A Correlation Method for Handling Infrequent Data in Keystroke Biometric Systems

Steve Kim, Sung-Hyuk Cha, John V. Monaco, and Charles C. Tappert

Department of Computer Science, Pace University
1 Pace Plaza, New York, NY 10038, USA
skim@gmail.com, {scha, john.v.monaco, tappert}@pace.edu
Linguistic Hierarchy Model
Touch-type Hierarchy Model
Keystroke feature acquisition

Keystroke duration

\[ S_x = \{ r_i - p_i \mid (r_i, p_i, k_i) \in A \land k_i = x \} \]
Extracting sufficient co-exist table with $t_1 = 7$. 

<table>
<thead>
<tr>
<th>$i$</th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>12.5 (14)</td>
<td>14.2 (12)</td>
<td>17.3 (5)</td>
</tr>
<tr>
<td>2</td>
<td>11.4 (21)</td>
<td>13.6 (10)</td>
<td>18.1 (12)</td>
</tr>
<tr>
<td>3</td>
<td>17.1 (7)</td>
<td>12.9 (4)</td>
<td>12.2 (17)</td>
</tr>
<tr>
<td>4</td>
<td>10.2 (10)</td>
<td>16.8 (3)</td>
<td>11.6 (11)</td>
</tr>
<tr>
<td>5</td>
<td>15.1 (5)</td>
<td>18.2 (12)</td>
<td>- (0)</td>
</tr>
<tr>
<td>6</td>
<td>12.4 (12)</td>
<td>11.7 (9)</td>
<td>14.8 (4)</td>
</tr>
<tr>
<td>7</td>
<td>- (0)</td>
<td>12.3 (11)</td>
<td>17.1 (21)</td>
</tr>
<tr>
<td>8</td>
<td>10.2 (18)</td>
<td>15.9 (7)</td>
<td>19.2 (11)</td>
</tr>
<tr>
<td>9</td>
<td>16.2 (11)</td>
<td>11.3 (14)</td>
<td>12.8 (4)</td>
</tr>
<tr>
<td>10</td>
<td>19.4 (4)</td>
<td>18.1 (17)</td>
<td>17.6 (12)</td>
</tr>
</tbody>
</table>

Keystroke sample

Mean key duration in each sample (frequency)

Coexistence between key durations

$R$

$D_{AB}$

$D_{AC}$

$D_{BC}$
Good correlation
(0.908) b/w D and E

Bad correlation
(0.561) b/w A and I
Correlations among vowels
Max Correlation vs. Key Frequency.
Shifting features in sample with infrequent ‘E’ keys

Duration $q$ from key ‘D’ is translated to $q_c$ with the linear regression line previously found. This is a good estimate of ‘E’

Using the existing fallback models gives $q_d$, a poor estimate of ‘E’

2-Feature space of keys D and E
Shifting from $q (3.98, 4.16)$ to $q_c (2.47, 5.13)$

$\begin{align*}
q'_e &= 0.77q_d + 3.36 \\
q'_d &= q_e - 2.82
\end{align*}$
Flow chart of proposed correlation-based fallback table model
Hierarchical fallback

\[ \text{fallb}(x) = \begin{cases} 
\{ r_i - p_i \mid k_i \in \text{leaf}(x) \} & \text{if } |S_x| > t \\
\text{fallb}(\text{parent}(x)) & \text{otherwise}
\end{cases} \]

Linear regression fallback

\[
S_{x,l} = \{ \alpha_{x,l} (r_i - p_i) + \beta_{x,l} \mid (r_i, p_i, k_i) \in A \land k_i = k_{x,l} \}
\]

\[
cft(S_x, l) = \begin{cases} 
S_x & \text{if } |S_x| > t \\
cft(S_x \cup S_{x,l}, l+1) & \text{otherwise}
\end{cases}
\]
Experimental Results

EER of each fallback model as a function of $|S_x|$
## Fallback Model EER Table.

| $|S_x|$ | Default | Linguistic | Physiologic | Regression |
|-----|---------|------------|-------------|------------|
| 50  | 22.88   | 22.60      | 21.80       | 20.47      |
| 100 | 17.34   | 18.06      | 16.37       | 17.84      |
| 200 | 11.80   | 12.36      | 11.00       | 11.34      |
| 300 | 9.74    | 9.51       | 8.60        | 8.50       |
| 400 | 7.52    | 6.93       | 7.56        | 6.52       |
| 500 | 6.80    | 6.62       | 6.54        | 6.15       |
| Max | 4.54    | 4.86       | 4.70        | 4.31       |
Conclusions

• Modest improvements over existing methods
• No expert is needed to construct the fallback models
• More importantly: a language independent method of dealing with infrequent keystroke data