Continuous Authentication with Cognition-centric Text Production and Revision Features*

Hilbert Locklear† hlocklea@nyit.edu
Sathya Govindarajan† sathyag413@gmail.com
Zdeňka Sitová† sitovaz@mail.muni.cz
Adam Goodkind‡ a.goodkind@gmail.com
David Guy Brizan‡ DBrizan@gc.cuny.edu
Andrew Rosenberg‡ andrew@cs.qc.cuny.edu
Vir V. Phoha§ phoha@latech.edu
Paolo Gasti† pgasti@nyit.edu
Kiran S. Balagani† kbalagan@nyit.edu

†New York Institute of Technology ‡City University of New York Graduate Center §Louisiana Tech Univeristy

Abstract

Most continuous user authentication techniques based on typing behavior rely on the keystroke dynamics or on the linguistic style of the user. However, there is a rich spectrum of cognition-centric behavioral traits that a typist exhibits during different stages of text production (e.g., composition, translation, and revision), which to our knowledge, have not been considered for continuous authentication. We study the continuous authentication performance of 123 behavioral traits extracted from discrete cognitive units called bursts. We performed experiments on typing data collected from 486 volunteer subjects. Our findings include: (1) features from bursts delimited by pause events have significantly higher availability and authentication performance compared to bursts delimited by revision events; (2) bursts with pause durations of at least one second provide the best authentication accuracy and availability; and (3) fusing our features with traditional keystroke dynamics features reduced authentication error rates. We achieved an equal error rate between 13.37 and 4.55 percent for authentication windows as low as 30 seconds to 3.5 minutes.

1. Introduction

In typing-based *continuous* authentication, two types of features[1] have been predominantly studied—(1) keystroke dynamics [2], where authentication is based on *the timing of a user's key presses and releases* and (2) stylometry [3, 4], where authentication is based on *the linguistic style of the text produced by the user*. What differentiates our work from the state of the art is:

1. We approach the problem of typing-based continuous authentication by viewing the *typist* as a performer of the following text-generation tasks [5, 6] (not necessarily in this order): composition, translation, transcription, evaluation, and revision. Under this view, keystroke dynamics and stylometry are a *subset* of a rich spectrum of cognition-centric and possibly idiosyncratic behaviors a typist exhibits during composing, translating, transcribing, evaluating, and revising text. Because an individual acquires and gradually evolves these behaviors over time (through repetition or practice), we hypothesize these traits are intrinsic to the individual.
2. Because cognition-centric features can be sensitive to the context in which they are produced, we associate the features with cognitive contexts called *bursts*. A burst indicates the activity performed by the user (e.g., did the user produce or revise text?) and the amount of cognitive effort expended by the user in performing the activity. The features, extracted from bursts, indicate how the user performed the activity.

Contributions. The three main contributions of our paper are: (1) We present 123 cognition-centric features measured across four types of cognitive activity bursts. To our knowledge, ours is the first work to harness cognitive contexts for continuous behavioral authentication. (2) We present authentication performance of our features. Our results show that features associated with *pause-delimited*

*This work was supported in part by DARPA Active Authentication grants FA8750-12-2-0201 and FA8750-13-2-0274 and NYIT ISRC 2012 and 2013 grants. The views, findings, recommendations, and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the sponsoring agencies or the U.S. Government. Z. S. is a student of Faculty of Informatics, Masaryk University, Czech Republic. This work was done while Z. S. was visiting NYIT.

[1] A recent work [1] has also used keystroke acoustics.

bursts have lower equal error rates compared to *revision-delimited* bursts. Results also show that fusing our features with keystroke dynamics (hold and digraph latencies) *reduced* equal error rates, even when the authentication latencies were as low as 30 and 60 seconds. Thus we demonstrate that our features complement the discriminability offered by keystroke dynamics. (3) We present availability analysis of our features on free-text data collected from 486 users. We show that pause-delimited burst features have higher availability than revision-delimited burst features, in terms of two metrics: (i) proportion of authenticatable users and (ii) availability relative to key hold latencies.

Organization of Our Paper. In Section 2, we present the details on bursts. In Section 3, we discuss 123 behavioral features. In Section 4, we present details of experiments. In Section 5, we present availability and authentication performance results. In Section 6, we present related research. We conclude our work in Section 7.

2. Burst As a Unit of Cognitive Activity

A writer produces text in short multi-word segments called *bursts* [5, 7]. In this work, a burst is delimited by the occurrence of two events:[2] (1) *pause*—where it is assumed that the typist has momentarily paused to produce new text or evaluate previously typed text and (2) *pause followed by revision*—where it is assumed that the typist has paused to revise previously typed text. We define two events:

Pause (P)—If the duration between the release of a key and the press of next key is greater than or equal to p seconds, we consider the user to have paused.

Pause–Revision (PR)—If a pause of at least p seconds is followed by a revision (press of Backspace key at least m consecutive times), we consider the user to have paused to revise text. Identifying a PR event can be complicated by the many ways in which a user may pause and revise. We mitigate this complexity by utilizing two variables:

- *Revision Depth*—indicates the amount of leading edge revisions done by the user in continuously pressing the BACKSPACE key at least m times. Depth represents the intensity of the revision.
- *Revision Distance*—indicates the immediacy of revision after the occurrence of a pause. We measure this as the number of characters that fall between a pause and at least m contiguous BACKSPACE key presses.

We detect a PR if a revision of depth m occurs (i) after a pause of at least p seconds and (ii) within a distance d from the pause (p, d, and m are parameters). Below, we give an example to show how we detect PR. For illustration, we set our criteria for detecting a PR event as $d \leq 1$, $m \geq 2$, and $p \geq 2000$ms.

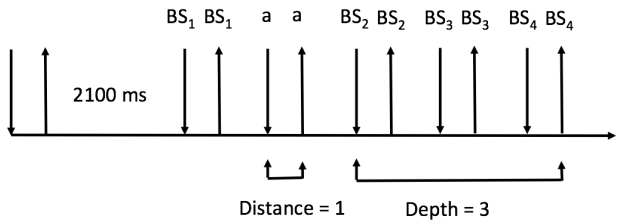

Figure 1. Detection of a PR event, where pause duration $p \geq$ 2000ms, revision distance $d \leq 1$, and revision depth $m \geq 2$.

In Figure 1, we give an example in which a PR event satisfies our criteria: pause of 2100ms, followed by a revision of depth 3 occurring within revision distance 1. The down arrows indicate key press and the up arrows indicate release. Note that the distance is 1 because the first BACKSPACE key, BS_1, is not a character and therefore not counted towards distance. The depth is 3 because there are three consecutive BACKSPACE key presses (BS_2, BS_3, and BS_4).

Burst Classification. Using P and PR events, we classify a burst distinctly into one of the four following types.

- PP: A burst which begins and ends with pauses (no revision involved).
- PR-PR: A burst which begins with a PR and ends with PR.
- PR-P: A burst which begins with PR and ends with P.
- P-PR: A burst which begins with a P and ends with PR.

Additionally, we refer to PR-P, P-PR, and PR-PR bursts collectively as "Combined Revision" burst.

Motivation for Using Bursts in Authentication. As postulated by Baaijen et al. [7] and assumed by many studies in written discourse (e.g., [5]), a pause is directly related to the conscious, intellectual activity of the user in regards to producing subsequent text or in evaluating the previous text. By extracting 123 features within bursts, we aim to capture the different aspects of text generation (composition, translation, transcription, evaluation, and revision) in relation to the cognitive effort expended by the user.

3. Cognition-centric Behavioral Features

From each burst, we extract two types of features: (1) pre- and post-pause features and (2) text production and revision features. Brief description of the features follow.

[2] In theory, many delimiters can be used to identify a burst (e.g., period, semi-colon, etc.), however, in this study we use pause and revision events because they are considered as good representations of cognition in the field of written communication.

3.1. Pre- and Post-Pause Features

These features aim to capture how the user produces, evaluates, or revises text in a burst, in relation to the amount of time the user paused before (*pre-pause*) or after (*post-pause*) producing the burst. We used the following 8 features for both pre-pause and post-pause, thus generating a total of 16 features. "Pause" below represents both pre-pause and post-pause.

Typing Speed: 2 features—(i) the ratio of the number of key presses to the duration of the pause and (ii) the ratio of alphanumeric key presses to the duration of the pause.

Token Count Duration: 1 feature—the ratio of the number of identifiable unique words in the burst to the duration of the pause.

Lexical Density: 1 feature—measures the content per grammatical and lexical unit as defined by the ratio of the number of word types to the number of all tokens. This measure is taken as a ratio to the duration of the pause.

Parts of Speech: 4 features—(i) the ratio of the number of recognizable verbs in the burst to the duration of the pause, (ii) the number of recognizable nouns in the burst to the duration of the pause, (iii) the number of recognizable modifiers in the burst to the duration of the pause, and (iv) the ratio of the number of recognizable modals in the burst to the duration of the pause

3.2. Text Production and Revision Features

We examine a burst for properties of word creation, lexical complexity, revision count, and keyboard proficiency. Within the four properties, we identify sub-properties, which allowed us to generate fine-grained features for achieving greater discriminability between users.

3.2.1 Word Creation

Word Creation is the measure of the token (group of characters delimited by whitespace and/or punctuation) content of a burst, without reference to the meaning of those tokens. This category provides insight into a user's raw ability to produce text and might broadly be an indication of the user's text production skill. Under Word Creation, we considered four sub-properties that resulted in 11 features. *Typing Speed:* 2 features–(i) the ratio of all key presses to the duration of the burst and (ii) the ratio of alphanumeric key presses to the duration of the burst. *Word Count:* 2 features–(i) the number of identifiable English words within the burst and (ii) the number of identifiable unique words within the burst. *Word Density:* 2 features–(i) the measure of the content per grammatical and lexical unit as defined by the ratio of the number of word tokens to the number of all tokens and (ii) the ratio of the content per grammatical and lexical unit to the duration of the burst. *Character Type:* 5 features–(i) ratio of the number of recognizable letter characters to total number of characters in the burst, (ii) ratio of the number of recognizable number characters to total number of characters in the burst, (iii) ratio of the number of recognizable uppercase letter characters to total number of characters in the burst, (iv) ratio of the number of recognizable space bar key presses to total number of characters in the burst, and (v) ratio of the number of recognizable vowel characters to total number of characters in the burst.

3.2.2 Lexical Complexity

Lexical Complexity provides multiple markers of a user's vocabulary sophistication, strength of text composition, language fluency, and understanding of grammatical constructs. Strong text composition emphasizes verbs over other parts of speech, while word sophistication measures character associations which are present in more complex words. Grammatical constructs identify familiarity with word contraction and possession. This category consists of four properties which results in 17 features:

Parts of Speech: The use of certain parts of speech emphasizes the sophistication of a user's composition ability. Here, we define 5 features: (i) ratio of the number of recognizable verbs in a burst to the duration of the burst, (ii) ratio of the number of recognizable nouns in a burst to the duration of the burst, (iii) ratio of the number of recognizable modifiers in a burst to the duration of the burst, (iv) ratio of the number of recognizable modals in a burst to the duration of the burst, and (v) ratio of the number of unique tokens in a burst to the duration of the burst.

Character Properties: A user's fluency with the English language can be measured by the complexity of her word construction. Word construction can be inferred by the user's use of certain letters and punctuation. We identify four properties and define 10 features – *Common Consonant:* The measure of a user's vocabulary sophistication through the use of common consonants which imply ordinary words. 3 features–(i) the ratio of common consonants to the total number of characters in the burst, (ii) the mean digraph latency between pressing any key and then pressing a common consonant key, and (iii) the ratio of common consonants timings to the total number of consonant timings in the burst. *Rare Consonant:* The measure of a user's vocabulary sophistication through the use of rare consonants which imply complex words. 3 features–(iv) the ratio of rare consonants to the total number of characters in the burst, (v) the mean digraph latency between pressing any key and then pressing a rare consonant key, and (vi) the ratio of rare consonants timings to the total number of consonant timings in the burst. *General Character:* The measure of a user's uncommon word production or sentence construction through the use of uncommon characters or punctuation. 4

features–(vii) the number of colons, semicolons, question marks, exclamation points, and quotation mark characters in the burst, (viii) the ratio of uppercase characters to all characters in the burst, (ix) the number of rare consonants in the burst, and (x) the ratio of common consonant timings to to rare common consonant timings in the burst.

Grammar: The measure of the user's familiarity with the structural rules of the English language. While grammatical understanding is abstract, there are syntax and phonetic constructs which can be applied to the text, within the burst, to determine a user's grammatical proficiency. We define 2 features: (i) the number of apostrophe "t" and apostrophe "s" character combinations within the burst and (ii) the number of words using oo, ough, and ew letter combinations within the burst (–these are known to capture phoneme awareness and orthographic complexity [8]).

3.2.3 Revision Count

Revision count is the measure of how often and how intense a user's revision is. Motivated by the evidence presented in [9], we believe that the amount of revision in a user's text indicates both typing ability and clarity of thought in a user's text production. Under this category, we define two features: (i) the number of Backspace key presses and (ii) the number of Delete key presses.

3.2.4 Keyboard Proficiency

Keyboard Proficiency measures a user's typing proficiency, while using the QWERTY keyboard. This category characterizes the user's speed and dexterity of hand/finger movement in a burst as well as the use of nonstandard keys. Keyboard Proficiency consists of three properties:

Function Key: measures a user's use of function keys or keyboard macros. This property is an indicator of the user's familiarity with typed-document construction or text processing software. Under this property we have 1 feature –the number of CTRL, ALT, ESC, and MENU keys.

Key Transition: measures a user's transition from one character type to another. Under this property, we have 49 features: *Revision*–Transition from the Backspace key to another Backspace, consonant, vowel, function, space bar, punctuation, or numeric key. *Alphabet*–Transition from any alphabet key to another alphabet, function, space bar, punctuation, or numeric key. *Numeric*–Transition from any numeric key to another numeric, consonant, vowel, function, space bar, or punctuation key. *Punctuation*–Transition from any punctuation key to another punctuation, consonant, vowel, function, space bar or numeric key. *Function*–Transition from any function key to another function, consonant, vowel, space bar or numeric key. *Space*–Transition from the space bar to punctuation, consonant, vowel, function, or numeric key.

Hand-Finger Speed: measures the time required for the user to type a key on the same row as the previously typed key or to type a key which is on a different row than the previously typed key. The feature measures latency between digraphs that are associated with a particular hand-finger combination. This property can be considered a measure of a user's typing skill because a lower latency between key and or key row transition times can be associated with correct QWERTY-keyboard typing procedure. Higher latencies imply poor hand positioning and low typing skill. The Hand-Finger Speed property groups all of its 15 features into three sub-properties of right hand, left hand, and normalized (both hands). These 15 features account for all five hand-finger (hand-thumb) combinations.

Hand-Finger Dexterity: measures the key press dwell times associated with a particular hand-finger combination. Dexterity, when applied to typing, is considered a fine motor skill and represents typing experience. Long dwell times, for a particular key associated with a specific hand-finger combination indicate lower "typing" dexterity and thus lower typing ability or keyboard familiarity. The Hand-Finger Dexterity property groups all of its 12 features into the three sub-properties of right hand, left hand, and normalized (both hands). These 12 features account for all four hand-finger combinations (thumbs are not considered).

4. Experiments

4.1. Data Collection

We performed experiments on free-text (i.e., composed text) typing data collected from 486 volunteer subjects at a university. Majority of subjects were students, but some faculty and staff also participated. The mean subject age was 21.65 years (4.37 standard deviation). 284 were male and 195 were female (gender information was not available for the rest). Data was collected on a desktop equipped with a QWERTY keyboard. Each subject participated in two sessions. We ensured that the two sessions did not occur on the same day. During each session, a participant composed and typed answers to 10 to 13 questions. Each question had an associated cognitive difficulty level and belonged to one of the six cognitive levels defined in [10]. The participant had to type an answer of at least 300 characters to each question. So answering all questions produced at least 3000 characters of text per session. Subjects answered different set of questions in each of the two sessions. While typing the answers, subjects were allowed to revise text by using the Delete/Backspace keys. Revision was optional. A participant's session ended after he/she typed answers to all questions. A keystroke sensor with 15.625ms clock resolution recorded and time stamped each keystroke event.

4.2. Experiment Design

Templates: We used the first session of typing data to build users' template. We experimented with 5 types of templates, containing: (i) P-P burst features, (ii) P-PR burst features, (iii) PR-P burst features, (iv) PR-PR burst features, and (v) "Combined Revision" burst features. We used templates containing 98 key hold and 676 digraph latency features [11] for performance comparison and fusion.

Authentication: We used second session of a user's data to create authentication vectors. We created an authentication vector by extracting features from typing data falling in a time-window of t seconds. We computed the match score using scaled Manhattan verifier, which was one of the top performers in a benchmark study on keystroke authentication [11]. We generated two types of scores, genuine (authentication vector matched against template of the same user) and zero-effort impostor (authentication vector of one user matched against the template of another). We used equal error rate to measure authentication performance.

Parameters: We experimented with two parameters: (i) t (we considered seven values: 30, 60, 90, 120, 150, 180, and 210 seconds). Note that t is an important parameter because it directly contributes to authentication delay; and (ii) We considered three values: 1, 1.5, and 2 seconds for pause duration (p). We demonstrate the impact of these parameters on availability and authentication performance.

Other settings: To detect a revision burst, we set the revision distance to 1 and depth to 3. These values were based on trial and error. For computing the template and authentication vector, we needed each feature to occur at least 3 times. When we used key hold and digraph features, we considered feature values greater than 500ms as outliers.

5. Results

5.1. Availability

We measure availability of bursts using two metrics: (1) *Percentage of authenticatable users*; and (2) *Percentage of available authentication vectors relative to key hold vectors*. Details follow.

Percentage of authenticatable users–measures the proportion of users for whom we were able to generate at least 1 genuine match. This measures how many users possessed the biometric. Some users were not authenticatable because either they did not have a *valid*[3] authentication vector or valid authentication vector was present but corresponding features were not present in the template. (In free-text authentication, there is no restriction on what a user can type, which leads to the possibility that features in template are not in authentication vector and vice-versa.)

[3]We considered an authentication vector *valid* if it had at least one feature with three or more instances in the template.

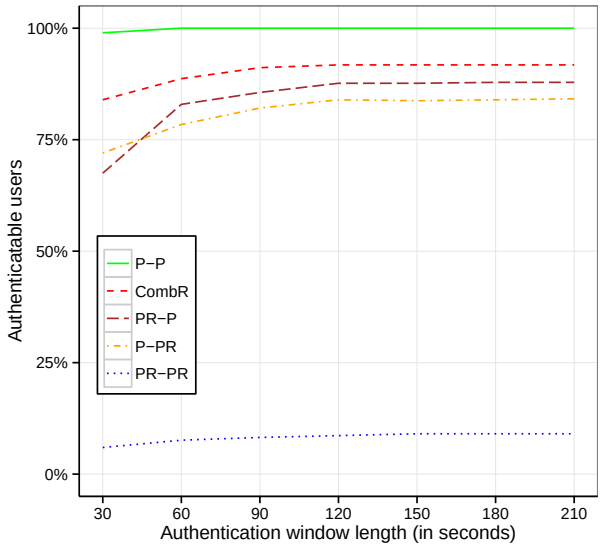

Figure 2. Percentage of authenticatable users for five types of bursts. The x-axis shows the length of authentication window (t) in seconds and y-axis shows the percentage of authenticatable users for different types of bursts.

From Figure 2, we make the following observations: P-P bursts have the highest percentage of authenticatable users (between 98.7 and 100), even when the authentication window size is low (30 and 60 seconds). The PR-PR bursts have the lowest percentage (between 5.97 and 9.05). Percentage of authenticatable users improved when all the revision bursts were grouped into "Combined Revision" burst. Percentage of authenticatable users for Combine Revision bursts still remained lower than P-P bursts.

Percentage of burst authentication vectors relative to key hold vectors–For each user, we calculated this measure as follows:

$$\frac{\text{no. of valid burst authentication vectors available}}{\text{no. of valid key hold vectors available}} \quad (1)$$

This measure compares how many valid burst authentication vectors were obtained from an authentication sample containing k keystrokes to the number of key hold vectors obtained when the same k keystrokes were used. In figures 3, 4, and 5, we report the average of this measure (computed across the authenticatable users).

We present the availability of P-P burst features relative to key hold vectors in Figure 3. We achieved the highest relative availability with pause duration $p \geq 1$ second.

In Figure 4, we show revision bursts with $p \geq 1$ second. With $p \geq 1.5$ and $p \geq 2$ seconds, the availability dropped further (figures not shown). From Figure 4, we note that the availability of bursts containing revision (i.e., P-PR, PR-P, and PR-PR) is considerably less than that of P-P bursts. Combining revision bursts (see Figure 5) im-

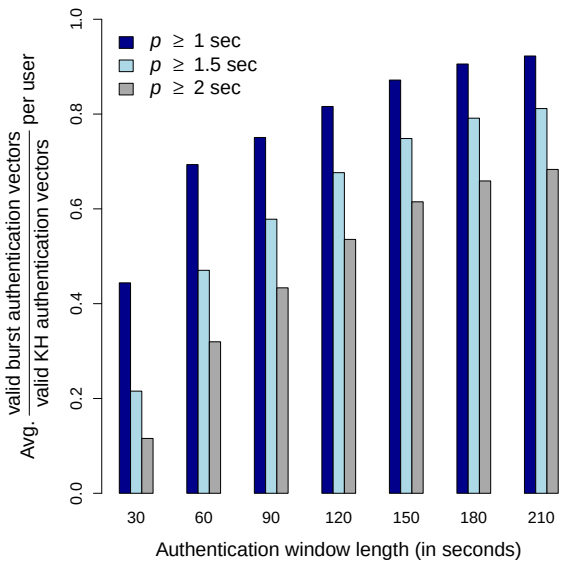

Figure 3. Availability of P-P burst vectors relative to key hold vectors for pause durations of at least 1, 1.5, and 2 seconds.

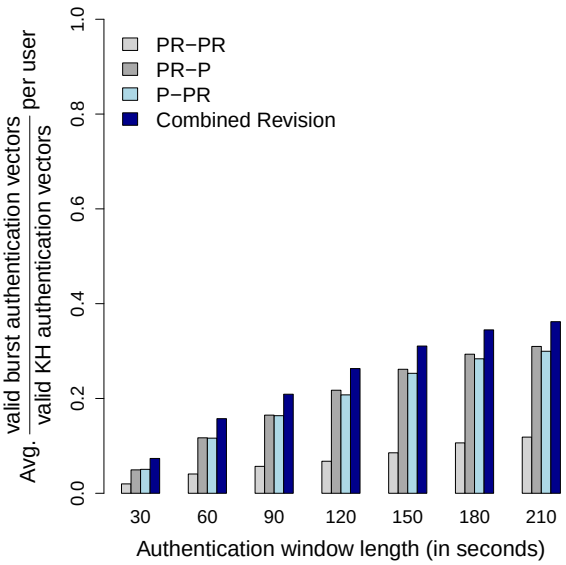

Figure 4. Availability of individual revision burst vectors relative to key hold vectors for pause durations of at least 1 second.

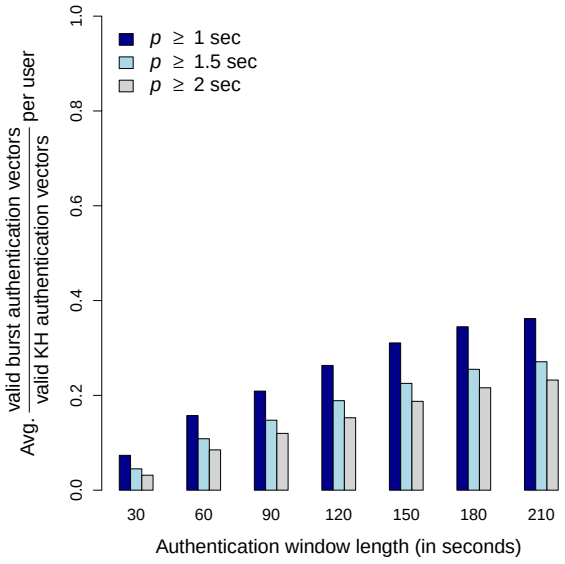

Figure 5. Availability of "Combined Revision" burst vectors relative to key hold for pause durations of at least 1, 1.5, and 2 seconds.

proved availability over individual revision bursts, but was still considerably lower than P-P bursts. Because of the low availability, in the following sections, we omit analysis on individual revision bursts.

5.2. Authentication Results with Burst Features

Before performing authentication with P-P and "Combine Revision" burst features, we performed feature selection using Fisher score (ratio of *between-user* to *within-user* distances) of each feature. The higher the Fisher score, higher the discriminability of a feature. We used training data (Session 1) to compute Fisher scores. Using Fisher scores of individual features, we created a subset of features using the following steps: (1) Compute $total$ by adding the Fisher scores of all features, (2) Sort features in decreasing order of Fisher scores; and (3) Select top k features so that the scores of the k features added accounts for x percentage of $total$. We tried subsets created by different x values (85, 88, 90, 95, and 98). For P-P burst features, $x = 95\%$ gave the best EERs for all three pause durations (i.e., 1, 1.5, and 2 seconds). For Combined Revision bursts, we achieved best results with $x = 90\%$ for $p \geq 1$ and $x = 98\%$ with p greater than 1.5 and 2 seconds. However, in each case the features selected for different p values were different.

From Figure 6, we observe the following: (a) P-P burst based features yield lower EERs than Combined Revision burst based features and (b) for both P-P and Combined Revision, pause of 1 second yielded the lowest EERs. The best performing burst feature subset, i.e., P-P burst features ($p \geq 1$) with $x = 95\%$, had 72 features comprising of 5 Word Creation, 8 Lexical Complexity, 1 Revising Style, 54 Keyboard Proficiency, and 4 Pre- and Post-pause features.

Hereafter, we report the authentication results of burst features with $p \geq 1$ because it has: (1) the highest percentage of authenticatable users (Figure 2), (2) the highest availability relative to key hold vectors (figures 3 and 5) and (3) the best authentication performance (Figure 6).

Comparison with Key Hold and Digraph. After performing feature subset selection, we used 98 key hold and

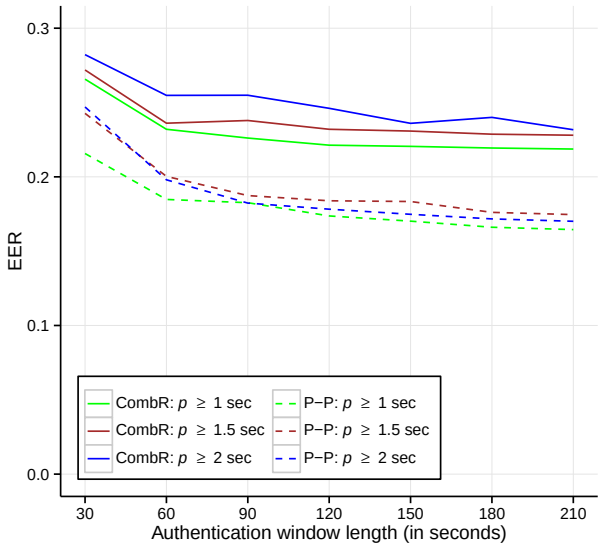

Figure 6. EERs of P-P and Combined Revision bursts. P-P burst performs better than Combined Revision bursts (top three EER plots are P-P, bottom three are Combined Revision).

676 digraph features. In Figure 7, we compare EERs of features extracted from bursts, with EERs of digraph and key hold features. Key hold outperforms both digraph and burst features. Digraph features perform worse than burst features for short authentication windows (worse than Combined Revision burst for 30-second authentication window and worse than P-P burst features for 30-, 60- and comparable for 90-second windows). However, for longer windows, digraphs perform better than burst features.

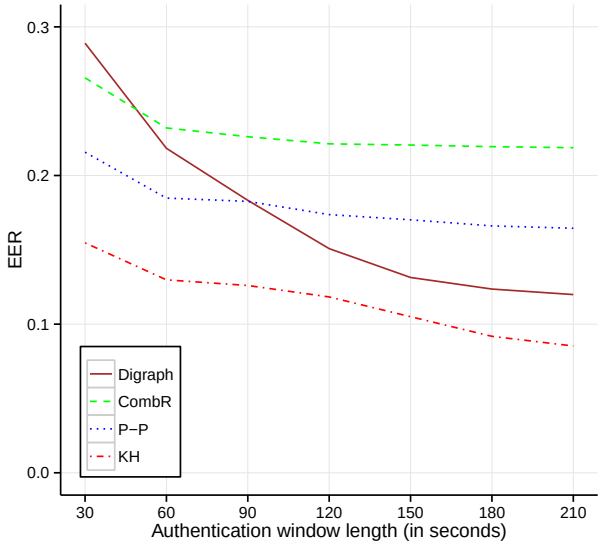

Figure 7. Comparison of EERs achieved with key hold, digraph, P-P burst features, and Combined Revision burst features.

5.3. Fusion Performance

We performed weighted sum score-level fusion of key hold, digraph, and burst verifiers. We choose weighted sum fusion because it is simple and has been demonstrated to perform well in biometric fusion [12].

In free text continuous authentication, all the features may not be available all the time. This leads to situations where scores form some verifiers are missing during fusion. We address this problem using a two step score-level fusion procedure (illustrated below with an example).

1. Let w_1, w_2, w_3, and w_4 ($w_1+w_2+w_3+w_4 = 1$) be the weights of four scores s_1, s_2, s_3, and s_4, output by four verifiers v_1, v_2, v_3, and v_4 respectively. If all scores are available, we fuse using: $w_1s_1+w_2s_2+w_3s_3+w_4s_4$.
2. If the score from one verifier, say v_2, is missing, we recalculate the weights as follows: $w_1 = w_1/(w_1 + w_3 + w_4)$, $w_3 = w_3/(w_1 + w_3 + w_4)$, and $w_4 = w_4/(w_1 + w_3 + w_4)$. Similarly, if scores from two verifiers, say v_2 and v_3 are missing, we recalculate the remaining scores as follows: $w_1 = w_1/(w_1 + w_4)$ and $w_4 = w_4/(w_1 + w_4)$.

The above procedure ensures weights sum to 1 even when some scores are missing and the proportion of weights is preserved after readjustment. We performed a grid search of weights and report results of the best combination in Figure 8. Our observations on fusion performance are: (1) Fusion of P-P burst scores with key hold improves performance compared to using only key hold features; (2) Fusion of digraph with key hold scores performs better than the fusion of P-P with key hold; (3) Fusing key hold, digraph, and P-P burst features gives the best EERs; and (4) Including Combined Revision in fusion with key hold, digraph, and P-P burst did not further improve EERs

6. Related Research

Due to space constraints, we discuss recent continuous authentication studies that have used bursts and combined keystroke dynamics with stylometry. For an overview of continuous authentication using keystroke dynamics and stylometry, refer to [2, 13] (and references therein).

In [14], the authors extracted keystroke biometric from short (1- to 3-minute) bursts. While [14], like ours, extracts features from bursts, the purpose and usage of bursts is completely different. [14] uses bursts to reduce the frequency of authentication checks, to potentially reduce false accept rate. On the other hand, we use different types of bursts to provide a "cognitive" activity contexts to our features.

The authors of [3] captured keystroke and linguistic features from typing data collected when students were taking an online test in a university-level course. The feature set included a total of 239 keystroke-based features (key press duration and key transition times) and 228 linguistic fea-

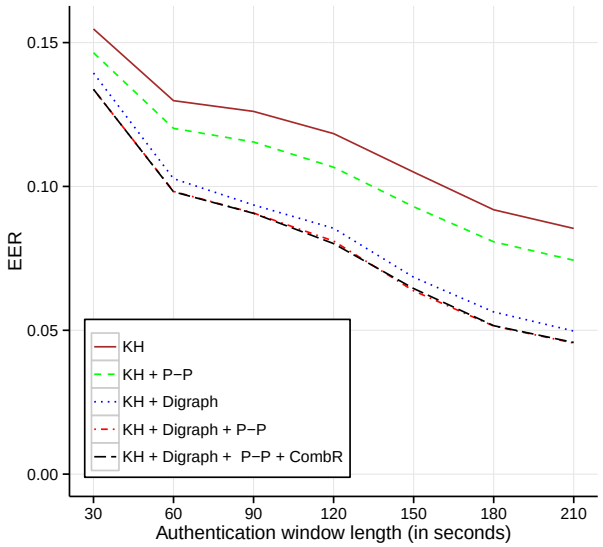

Figure 8. EERs of fusion for the best combination of weights. Fusing P-P with key hold and digraph gave the lowest EERs. Fusing Combined Revision bursts did not lower EERs any further.

tures: 49 character features, 13 word features, and 166 syntax features. The authors reported a classification performance of 99.96% and 100.00% with keystroke system, and 74% and 78% with stylometry system , on test input of 500 and 1000 words, respectively. The work in [4] is similar to [3] but used a smaller set of features. The conclusion reached in both [3] and [4] was that keystroke based authentication was far superior to that of the stylometry. In contrast to [3, 4], we perform context-aware feature extraction by capturing 123 feature measurements from different types of bursts rather than from arbitrary keystroke streams. These features have resulted in improvements in authentication accuracies (see Figure 8). Thus our work demonstrates that harnessing features that capture the cognitive aspects of text generation process is a promising new direction in continuous behavioral authentication.

7. Conclusions and Future Work

We presented the continuous authentication performance with 123 behavioral features, extracted from text production (i.e., P-P) and revision (i.e., P-PR, PR-P, PR-PR, and Combined Revision) bursts. Our results on free text data from 486 users showed that P-P burst features have higher availability and lower authentication error rate compared to revision burst features. In addition, fusion of P-P burst features with keystroke dynamics features improved authentication error rate, even for low authentication window sizes (which is desirable for early detection of impostors). Our future work will involve two important directions: (1) quantifying the impact of the cognitive load of the text typed by the user on continuous authentication performance with burst features and (2) investigating the resilience offered by burst features against non-zero effort attacks such as those in [15].

References

[1] J. Roth, X. Liu, A. Ross, and D. Metaxas, "Biometric authentication via keystroke sound," in *2013 Intl. Conf. on Biometrics,*, June 2013, pp. 1–8.

[2] A. Ahmed and I. Traore, "Biometric recognition based on free-text keystroke dynamics," *IEEE Trans. on Cybernetics*, vol. 44, no. 4, pp. 458–472, April 2014.

[3] J. Monaco, J. Stewart, S.-H. Cha, and C. Tappert, "Behavioral biometric verification of student identity in online course assessment and authentication of authors in literary works," in *2013 IEEE Sixth Intl. Conf. on Biometrics: Theory, Applications and Systems*, Sept 2013, pp. 1–8.

[4] J. Stewart, J. Monaco, S.-H. Cha, and C. Tappert, "An investigation of keystroke and stylometry traits for authenticating online test takers," in *2011 Intl. Joint Conf. on Biometrics*, Oct 2011, pp. 1–7.

[5] N. A. Chenoweth and J. R. Hayes, "Fluency in writing: Generating text in l1 and l2," *Written Communication*, vol. 18, no. 1, pp. 80–98, 2001.

[6] ——, "The inner voice in writing," *Written Communication*, vol. 20, no. 1, pp. 99–118, 2003.

[7] V. M. Baaijen, D. Galbraith, and K. de Glopper, "Keystroke analysis: Reflections on procedures and measures," *Written Communication*, vol. 29, no. 3, pp. 246–277, 2012.

[8] M. S. Seidenberg, G. S. Waters, M. A. Barnes, and M. K. Tanenhaus, "When Does Irregular Spelling or Pronunciation Influence Word Recognition?" *J. of Verbal Learning and Verbal Behaviour*, vol. 23, pp. 283–404, 1984.

[9] M. Schwartz, "Revision profiles: Patterns and implications," *College English*, vol. 45, no. 6, pp. pp. 549–558, 1983.

[10] D. R. Krathwohl, "A revision of bloom's taxonomy: An overview," *Theory Into Practice*, vol. 41, pp. 212–218, 2002.

[11] S. Killourhy and R. A. Maxion, "Comparing anomaly-detection algorithms for keystroke dynamics," in *IEEE/IFIP Intl. Conf. on Dependable Systems and Networks*, Sept. 2009, pp. 125–134.

[12] A. Jain and A. Ross, "Information fusion in biometrics," *Pattern Recognition Letters*, vol. 24, pp. 2115–2125, 2003.

[13] A. Fridman, A. Stolerman, S. Acharya, P. Brennan, P. Juola, R. Greenstadt, and M. Kam, "Decision fusion for multimodal active authentication," *ITPro*, vol. 15, pp. 29–33, Jul. 2013.

[14] J. Monaco, N. Bakelman, S.-H. Cha, and C. Tappert, "Developing a keystroke biometric system for continual authentication of computer users," in *European Intelligence and Security Informatics Conference*, Aug 2012, pp. 210–216.

[15] K. Rahman, K. Balagani, and V. Phoha, "Snoop-forge-replay attacks on continuous verification with keystrokes," *IEEE Trans. on Info. Forensics and Security*, vol. 8, no. 3, pp. 528–541, March 2013.