# Analysis of the Acquisition Process for Keystroke Dynamics

***R. Giot*** A. Ninassi M. El-abed C. Rosenberger

GREYC - Ensicaen, Université de Caen, CNRS UMR6072

09/07/2012

# What are the objectives of this talk?

We are interested in the ***acquisition*** procedure for ***keystroke dynamics***

## Keystroke dynamics performance problem

- Recognition performance depends on the selected dataset
- The reasons can be various
  - Password(s) different
  - Individuals different
  - Acquisition procedure different
    - ***The way to give the password to individuals is different***

## Our contribution

Analysis of the typing factors which affect the performance

# Outline

# Keystroke dynamics is a behavioral modality I

## Benefits of keystroke dynamics

+ ***Low cost*** behavioral biometric
+ ***Not*** an ***intrusive*** biometric

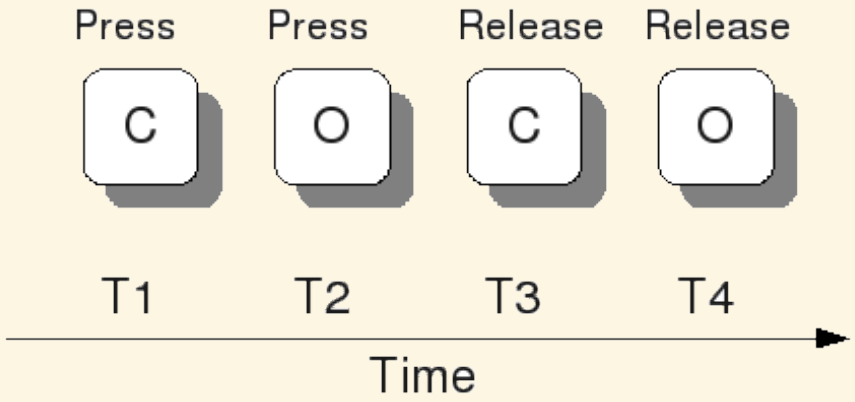
## Drawbacks of keystroke dynamics

- Lots of samples required during the enrollment
- Lots of factors impact the stability
- Subject to template ageing

# Keystroke dynamics is a behavioral modality II

## How does it work?

- Timestamps of keyboard events are tracked
- Various ***durations*** and ***latencies*** are computed
- Authentication verifies if the query vector looks like the enrollment vectors




T2-T1/T4-T3 Duration

T3-T1 Latency

T4-T2 Latency

T3-T1 Latency

# There are some public keystroke dynamics datasets I

## Properties of the public datasets

- Differences in number of users
- Differences in number of samples

## Remark on the acquisition procedure

There is ***no*** explanation of the way the text to type is given to the volunteer

# There are some public keystroke dynamics datasets II

## Problem about the acquisition procedure

- We can suspect that performance depends also on the acquisition procedure and the way text is presented
- Why?
  - On tactile smart-phones, we encountered bad performances.
  - We suspect it is because of the way we presented the text to type

# Main points of the protocol

## Objectives

1. ***Acquisition*** of a dataset which follows ***different acquisition*** scenarios
2. ***Comparison*** of the different scenarios, and their impact on performances

## Two advances are expected

1. Knowledge to create less noisy datasets under more realistic conditions
2. Selection of the most appropriate acquisition method to maximize the performance of keystroke based authentication systems

# Different kinds of passwords can lead different typing difficulty I

## Numbers based passwords

- ***Structured numbers***
  - Supposed to be commonly known or easy to remember
  - phone number
  - credit card number
  - ...
- ***Unstructured numbers***
  - Supposed to be difficult to memorise
  - Are not related to common numbers

# Different kinds of passwords can lead different typing difficulty II

## Text based passwords

- ***Known words***
  - Supposed to be easy to remember
  - The volunteers know the word
- ***Unknown words***
  - Supposed to be hard to remember
  - The volunteers do not know the word
  - Random word
  - Word with non alphabetic characters

# Different ways to ask to type a password can also lead to different typing difficulty

## Common ways of presenting the passwords

- Displayed in the ***GUI*** (can be read when typing)
- Displayed in the GUI with a ***different graphical presentation***
- Displayed in a ***dialog box*** (requires memorization)

## Specific ways for numerical data

- Displayed by ***groups*** of L digits
- ***Read*** in totality
- ***Read*** by ***groups*** of L digits
- ***Read digit*** by ***digit***

# Several constraints are needed to collect the dataset

- We need ***several passwords*** per type of password
- We need ***several sessions*** to verify reproducibility
- A session is composed of a list of events
- An Event is a pair of
  - Information to type
  - Way of presenting this information
- Each session must be different to limit habituation factor

# This gives a huge amount of combinations to test

## Remark

- Cartesian product between the type of passwords, the selected password, and the presentation schemes is very large
- By the way, we have not listed all the possibilities

## Consequence

- It would need too much time to test all the possibilities

$\Rightarrow$ That's why we use only a subset in our experiments

# 3 passwords per type of passwords are selected

## Numbers

- 118218
- 982491840
- 234567

## Known words

- voiture
- poisson
- appartement

## Unknown words

- vertuio
- ospsoni
- entappremat

# We have selected few kinds of presentation I

## Numbers

PF1 Displayed in the graphical user interface, all digits together.

PF2 Displayed in the graphical user interface, by packet of 3 digits.

PF3 Listened by packet of 3 digits.

## Known words

PC1 Displayed in the graphical user interface.

PC2 Listened

# We have selected few kinds of presentation

II

## Unknown words

PI1 Displayed in the graphical user interface.

PI2 Spelt

# Summary of the obtained dataset

- AZERTY keyboard
- 28 volunteers, all French
- 2 sessions
- 1 week between each session
- 10 presentations of each event
- 21 events/105 inputs per session

# Validation of the results I

## Keystroke dynamics distance function

$$1 - \frac{1}{N} \sum_{i=1}^{N} exp\left(-\frac{|q_i - \mu_i|}{\sigma_i}\right)$$

- $q$, query
- $\mu$, enrolled mean
- $\sigma$, enrolled standard deviation

## Selected error rates

- Equal Error Rate
  - Performance for one operational point
- Failure To Acquire Rate
  - Often important in keystroke dynamics
  - Can be annoying for the user

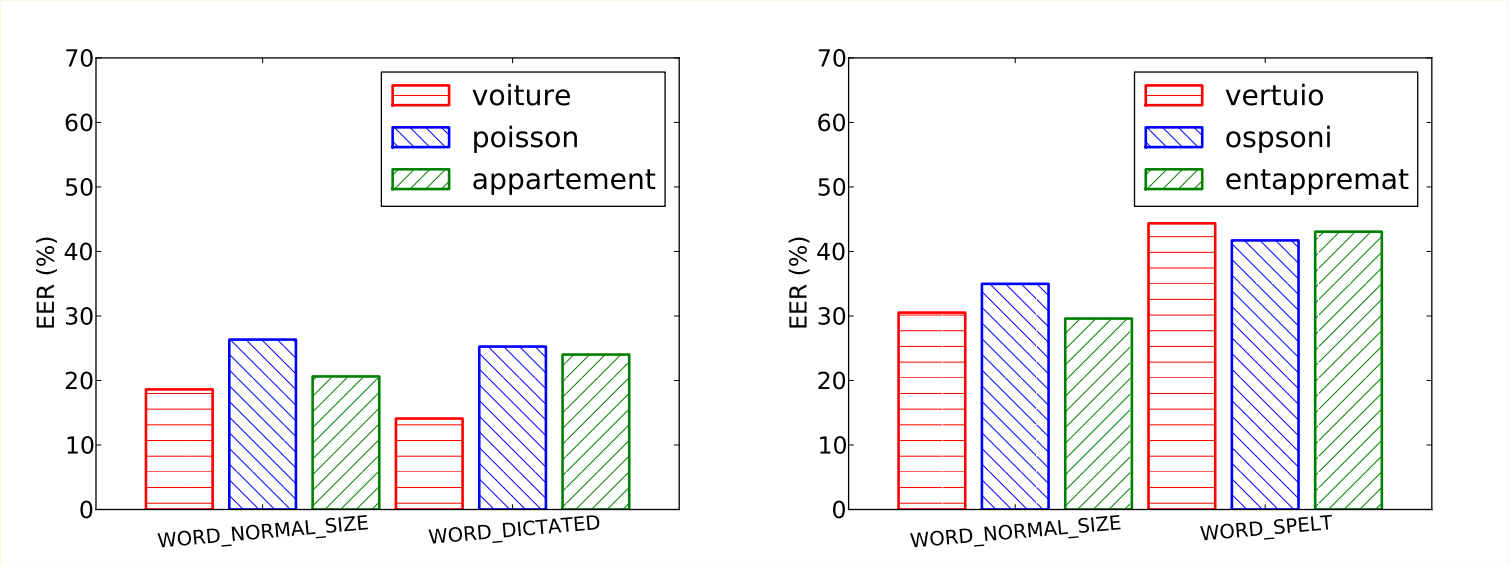
# Validation of the results II

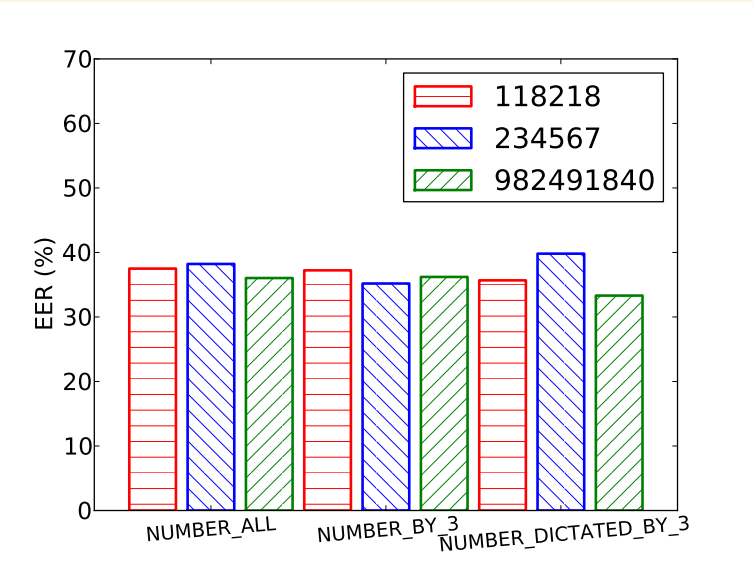
## Statistical verification

Kruskal-Wallis (KW)

- Verification if the samples (set of EER or FTAR) are from the same population
- H0 = same population
- H1 = different population

# EER analysis



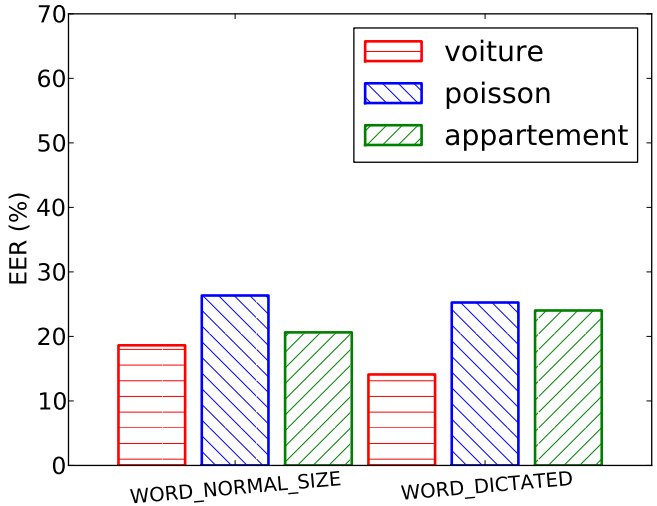
# Known words give better performance than randomized ones and numbers

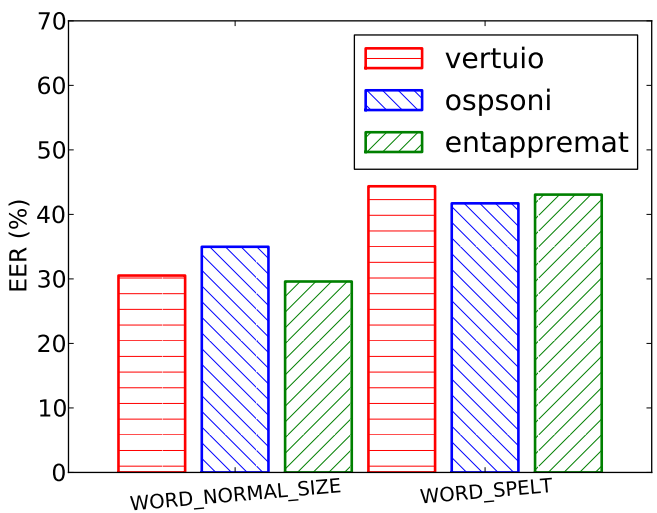
- known/unknown
  p-value=0.00394

⇒ We must use real words instead of complex ones

- known/numbers
  p-value=0.00146

⇒ Not a good idea to use keystroke dynamics with PIN codes



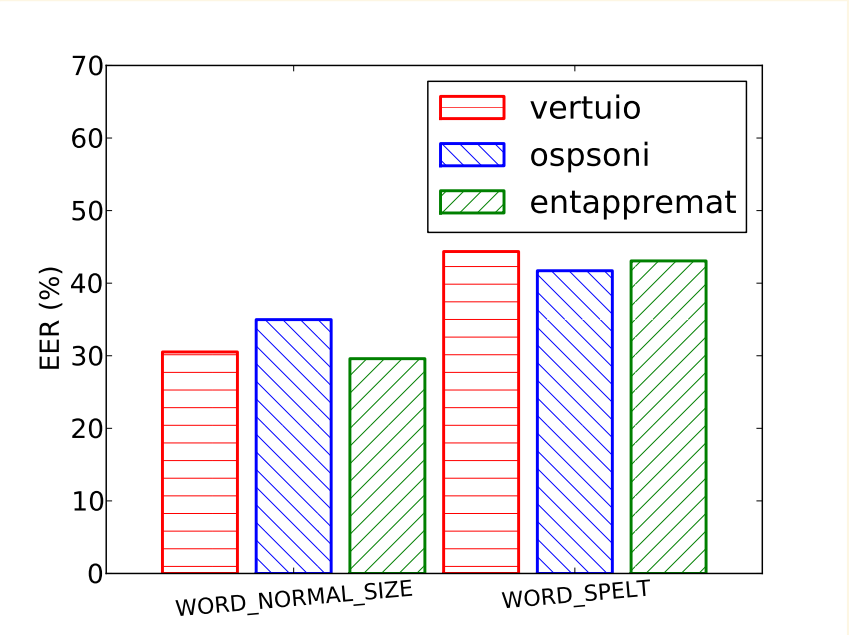



# Oral presentation (spelt) increases the EER for randomized words

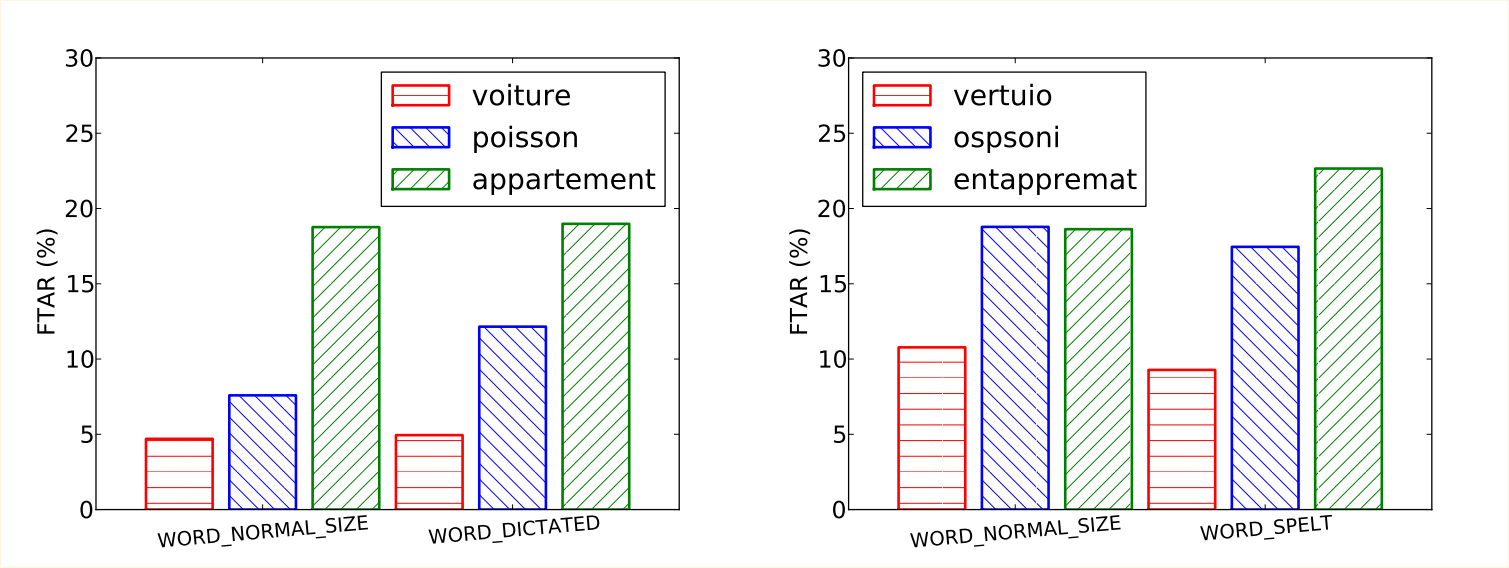
- p-value=0.049

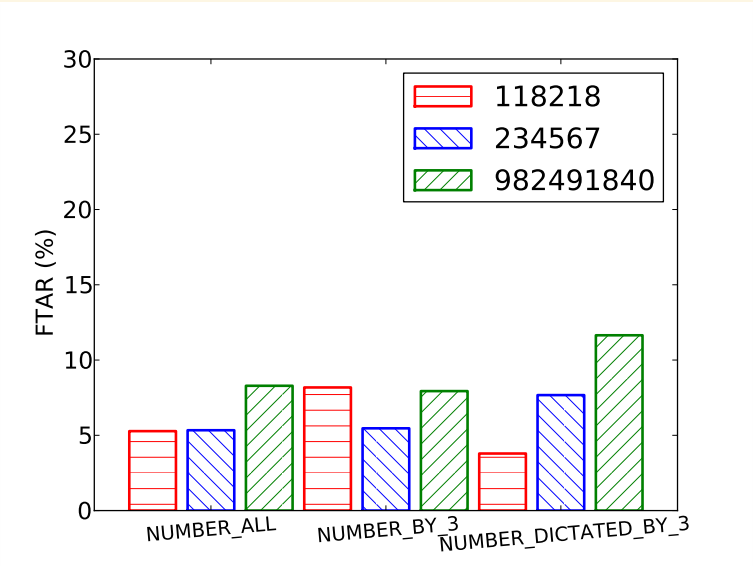
⇒ We are not use to hearing spelt messages when using a computer

- However, there is no difference for known words and numbers



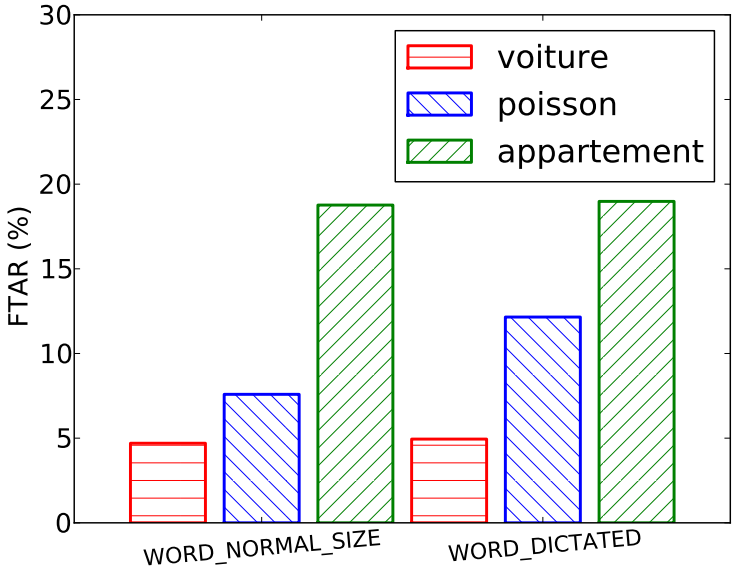
# Failure To Acquire Rate analysis

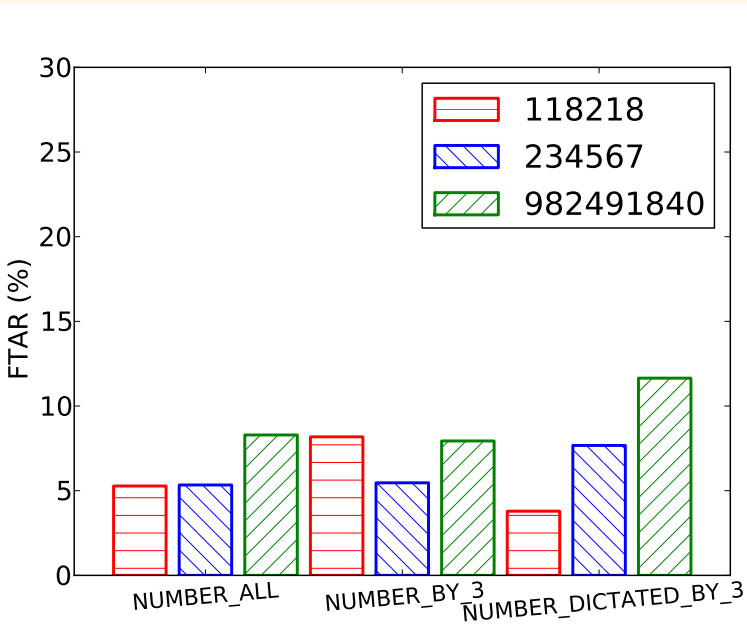


# FTAR depends more on the type of text

- FTAR is lower for numbers than texts
  - however, it is not statistically significative (p-value=0.1023)
- FTAR is lower for known words than other words
  - It can be explained by typing habits
  - however, it is not statistically significant (p-value=0.26)
- Oral presentation has no influence on the FTAR



# Various remarks about the performances

## Average EER=32%

- Not enough samples for the enrollment
- Big performance difference between the best and worst performing event

## Average FTAR=11%

- Similar than other datasets in the lab
- Keystroke dynamics is probably the biometric modality having the highest FTAR

## There is no correlation between EER and FTAR

Pearson correlation coefficient of 0.005

# Conclusion I

## We have analysed the performance of keystroke dynamics

- by varying the kind of text to type
- by varying the way to present the text
- by varying the value of the text to type

## We have shown that

- It is better to choose passwords
  - short
  - simple, known word
  - not numbers
- During the acquisition procedure of the database, there is no preference between written or vocal presentation

# Conclusion II

## Results are interesting for

- People planning to ***acquire*** new datasets
- People planning to write ***rules*** to use keystroke dynamics systems
- People planning to produce systems with ***spontaneous passwords***

## Limitations of the study

- Database too short in numbers of users
- Database too small in numbers of different events

$\Rightarrow$ Slightly different results are expected with a bigger database

$\Rightarrow$ a bigger dataset is mandatory to explore deeply

Thanks for your attention

http://www.epaymentbiometrics.ensicaen.fr/