section, we present our algorithm for calculating the similarity score between two sound streams.

Figure 2 provides an overview of the architecture of our proposed algorithm. There are three different stages in our algorithm: training, enrollment, and authentication. In each stage, given a raw acoustic signal, we first isolate the keystroke portion from the silent periods in the temporal domain, and MFCC features are extracted from each of the resultant keystroke segments. During the *training* stage, we perform a K-means clustering of the MFCC features of all keystroke segments, where the centroids of clusters are considered as the virtual letters in a virtual alphabet. We also extract the digraph and letter frequencies, top N digraphs, and compute score distributions for fusion. In the *enrollment* stage, we use the virtual alphabet to convert the MFCC features into virtual letters, from which three different feature sets, viz., digraph statistic, histogram of digraphs and intra-letter distance, are designed to collectively form a biometric template for the user. Finally, in the *authentication* stage, the score functions are computed for a probe signal, and score-level fusion is used to make the final authentication decision. In the following we will present each component of this architecture in detail.

3.1. Temporal Segmentation & Feature Extraction

For an acoustic typing signal, $g(t)$, composed of keystrokes and silent periods, the first task before feature extraction is to identify portions in the stream where a keystroke occurs since it is assumed that the keystroke carries the discriminative information about an individual typist. As shown in Figure 3, a keystroke is composed of key pressing and key release, where the former comprises of the sound emitted when a finger touches the surface of a key as well as the sound due to the key being pressed down. We denote a keystroke as $\mathbf{k}_i$, with a start time of t_i and a duration of L. In our algorithm, L is set to be 40 ms, since it covers the full length of most keystrokes.

Motivated by [20], we conduct the temporal segmentation based on the observation that the energy of keystroke is concentrated in the frequencies between 400 Hz and $12K$ Hz, while the background noise occupies other frequency ranges. We compute the 5-ms windowed Fast Fourier Transform (FFT) of an acoustic signal using a sliding window approach, where the magnitudes of outputs in the range of 400 Hz and $12K$ Hz are summarized to produce an aggregate curve of the FFT power. By setting a threshold on the curve, we identify the start of keystrokes, as shown in Figure 4. Ideally this temporal segmentation should detect all keystrokes, instead of the background noise. Our current simple threshold-based method may not achieve this reliably as yet, and in the future we will investigate more advanced methods such as adaptive thresholding or supervised learning approaches.

Once the start of a keystroke t_i is determined, we convert the acoustic signal within a keystroke segment, $g(t_i, ..., t_i + L)$, to a feature representation $\mathbf{f}_i$ for future processing. The standard MFCC features are utilized due to its wide applications in speech recognition and demonstrated effectiveness in key recovery [20]. Our MFCC implementation uses the same parameter setup as [20]. That is, we have 32 channels of the Mel-Scale Filter Bank and use 16 coefficients and 32 filters with 10-ms windows that are shifted by 2.5 ms. The resultant feature of a keystroke $\mathbf{k}_i$ is a 256-dimensional vector $\mathbf{f}_i$.

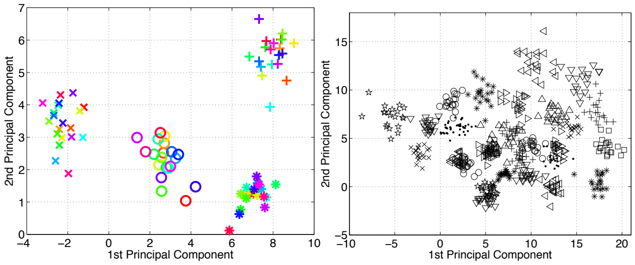
3.2. Virtual Alphabet via Clustering

Most of the prior work in keystroke dynamics uses digraphs statistics, which is the mean and standard deviation of delays between two individual keys or two groups of keys, or trigraphs, which is the delay across three keys, to model keystroke dynamics. It has also been shown that word-specific digraph, which is computed on the same two keys, is much more discriminative than the generic digraph [14]. In keystroke dynamics, such word information is readily available since key logging records the letter associated with each keystroke. However, this is not the case in our keystroke acoustic signal. Hence, in order to enjoy the benefit of digraph in our application, we strive to achieve two goals from the acoustic signal:

- Estimate the starting time of each keystroke;
- Infer or label the letter pressed at each keystroke.

While the first goal is addressed by the temporal segmentation process, the second goal requires knowledge of the pressed key in order to construct word-specific digraphs. However, as shown in acoustic cryptanalysis [6], precisely recognizing the typed key itself is an ongoing research topic. Hence, we take a different approach by aiming to associate each keystroke to a unique *label*, with the hope that different physical keys will correspond to different labels, and different typists will generate different labels when pressing the same key. We call each label a *virtual letter*, the collection of which is called a *virtual alphabet*. Learning the virtual alphabet is accomplished by applying K-means clustering to the MFCC features of all keystrokes in the training set.

Assuming the training set includes streams from J subjects, each with I_j keystrokes, the input of the K-means is $\{\mathbf{f}_i^j\}$, where $j \in [1, J]$ and $i \in [1, I_j]$. The K-means clustering partitions the input samples into K clusters, with centroids $\mathbf{m}_f(k)$. We set K to be 30 considering that the total number of common keys on a typical keyboard is around 30.



(a) First four virtual letters. (b) All 30 virtual letters.

Figure 5. The mean MFCC features $\bar{\mathbf{f}}_k$ of each of 20 subjects within virtual letters, plotted along the top-2 principle components reduced from the original 256-dimensional space. The symbol represents a letter and the color in (a) indicates a subject (best viewed in color).

During all three stages, given an original acoustic signal represented as a collection of keystrokes $\mathbf{K} = \{\mathbf{k}_i\}$, we compute the Euclidean distance between the MFCC feature of a keystroke $\mathbf{f}_i$ and each of the cluster centroids. The centroid with the minimal distance will assign its index to the keystroke as its corresponding virtual letter, i.e., $l_i = k$. Thus, in our algorithm a virtual letter is simply a number between 1 and K. Finally, we can represent a keystroke as a triplet $\mathbf{k}_i = \{t_i, \mathbf{f}_i, l_i\}$.

3.3. Score Functions

Given the aforementioned feature representation scheme, we next investigate a set of score functions to compute the similarity measure between two sets of features, listed as follows.

Digraph statistic As a representation of individual typing characteristics, digraph statistic refers to the statistics of the latency, in terms of mean and stand deviation, within frequent pairs of letters. A pair of letters is named digraph, with examples such as *(t,h)*, *(h,e)*. In our algorithm, a virtual alphabet with K letters can generate K^2 digraphs, such as $(2, 5)$, $(7, 1)$. During the training stage, we count the frequency of each digraph by passing through adjacent keystrokes, l_{i-1}^j and l_i^j, in the entire training set. Then we generate a list of top N most frequent digraphs, each denoted as $\mathbf{d}_n = \{k_{n1}, k_{n2}\}$. Given the $\{\mathbf{k}_i\}$ representation of an acoustic signal, for each one of N digraphs, we then compute the mean, $\{m_n\}$, and the standard deviation, $\{\sigma_n\}$, of the time difference variable $\Delta t = t_i - t_{i-1}$ where $l_i = k_{n2}$ and $l_{i-1} = k_{n1}$. Finally, the similarity score between two arbitrary length signals, $\mathbf{K}$ and $\mathbf{K}'$, is computed using the following equation:

$$S_1(\mathbf{K}, \mathbf{K}') = \sum_{n=1}^{N} w_1^n \left[\sum_{\Delta t} \sqrt{\mathcal{N}(\Delta t; m_n, \sigma_n^2) \mathcal{N}(\Delta t; m_n', {\sigma_n'}^2)} \right], \tag{1}$$

```
Input: A stream $g(t)$, top digraphs $\mathbf{d}_n$, cluster
       centroids $\mathbf{m}_f(k)$.
Output: A feature set $\mathbf{F}$.
Locate keystrokes $\mathbf{t} = [t_1, ..., t_i, ...]$ at times of high
energy using FFT,
foreach keystroke time $t_i$ do
    $\mathbf{f}_i$ = MFCC($g(t_i, ..., t_i + L)$),
    $l_i = argmin_k \|\mathbf{m}_f(k) - \mathbf{f}_i\|_2$,
foreach digraph $\mathbf{d}_n$ do
    $\mathbb{T} = \{t_i - t_{i-1} : l_i = k_{n2} \ \& \ l_{i-1} = k_{n1}, \forall i \in$
    $[2, |\mathbf{t}|]\}$,
    $m_n$ = mean($\mathbb{T}$),
    $\sigma_n$ = std($\mathbb{T}$),
    Compute histogram of digraphs $h_n$ via Eqn. (2),
foreach letter $k$ do
    Compute $\bar{\mathbf{f}}_k$ via Eqn. (4),
return $\mathbf{F} = \{m_n, \sigma_n, h_n, \bar{\mathbf{f}}_k\}$.
```

Algorithm 1: Feature extraction algorithm.

which basically summarizes the overlapping region between two Gaussian distributions of the same digraph, weighted by w_1^n. The overlapping region is computed via the Bhattacharyya coefficient, and w_1^n is the overall frequency of each digraph learned from the training set. We set N to be 50 in our algorithm.

Histogram of digraphs In our virtual alphabet representation, different subjects may produce different virtual letters when pressing the same key. This implies that different subjects are likely to generate different digraphs when typing the same word. Hence, it is expected that we can use the frequencies of popular digraphs as a cue for authentication. Given this, we compute the histogram $\mathbf{h} = [h_1, h_2, ..., h_N]^T$ of top N digraphs, which are the same as the ones in digraph statistic, for each acoustic signal. That is,

$$h_n = \frac{\sum_i \delta(l_i = k_{n2})\delta(l_{i-1} = k_{n1})}{|\mathbf{K}| - 1}, \tag{2}$$

where δ is the indicator function and the numerator is the number of digraph $\mathbf{d}_n = \{k_{n1}, k_{n2}\}$ within a sequence of length $|\mathbf{K}|$. The similarity score between two signals is simply the inner product between two histograms of digraphs,

$$S_2(\mathbf{K}, \mathbf{K}') = \mathbf{h}^\mathbf{T}\mathbf{h}'. \tag{3}$$

Intra-letter distance We use the virtual letter to represent similar keystrokes emerging from different keys pressed by different subjects. Hence, within a virtual letter, it is very likely that different subjects will have different distributions. Figure 5 provides evidences for this observation by showing the mean MFCC features of 20 training subjects within virtual letters. It can be seen that 1) there is distinct inter-letter separation among virtual letters; 2) within each virtual letter, there is intra-letter separation due to individuality. Hence, we would like to utilize this intra-letter separation in our score function. For an acoustic signal, we compute the mean of $\mathbf{f}_i$ associated with each virtual letter, which results in $K = 30$ mean MFCC features, as follows:

```
Input: A probe stream $g'(t)$, biometric template $\mathbf{F}$,
       top digraphs $\mathbf{d}_n$, cluster centroids $\mathbf{m}_f(k)$,
       score distribution $m_{sv}, \sigma_{sv}$, threshold $\tau$.
Output: An authentication decision $d$.
Compute feature set $\mathbf{F}'$ for probe $g'(t)$ via Alg. 1,
Compute digraph statistic score $S_1$ via Eqn. (1),
Compute histogram of digraphs score $S_2$ via Eqn. (3),
Compute intra-letter distance score $S_3$ via Eqn. (5),
Compute normalized score $S$ via Eqn. (6),
if $S > \tau$ then
    return $d$ = genuine.
else
    return $d$ = impostor.
```

Algorithm 2: Authentication algorithm.

$$\bar{\mathbf{f}}_k = \frac{1}{|l_i = k|} \sum_{l_i = k} \mathbf{f}_i. \tag{4}$$

Given two acoustic signals $\mathbf{K}$ and $\mathbf{K}'$, we use Equation (5) to compute the Euclidean distance between the corresponding means and sum up by using a weight w_3^n, which is the overall frequency of each virtual letter among all keystroke segments and is computed from the training set. The sign -1 is to make sure that, on average, the genuine probe has a larger score than the impostor probe.

$$S_3(\mathbf{K}, \mathbf{K}') = -\sum_{k=1}^{K} w_3^n ||\bar{\mathbf{f}}_k - \bar{\mathbf{f}}'_k||_2. \tag{5}$$

So far we have constructed a new feature set for one acoustic signal, denoted as $\{m_n, \sigma_n, h_n, \bar{\mathbf{f}}_k\}$ where $n \in [1, N]$ and $k \in [1, K]$. We summarize the feature extraction algorithm in Algorithm 1. If the acoustic signal is from the gallery stream, we call the resultant feature set as a *biometric template* of the gallery subject, which is computed during enrollment and stored for future matching with a probe stream.

Score fusion Once the three scores are computed, we fuse them to generate one value to determine whether the authentication claim should be accepted or rejected. In our system we use a simple score-level fusion where the three normalized scores are added to create the overall similarity score between two streams,

$$S = \sum_{v=1}^{3} \frac{S_v - m_{sv}}{\sigma_{sv}}. \tag{6}$$

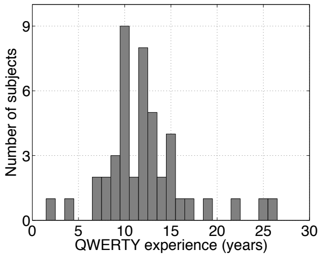


Figure 6. Distribution of subjects' experience with keyboard.

The score normalization is conducted by using the mean m_{sv} and standard deviation σ_{sv} of the score distribution learned from the training data, such that the normalized scores fall in a standard normal distribution. We summarize the algorithm during the authentication stage in Algorithm 2. In the future, more advanced fusion mechanisms can be utilized, especially by leveraging previous work in the biometrics fusion domain [5].

4. Experiments

In this section, we present an overview of the database collected for this experiment which can be used for other typing based biometrics. We also present our specific experimental setup and the results of our static-text experiments.

4.1. Keystroke Sound Database

Since keystroke sound is a novel biometric modality without any prior database, we develop a capture protocol that ensures the collected database is not only useful for the current research problem, but also beneficial to other researchers on the general topic of typing-based biometrics.

When developing the protocol, our first goal was to study whether the keystroke sound should be based on static text or free text, i.e., does a subject have to type the same text to be authenticated? The same question was addressed in keystroke dynamics, where there is substantially more research effort on static text than free text [2]. To answer this question, we request each subject to perform typing in two sessions. In the first session, a subject types one paragraph four times (the first paragraph of "A Tale of Two Cities" by Charles Dickens displayed on the monitor), with a 2–3 seconds break between trials. Multiple typing instances of the same paragraph enable the study of static text based authentication. In the second session, a subject types a half-page letter with arbitrary content to his/her family. It can be seen from this session that most users make spontaneous pauses during typing, which mimics well the real-world typing scenario. Normally a subject spends around 5–8 minutes on each session, depending on the typing speed.

A second consideration is the recording environment. The background environment plays an important role in the usability of the data. Low frequency pitches from computers, heaters, and lights can create distractions to the sound of the actual typing. The placement of the sensor relative to the keyboard as well as in the room can also introduce differences due to echoes. All subjects in the study performed the typing with the same computer, keyboard, and microphone, in the same room, under reasonably similar conditions. Care was taken to use non-verbal communication during the experiments to eliminate human speech from corrupting the audio stream. Nevertheless, standard workplace noises still exist in the background, e.g. doors opening, chairs rolling, and people walking.

Age	10–19	20–29	30–39
Number of subjects	11	30	4

Table 1. Age distribution of subjects.

The third consideration is the hardware setup. We use a US standard QWERTY keyboard for data collection. Although there are many available options for microphones, we decide to utilize an inexpensive webcam with embedded microphone, which is centered on the top of the monitor and pointed toward the keyboard. This setup allows us to capture both the video of hand movement and the audio recording of keyboard typing. Thus, a multi-modal (visual and acoustic) typing-based continuous authentication system can be developed in the future based on this database. The sound is captured at $48,000$ Hz with a single channel.

Thus far our database contains 45 subjects with different years of experience in using the keyboard. The subjects are students and faculty at Michigan State University, whose typing experience and age are displayed in Figure 6 and Table 1. The database of keystroke sound is available at http://www.cse.msu.edu/~liuxm/typing, and is intended to facilitate future research, discussion, and performance comparison on this topic.

4.2. Results

We use the static text portion of the database for our experiments and leave the use of the free text for future work. Our algorithm requires a separate training dataset for the purpose of learning a virtual alphabet, top digraphs, and the statistics of score distributions. Hence, we randomly partition the database into 20 subjects for training and the remaining 25 subjects for testing our algorithm.

For each subject, we use the first typing trial as the gallery stream and portions of the fourth typing trial as the probe stream. We choose the first and last paragraph to maximize the intra-subject differences for the static text. For the probe streams, we need to balance two concerns. Firstly, we want to use as many probes as possible to enhance the statistical significance of our experiments, which requires that we use a shorter partial paragraph. Secondly, we want to use longer probe streams to allow accurate calculations of features for a given subject. We decide to form 7 continuous probe streams for each user by using 70% of the fourth paragraph starting at the 0%, 5%, 10%, 15%,

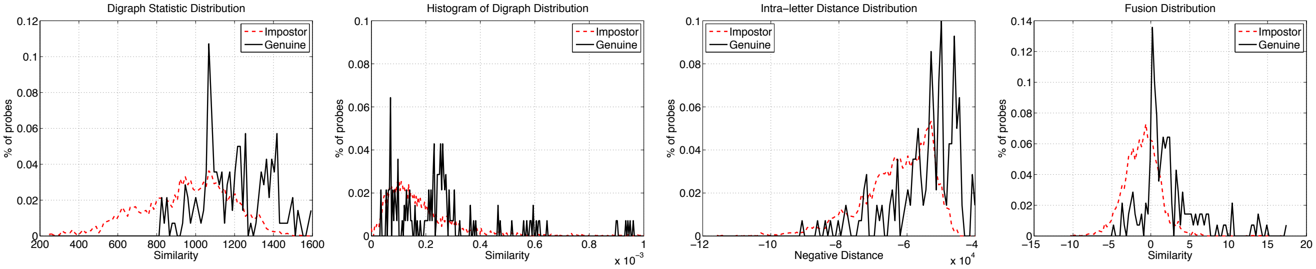


Figure 7. Distribution of the individual score functions, S_1, S_2 and S_3, and the fused score S, for genuine and impostor probes.

20%, 25%, and 30% mark of the paragraph. This overlap of text streams allows us to balance both concerns, while also simulates a continuous environment where the algorithm works with an incoming data stream and the results of the current probe build on the results of the previous probes. The average probe length is 62 seconds. The total number of training probe streams is 140 with 3500 training cases. The total number of testing probe streams is 175 with 4375 testing cases.

Evaluation metrics We use the standard biometrics authentication metric, the ROC curve, to measure performance. The ROC curve has two axes, False Positive Rate (FPR), the fraction of impostor pairs incorrectly deemed to be genuine, and True Positive Rate (TPR), the fraction of genuine pairs correctly deemed to be genuine. We use the Equal Error Rate (EER) to succinctly summarize the ROC curve. We also use the probability distributions of the genuine and impostor scores to study the performance.

Feature performance comparison Figure 7 illustrates the separability of the three score functions and the overall fused score on the testing data. Figure 8 displays the same information in the standard ROC format. We can make a number of observations. Firstly, the individual score function distributions all display overlap between the genuine and impostor probes. The task of identifying a single feature representation scheme to authenticate users via keystroke sounds proves challenging. Digraph statistics, histogram of digraphs, and intra-letter distance provide 31%, 32%, and 33% EER, respectively. Secondly, despite the overlap, there is still some separation between the genuine and impostor probes. We see Gaussian-like distribution for digraph statistics and intra-letter distance, which implies that in real-world applications with unseen data, the proposed score functions will be useful. The histogram of digraphs displays a bimodal Gaussian distribution for the genuine class, which indicates that it will behave on the extremes of either providing very useful information or limited discriminability. Furthermore, by using fusion, we create a new fused score, which produces better results and indicates that the individual score functions capture different aspects of the subject typing behavior. The result with the fused score has an EER of 25%.

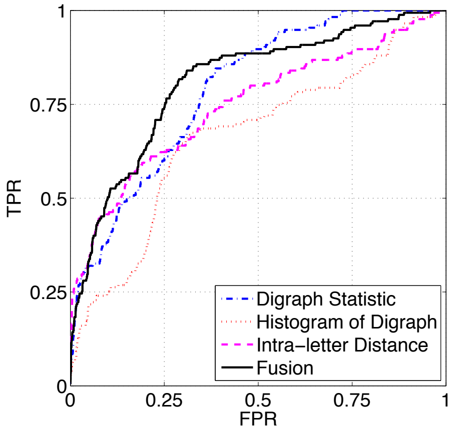


Figure 8. ROC curves of individual score functions as well as the final fused score.

Computation efficiency Since a probe stream is an one-dimensional signal, our algorithm can operate in real time, with very minimal CPU load, which is a very favorable property for continuous authentication. For a 60-second probe stream, it takes approximately 20 seconds to create the feature representation with the majority of the time spent identifying the locations of keystrokes. Once the features have been extracted, it takes less than 0.1 seconds to compute the score functions against a biometric template. A future work is to design an incremental score function. This is important since we would like to perform authentication in the online mode, as the sound stream is continuously received.

5. Conclusions

In this paper, we presented keystroke sound as a potential biometric for use in continuous authentication systems. The proposed biometric is non-intrusive and privacy-friendly. It can be easily implemented with standard computer hardware and requires minimal computational overhead. To facilitate this research, we collected both acoustic and visual typing data from 45 individuals. We designed multiple approaches to compute match scores between a gallery and probe keystroke acoustic stream. In particular, we proposed a fusion of digraph statistics, histogram of digraphs, and intra-letter distances to authenticate a user. We obtained an initial EER of 25% on a database of 45 subjects. This suggests that more sophisticated features have to be investigated in the future. While these are preliminary results, with the accuracy far below well-researched biometric modalities such as face or fingerprint, we wish to emphasize that the intent of this feasibility study is to open up a new line of exciting research on typing-based biometrics. We anticipate other interested researchers to commence working on keystroke acoustics as a biometric modality - either by itself or in conjunction with other modalities - in continuous authentication applications.

Our future work on this topic will cover the following aspects. First, we will increase the subject size of our database to provide more training samples. We will continue to collect data from a larger number of users as well as acquire more samples from the current users to gain a better understanding of the inter- and intra-class variability of this biometric. Second, we will explore methods for better identifying keystrokes through supervised learning or context-sensitive thresholding. This will allow for more robust authentication in the presence of background noise typical of a normal work environment. Third, we will explore the changes that occur when comparing free text, by utilizing the second session of our database. Finally, we will investigate new feature representation schemes to address the intra-class variability of this biometric.

It is interesting to note a potential attack on this biometric system that involves recording a genuine user's typing sounds and then replaying the recording while the impostor is using the computer in a silent manner through the mouse. A means of foiling this attack is to log only the times of keystrokes and ensure alignment of the detected keystroke times from the audio with the actual keystroke times. Alternatively, positioning a webcam toward the keyboard to analyze the video of typing can foil this attack as well. It is also desirable to completely fuse this method with standard keystroke dynamics methods where the audio and virtual digraphs add extra discriminating information to the physical digraphs.

References

[1] D. Asonov and R. Agrawal. Keyboard acoustic emanations. In *Proc. of IEEE Symposium on Security and Privacy*, pages 3 – 11, May 2004. 3

[2] S. P. Banerjee and D. Woodard. Biometric authentication and identification using keystroke dynamics: A survey. *J. of Pattern Recognition Research*, 7(1):116–139, 2012. 1, 3, 6

[3] F. Bergadano, D. Gunetti, and C. Picardi. User authentication through keystroke dynamics. *ACM Trans. Inf. Syst. Secur.*, 5(4):367–397, Nov. 2002. 2

[4] S. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Trans. Acoustics, Speech and Signal Processing*, 28(4):357 – 366, Aug 1980. 2

[5] A. Jain, A. Ross, and S. Prabhakar. An introduction to biometric recognition. *IEEE Trans. CSVT*, 14(1):4 – 20, Jan. 2004. 6

[6] A. Kelly. *Cracking Passwords using Keyboard Acoustics and Language Modeling*. Master thesis, University of Edinburgh. 2010. 3, 4

[7] K. S. Killourhy and R. A. Maxion. Comparing anomaly detectors for keystroke dynamics. In *Proc. of 39th International Conference on Dependable Systems and Networks (DSN-2009)*, pages 125–134, 2009. 3

[8] F. Monrose, M. Reiter, and S. Wetzel. Password hardening based on keystroke dynamics. *International Journal of Information Security*, 1(2):69–83, 2002. 2

[9] F. Monrose and A. Rubin. Authentication via keystroke dynamics. In *Proceedings of the 4th ACM conference on Computer and communications security*, pages 48–56, 1997. 2

[10] T. Mustafic, S. Camtepe, and S. Albayrak. Continuous and non-intrusive identity verification in real-time environments based on free-text keystroke dynamics. In *ICJB*, 2011. 3

[11] K. Niinuma, U. Park, and A. K. Jain. Soft biometric traits for continuous user authentication. *IEEE Trans. Information Forensics and Security*, 5(4):771–780, Dec. 2010. 2

[12] K. Revett, H. Jahankhani, S. T. Magalhes, and H. M. D. Santos. *Global E-Security*, volume 12 of *Communications in Computer and Information Science*, chapter A Survey of User Authentication Based on Mouse Dynamics, pages 210–219. Springer Berlin Heidelberg, 2008. 2

[13] D. Shanmugapriya and G. Padmavathi. A survey of biometric keystroke dynamics: Approaches, security and challenges. *International Journal of Computer Science and Information Security*, 5(1):115–119, 2009. 1

[14] T. Sim and R. Janakiraman. Are digraphs good for free-text keystroke dynamics? In *CVPR, Workshop on Biometrics*, 2007. 2, 4

[15] T. Sim, S. Zhang, R. Janakiraman, and S. Kumar. Continuous verification using multimodal biometrics. *IEEE T-PAMI*, 29(4):687–700, Apr. 2007. 1, 2

[16] C. C. Tappert, S.-H. Cha, M. Villani, and R. S. Zack. A keystroke biometric systemfor long-text input. *Int. J. of Information Security and Privacy (IJISP)*, 4(1):32–60, 2010. 3

[17] M. Turk. Grand challenges in biometrics. Presented as keynote at IEEE Int. Conf. on Biometrics: Theory, Applications and Systems (BTAS 10), Washington, DC, 2010. 1

[18] N. Zheng, A. Paloski, and H. Wang. An efficient user verification system via mouse movements. In *Proc. of the 18th ACM conference on Computer and communications security*, pages 139–150, New York, NY, USA, 2011. ACM. 2

[19] Y. Zhong, Y. Deng, and A. K. Jain. Keystroke dynamics for user authentication. In *CVPR, Workshop on Biometrics*, 2012. 3

[20] L. Zhuang, F. Zhou, and J. D. Tygar. Keyboard acoustic emanations revisited. In *Proc. of the 12th ACM conference on Computer and communications security*, pages 373–382, 2005. 3, 4