Multimed Tools Appl
DOI 10.1007/s11042-017-4571-7

Improving static audio keystroke analysis by score fusion of acoustic and timing data

Matúš Pleva[1] · Patrick Bours[2] · Stanislav Ondáš[1] · Jozef Juhár[1]

Received: 1 August 2016 / Revised: 18 January 2017 / Accepted: 3 March 2017

Abstract In this paper we investigate the capacity of sound & timing information during typing of a password for the user identification and authentication task. The novelty of this paper lies in the comparison of performance between improved timing-based and audio-based keystroke dynamics analysis and the fusion for the keystroke authentication. We collected data of 50 people typing the same given password 100 times, divided into 4 sessions of 25 typings and tested how well the system could recognize the correct typist. Using fusion of timing (9.73%) and audio calibration scores (8.99%) described in the paper we achieved 4.65% EER (Equal Error Rate) for the authentication task. The results show the potential of using Audio Keystroke Dynamics information as a way to authenticate or identify users during log-on.

Keywords Biometrics · Keystroke dynamics · Timing analysis · Acoustical analysis · Authentication · Identification

✉ Matúš Pleva
matus.pleva@tuke.sk;
http://www.kemt.fei.tuke.sk

Patrick Bours
patrick.bours@ntnu.no;
http://www.ntnu.edu/iik

Stanislav Ondáš
stanislav.ondas@tuke.sk

Jozef Juhár
jozef.juhar@tuke.sk

[1] Technical University of Košice, Department of Electronics and Multimedia Communications, FEI Letná 9, 041 20 Košice, Slovakia

[2] NISlab - Norwegian Information Security Laboratory, Department of Information Security and Communication Technology, Postboks 191, 2802 Gjøvik, Norway

Published online: 29 March 2017

1 Introduction

Password systems are widespread and used everywhere on a daily basis. The main advantage is the easy usage, but that is often offset by users selecting weak passwords and hence rendering the system weak [13]. Analysis on a database of 32M leaked passwords [8] has shown that many users select passwords that can be easily broken by a dictionary attack. Biometric systems provide higher security, but generally at the cost of expensive sensors. Keystroke Dynamics (KD) is a way to combine passwords and biometrics at no additional cost, because the device necessary for this is already present at a common laptop or desktop computer. Keystroke Dynamics does not look so much at what a user types but how a user types [24]. So far, most used features in KD are timing related, as this information is the easiest to collect [2, 23]. The Operating System (OS) of a computer can provide information on when a key is pressed down (key-down time) and when it is released (key-up time). In some research also key pressure [18, 22] has been considered for KD features. On a computer this will, however, require a special keyboard, which means that the advantage of no extra cost is lost.

Relatively little research has been done using the sound produced when typing on a keyboard [7, 15, 20, 21]. In this paper we will investigate this area and see if we can identify the user by the sound of typing his/her password. The novelty of this paper lies in the fact that it will compare the performance of "classic" timing-based KD with the relatively new audio-based KD. Our comparison is both with respect to authentication as well as identification. In addition we will investigate the performance change when timing and audio information is combined. Most laptop and desktop computers have a microphone and camera already present, hence it will be relatively easy to collect audio information.

The *contributions* in this paper can be summed up as follows:

- Improvement of timing-based KD by outliers detection.
- Improvement of audio-based KD by audio scores calibration.
- Fusion of the timing-based and the calibrated audio-based KD.

The remainder of this paper is organized as follows. In Section 2 we will give an overview of the research done in Keystroke Dynamics with the focus on using audio information. We will also provide some additional ideas on how typing audio information can be used. Section 3 will describe the setup of our experiment as well as describe the data that was collected. Sections 4 and 5 will provide the description of how the analysis and fusion of our data has been performed as well as the results of this analysis with comparisons. We conclude this report with future research suggestions in Results discussion Section 6 and in Conclusions Section 7.

2 State of art

Keystroke Dynamics is a behavioral biometric modality that is used to authenticate or identify individuals based on their typing rhythm [2]. The major advantage of KD as a biometric is that it can be software based and no extra hardware is required. Most KD systems use key-up and key-down timing information from which the duration and latency features can be derived. Duration is a single key feature and represents how long a key has been kept down, i.e. the elapsed time between key-down and key-up of that particular key. Latency is a feature of 2 consecutive keys and represents the time during which the keyboard has not been used, i.e. the time between key-up of the first key and key-down of the second key. Alternative definitions of latency exist but this is not relevant for this paper. Note that latency can also be negative if the second key is pressed before the first is released. A comparative study of statistical and machine-learning approaches has been conducted on a common database to see which analysis technique performs best [10]. On a database of 51 persons and 400 samples per person, the best performing analysis method was the Scaled Manhattan Distance, with a 9.6% Equal Error Rate (EER).

In some research [18] it has been shown that including key pressure will lead to better performance results. This, however, requires specialized hardware to measure this kind of information. If KD is performed on a tablet with a touch screen, then such information is available. In our research we will not focus on pressure or touch screens. We will exhibit the fact that the sound of typing is different for different persons. Some people gently press each key separately, while others might type fast and hit keys hard. It has been noted before that the sound of typing differs per person, and, per key typed. This information can be used to recognize either the person that is typing, or what this person is typing. Early research was mostly focused on retrieving the text from the keystroke audio [1, 11, 12]. Interestingly, it is not just the sound that can be recorded, but also, due to the typing on the keyboard, the vibration of a laptop screen [3] can be captured by a laser microphone and used to reconstruct the typed text. An assumption that is often made is that the user types English text and no typing corrections are considered.

Little research has been done in using the typing audio for authentication or identification of people. To the best of our knowledge, this was first investigated in [7], where the authors combined keystroke timing information with typing sound information. In their investigation they experimented participants needing to type the password “kirakira”. Audio data was collected through a separate microphone placed at the base of the laptop keyboard. Ten users provided 10 typing samples each, and of these samples, 5 were used to train a Self Organizing Map (SOM), while the other 5 were used for testing. They evaluated the performance of the system in an authentication setting. In a follow-up publication [15] they used Supervised Pareto SOM to improve their results. The main difference between the research in [7, 15] and our research is threefold. First of all, we use a different analysis technique on the audio data, furthermore we compare performance between timing KD and audio KD and finally our dataset is much larger (50 versus 10 users and 100 versus 10 samples per user). The overlap lies in the fact that they, just as we, have used a fixed password, i.e. we considered static authentication in both cases [5].

In [21] the authors extended their initial work [20] on keystroke sound. They did not investigate the possibility for static authentication, but concentrated on continuous authentication [4] to see how the sound of typing could be used. They collected audio data of 50 persons, typing either a fixed text (i.e. the first paragraph of ”A Tale of Two Cities” by Charles Dickens, containing 613 characters, shown on screen) or a free text (i.e. they were tasked to type a half page email to family without further instructions on the content). Audio data was collected using the microphone of an inexpensive webcam mounted at the top of the laptop screen, where the microphone was directed towards the keyboard. The authors achieved an EER of approximately 11% in their experiments. They also briefly looked at an identification task where they reached a rank-1 accuracy of approximately 75%. The main differences between ours and the research from Roth et al. is that they focused on long texts, i.e. on continuous authentication, while our focus is on short passwords, i.e. static authentication.

3 Data collection

Because there does not exist any publicly available dataset for the desired research, we had to collect new data. The setup for our data collection is similar to what was done in the works of Dozono et al. [7], Nakakuni et al. [15] and Roth et al. [20, 21]. The difference with Dozono & Nakakuni [7, 15] is the size of the collected database and with Roth [20, 21] it is the amount of data per sample. The keyword used as passphrase could be retrained for the real password used by the user. The word “password” or “kirakira” (in other databases [7]) is used mainly for proof of concept.

We focus on the password scenario where all users type the same password. So the chosen word was “password” and the participants in the experiment were not allowed to see what they typed. This was enforced by moving the screen away from the participants. Only the experiment supervisor could see what a participant typed. We are well aware that “password” is weak as a password, but we have deliberately chosen a password that is easy for everybody to type to avoid major changes in typing behaviour of the participants because they had to learn a completely random complicated password.

Each of the 50 participants (10 female, 40 male, average age 26, mainly staff and students of Gjøvik University College - currently the Norwegian University of Science and Technology) had to type the password 100 times, divided into 4 session of 25 times. The experiment supervisor checked how many correct typings of the password occurred in each session and stopped the participants when the target was reached. Between sessions the participants could relax for a few minutes before they continued to the next session.

The experiment took place in a semi-controlled environment, where there was limited background noise, but noise from neighboring rooms could not be controlled, only recorded. Generally the noise level from adjacent rooms was very low during the experiment. Of course the hardware and environment sounds play important role for audio-based KD analysis but for authentication task the calibration and training is needed anyway, so all keyboards & environments used during the training should be reliable for this purpose.

One major difference between our research and other related research is that we have used a desktop keyboard instead of a laptop keyboard. Figure 1 shows the setup that was used in the experiment. We used a simple webcam microphone (Logitech model QuickCam

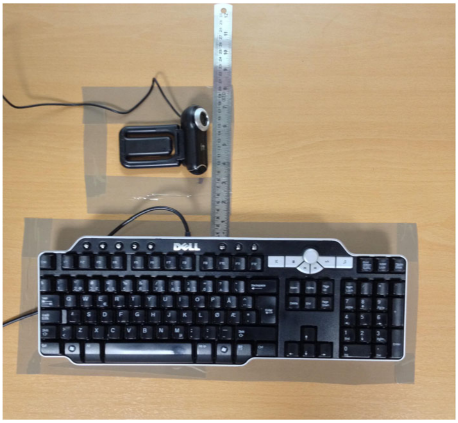


Fig. 1 Experiment setup used to collect the audio data

Pro 9000) to collect the typing sound of the desktop keyboard (DELL model SK-8135). The webcam and keyboard were placed in the area marked by tape, indicating also the location of the microphone on the webcam. The microphone was placed near the middle of the keyboard at a distance of approximately 10 cm. Minor moves of the equipment could be observed, but no large movements were possible.

The circumstances per session were constant, so there were no session-dependent circumstances. Specifically, we controlled the level of external noise, which is not controlled in a real-life scenario, but this will be part of the future work. Additionally such a setup would imply that the keyboard is not standard and users would be using their own personal keyboard with specific characteristics as far as producing noise while typing is concerned.

Besides the audio data collected, we also recorded the timing information from the typing data. This data would be used to calculate the performance of static timing-based KD to see the difference when compared to the performance using the audio typing data.

The password was "password" and coded as $k_1, k_2, ..., k_8$, e.g $k_5 =$ 'w' and $k_3 = k_4 =$'s'. The data for each key m ($m = 1..8$) of the "password" from user i ($i = 1..50$) collected in typing j ($j = 1..25$) of session l ($l = 1..4$) was:

1. The key-down time: $t^{down}_{i,j,l,m}$;
2. The key-up time: $t^{up}_{i,j,l,m}$;
3. The audio file $S_{i,j,l}$ for the full typing of the password.

From the key-down and key-up timings duration and latency can be calculated:

1. $dur_{i,j,l,m} = t^{up}_{i,j,l,m} - t^{down}_{i,j,l,m}$ for $m = 1..8$; and
2. $lat_{i,j,l,m} = t^{down}_{i,j,l,m+1} - t^{up}_{i,j,l,m}$ for $m = 1..7$.

Durations and latencies (see examples in Table 1 and the "p" key duration histogram in Fig. 2) are used to measure the performance of the timing-based KD system which will be used to compare against the performance of the audio-based KD system.

A graphical representation of one audio file is given in Fig. 3. We can clearly distinguish the sound produced by each pressing down and releasing of a key. This particular user had relatively high latencies (distance between key release of one key and key down of the next key), except between the two s'es. Various other users typed faster, resulting in some negative latencies, meaning that the sound of key down of the next key was recorded before the sound of the key up of the current key.

Table 1 Example of Duration and Latencies for 2 users of the database

Duration [key]	User1	User2	Latency [keys]	User1	User2
(p)	78 ms	157 ms	(p,a)	93 ms	62 ms
(a)	63 ms	156 ms	(a,s)	141 ms	172 ms
(s)	46 ms	141 ms	(s,s)	79 ms	203 ms
(s)	46 ms	156 ms	(s,w)	125 ms	344 ms
(w)	63 ms	141 ms	(w,o)	109 ms	187 ms
(o)	78 ms	110 ms	(o,r)	79 ms	375 ms
(r)	62 ms	93 ms	(r,d)	109 ms	203 ms
(d)	63 ms	110 ms	–	–	–

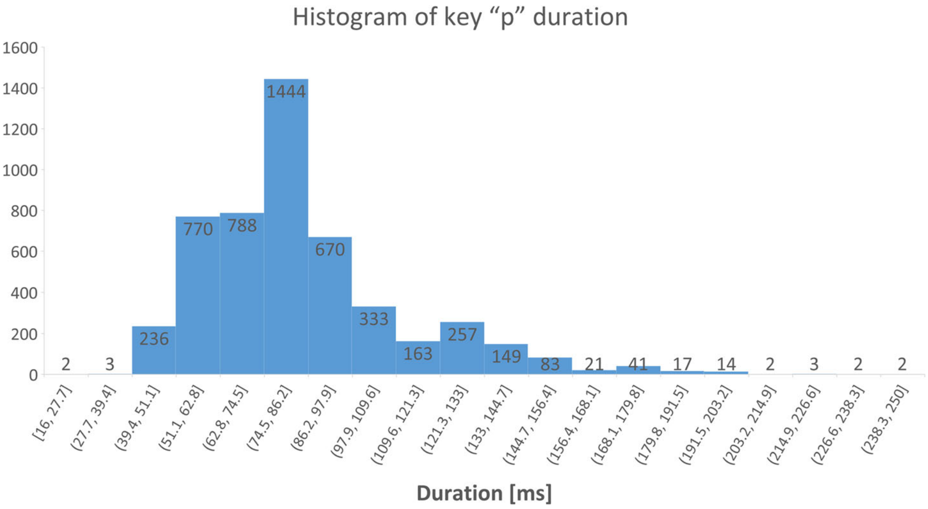


Fig. 2 Histogram of key "p" duration for the word "password"

The database is currently in preparation for release in a competition at a biometric conference. The database needs to fulfill Norwegian policies and any personal information related to the participants will be carefully removed. The Competition should benchmark state-of-the art algorithms on the data captured and will serve for public evaluation of the audio and timing keystroke dynamics analysis.

4 Data analysis

This section is split into two parts. In the first part we will focus on the baseline analysis based on the timing information, while the second part focuses on performance on audio related information.

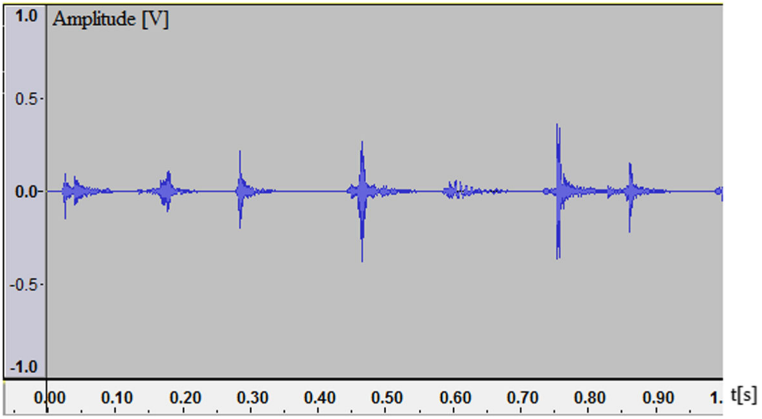


Fig. 3 Example WAV file for typing "password"

4.1 Timing based analysis

We used the collected timing information to calculate the duration and latency features for all 100 typings of each of the 50 participants. We used 1 session for template creation and 3 sessions for testing, and used cross-validation to assure that all 4 sessions are used for template creation. Alternatively we used 3 sessions for template creation and 1 for testing in a similar setup [5]. We evaluated the performance of the system both for authentication and for identification. The template of a user consisted of the mean and standard deviation for each of the 8 durations and 7 latencies [2].

The distance metric used was the Scaled Manhattan Distance (SMD) as this is the best performing distance metric according to [10]. If a template is denoted by $T = ((\mu_1, \sigma_1), (\mu_2, \sigma_2), \ldots, (\mu_{15}, \sigma_{15}))$ and the test input is denoted by $t = (t_1, t_2, \ldots, t_{15})$, then the Scaled Manhattan Distance is equal to:

$$d(T, t) = \sum_{i=0}^{15} \frac{|\mu_i - t_i|}{\sigma_i}.$$

When creating the template we removed outliers, by ignoring data samples that were more than 2 standard deviations away from the mean.

4.2 Audio based analysis

Similarly as in the previous section the acoustic data were split into a training and a testing set. To capture the nature of the acoustic signal, the MFCC (Mel-Frequency Cepstral Coefficients) features were applied [9]. Figure 4 shows the MFCC features for a single typing of a user (i.e. single recorded audio file). These features were extracted from 25ms Hamming windows with a 10ms frame shift. The Mel-filter bank was created by 26 filters and the final number of cepstral coefficients was set to 12. We also used log energy, and first and second time derivatives of the 12 static coefficients as features. One signal frame was finally described by a 39 dimensional MFCC feature vector (MFCC$_{EDA}$).

The Hidden Markov Model (HMM) based approach was employed for the classification. Each user was modeled by *ergodic* HMMs from 1 to 7 states (see 2 state example in Fig. 5) and from 1 to 1024 Gaussian Mixtures in each PDF (Probability Density Function). Each model was first initialized using initial means &variances values and re-estimation using Viterbi algorithm reassigning of the observation vectors to HMM states. Then the Baum-Welch re-estimation procedure was iteratively applied until the estimates change was smaller than 0.0001. The number of mixtures was then duplicated and the Baum-Welch procedure repeated. All user models were created and evaluated in off-line tests (using the HTK

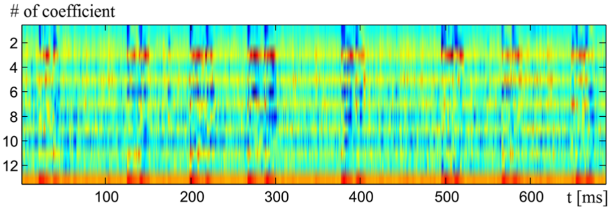


Fig. 4 Example of 13 static MFCC features for one user

Fig. 5 Example of transition matrix of 2 state ergodic HMM model with additional enter and exit state

```
<TransP> 4
   0.0 1.0 0.0 0.0
   0.0 0.6 0.4 0.0
   0.0 0.3 0.5 0.2
   0.0 0.0 0.0 0.0
<EndHMM>
```

toolkit [25]), but generally 5, 6 and 7 states achieved worse results compared to HMMs with a lower number of states, therefore these results are not presented here. It is mainly because of lack of the data for successful training of these models. In the authentication task we identified the problem of inconsistent output probability between the testing utterances so we used a number of frames and log energy for normalization of the recognition score (log probability of the particular model) and calibration approach later on. No adaptation of the model was realized because of the considered application scenario. The planned application should have all the models prepared for scoring after each password capture, so the models of all users are trained using the training utterances from the initialization phase. The possible universal model adaptation for decreasing the number of necessary training utterances will be evaluated in the future work.

5 Results

In this section, results of the identification and authentication analysis with timing-based as well as audio-based features, results of analysis with fused features and the comparison of obtained results will be presented. These are different analysis methods and the performance reporting should not be mixed. A good performance for identification will not automatically indicate a good performance for authentication or vice versa. We choose the Accuracy for Identification and Equal Error Rate (EER) for authentication tasks.

As in the previous section, subsections will provide separate results of the identification task and authentication task for audio and timing information analysis. Next, the comparison of the results and fusion of timing and audio results for authentication task will be provided.

The 4-fold *cross-validation* was realized for all test using always 4 combinations of training and testing recording sets, which means all recordings were used in one of the setups as training and as testing in another one. The average of these 4 tests is the *cross-validated* result.

5.1 Identification task

5.1.1 Timing based results

When evaluating the identification task, with the same template and distance metric as in the authentication task, we obtained a rank-1 accuracy of 56.7% when using 1 session for training and 64.6% when using 3 sessions for training. Full Cumulative Matching Characteristic Curve (CMC) for the latter case is given in Fig. 6. The CMC curve describes the Rank-N vs Accuracy plot, which means that the tested user was successfully identified among the N top scores.

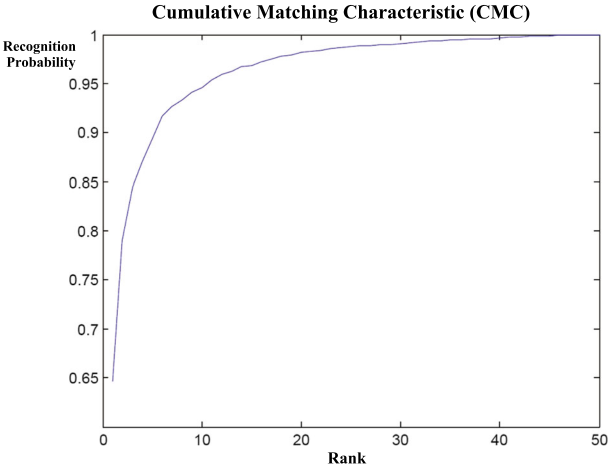


Fig. 6 Cumulative Matching Characteristic Curve for 3 training sessions and 1 test session of the timing identification analysis (Rank-N vs Accuracy)

We clearly see that the identification accuracy is not very high and only increases slightly when using three sessions for training and the remaining session for testing.

5.1.2 Audio based results

When using the audio based information in an identification setting, we noticed that the results were much better compared to timing analysis. The accuracy highly depended on the number of Gaussian Mixtures (GM) in PDFs and the number of HMMs used. Table 3 gives a partial overview of the accuracies obtained for various values of PDFs, GMs used and various HMM states. The Table 2 shows the results when using 3 sets for training and 1 for testing. It can be seen that the best result is obtained for 3 HMM states in combination with 128 GM in a PDF, but other settings give results that are almost as high.

We have also done another similar test, but this time using only a single session for training the system, were the other 3 sessions are used for testing. In this case the results are significantly lower, as can be seen in Table 3. We also noted, when comparing the two tables, that the range of accuracy values is much broader when using only 1 session for training which is mainly because of the lack of the training data amount. We want to explore this phenomenon in the future work where we'll have larger amount of data and use the universal model adaptation for decreasing the amount of data needed to train the user model.

Table 2 Accuracy of audio analysis *cross-validated* identification results when using 3 sessions for training and 1 for testing

PDF type \ # HMM states	1	2	3	4	5
64	94.6%	95.7%	96.8%	94.9%	92.7%
128	96.0%	96.8%	**97.3%**	96.6%	94.4%
256	96.9%	97.0%	96.7%	96.2%	94.4%
512	97.2%	95.6%	92.8%	92.2%	90.0%
1024	96.1%	83.4%	71.5%	62.0%	70.8%

Table 3 Accuracy of audio analysis *cross-validated* identification results when using 1 session for training and 3 for testing

PDF type \ # HMM states	1	2	3	4	5
64	88.3%	89.6%	90.0%	86.5%	81.7%
128	**90.6%**	88.5%	86.2%	82.9%	77.1%
256	90.0%	76.7%	65.4%	59.1%	58.1%
512	83.5%	39.1%	26.0%	25.1%	29.2%

In Table 4 you can see the cross-validated results of the best audio models for each scenario. It is interesting that the randomly selected recordings bring the best results. This means the typing behavior slightly changed between the sessions. For example, we discovered that when using only one session for training the system achieved 90.62% (cross-validated) and within this scenario the best result was 92.91% with the second session used for training, and the worst was 88.93% with the fourth session in training. It means that the first session when the user experienced the typing of the password for the first time was not the worst but not the best either. Nevertheless, in real-life application, the user chooses a password which is familiar to him, and therefore it should be easier to type. The problem of poor results with last session in training should not affect a real-life application because the users usually do not lose their typing habits over time.

5.2 Authentication task

5.2.1 Timing based results

The obtained Equal Error Rate (EER) when using 1 session for training and 3 for testing was equal to 14.4%, which dropped to 11.7% when 3 sessions were used for training and 1 for testing. The EER for authentication is at an acceptable level, more so given the fact that it is a short password (only 8 characters, i.e. 15 features) and that it is a common English word that most likely is easy to type for all participants.

5.2.2 Audio based results using calibration

We first evaluated the audio information for authentication purposes. However, when using a single session for training and the remaining three sessions for testing, we found that the EER was as high as 21.1%. Even using 3 sessions for training and the last session for testing did not improve the results significantly, as the EER decreased only to 19.1%. These results are obviously too poor for practical purposes, and more importantly they are worse than the results we obtained based on the timing analysis.

Comparing to identification results, one can clearly identify the problem of output probability inconsistency between the testing utterances. For identification task we used the

Table 4 Comparing *cross-validated* audio analysis identification accuracy results when using the best acoustic models

# of test sessions	Best model	Accuracy
3 (1 training sess.)	1-state 128-PDF	90.62%
2 (2 training sess.)	1-state 256-PDF	95.64%
1 (3 training sess.)	3-state 128-PDF	97.30%
25 random recordings	1-state 256-PDF	**99.33%**

Table 5 Comparing *cross-validated* rank-1 *identification accuracy* results when using the best acoustic models and general timing analysis models

# of test sessions	Timing analysis Accuracy	Audio analysis Accuracy (Model)
1	64.6%	**97.30%** (3-state 128-PDF)
3	56.7%	90.62% (1-state 128-PDF)

probabilities of every user model enrolled in the system for computing the current utterance probability. Then the probabilities were normalized by the number of frames and energy of the test recording (supported by HTK Tools) for authentication task. As they were still in different range for every recording, no meaningful threshold could be used for authentication task, so we decided to use the first user as a benchmark. It means that the first user could be used in a real system to calibrate the setup (environment sound), and then this calibration model could be used for normalization of the gathered tested user model probability. The main idea is that using the two results from two models on the same recording will bring us a benchmark information about levels of the scores on that particular recording.

We tried computing the distance between the calibration and genuine model probability and then the distances of the genuine (1 for every test recording) and the impostor (48 for every test recording) models were used to compute the final EER using the formula below.

$$Logprob_{calibrated} = \frac{Logprob_{actual}}{Logprob_{calibration\_user}} \times 50$$

The authentication results were varying between 9.4% and 14.8% EER. The best cross-validated result of 11.6% was achieved for 512 PDF with only 1 state HMM model and 3 training (enrolling) sessions. The worst (21%) for 3 training sessions was achieved with 3 state 1024 PDF HMM model. For the more realistic scenario of only 1 training session the best cross-validated result of 16.6% EER was achieved using 3 states 64 PDF HMM.

5.3 Comparison of timing and audio analysis

In this section we will make a comparison between the performance results based on timing information and the audio information (Tables 5 and 6).

What we can clearly observe is that audio-based and timing-based KD both perform differently. Most notably, timing-based KD performs significantly better in an authentication task (see Table 6), while in an identification task the performance of audio-based KD is much better (see Table 5). Given the high performance of audio KD in case of identification, we assume that we should be able to gain better performance in authentication as well. The main hurdle at this moment is that the distance scores need to be normalized.

Table 6 Comparing *cross-validated* EER *authentication* results when using general Timing & Audio analysis models with the calibrated results of the best acoustic models

# of test sessions	Timing EER	Audio EER	Best calibrated audio EER
1	11.7%	19.1%	**11.6%**
3	14.4%	21.1%	16.6%

Table 7 EER authentication results when using *best acoustic models* and *fusing* them with relevant timing models (same train/test sessions) using simple linear approach and Bosaris toolkit

# of test sessions	Timing analysis system EER	Calibrated audio EER	Linear Fused system EER	Bosaris toolkit fused EER
1	9.91%	8.99%	**4.65%**	4.71%
3	12.08%	14.34%	7.54%	7.30%

5.4 Fusion of the timing and audio analysis results for authentication task

First of all, we chose the best models from audio authentication setup for 1 and 3 training sessions. The chosen models were used for fusion of the calibrated results with timing-based analysis distances. The fusion was done using simple multiplication of distances after putting them in the same ratio from 0 to 200. It was necessary to suppress the results of the first user which was used for audio probability calibration described above (also for Timing analysis). So the results are for 49 user authentications without cross-validation.

We used Bosaris[1] toolkit [6] with widely used fusion approach for speaker identification/authentication [16] or Query by Example Search on Speech [19] for results comparison. We used a half of the testing set as development subset for fusion function training and applied it to the rest of the testing set - evaluation subset. After that we swapped the subsets and the average EER is presented in the Table 7.

For the fusion of 1 test session we chose a scenario where session 3 was used for testing and 1, 2 and 4 for training of the 1 state 512 Gaussian mixtures HMM model. For 3 test session scenario we chose the session 2 for training and sessions 1, 3 and 4 for testing of the 3 state 64 Gaussian mixtures model. The results of the original timing and calibrated audio distances compared with fused ones are depicted in the Table 7 and compared using Detection Error Trade-off (DET) Curve in Fig. 7. The Bosaris toolkit results as DET curves are presented in Fig. 8a and b. It is clear from the data that the fusion of the timing and calibrated audio systems provides significantly better results then each of them alone.

6 Results discussion

The fusion of the timing and calibrated audio analysis results for authentication task leads to 4.65% EER (timing 9.91%, audio 8.99%) for 3 training sessions and 7.30% EER (timing 12.08%, audio 14.34%) for 1 training session (25 utterances) using half of the testing data as development subset using Bosaris toolkit. Both are approximately one half of the separate timing and audio analysis EER.

For identification task, the best result achieved by audio analysis with one state 256 PDF model trained on 75% randomly selected recordings was 99.33%, which became 97.03% after cross-validation (where entire sessions were test sets). When using two sessions for training the accuracy decreased to 95.64% after cross-validation. With only one session for training the system achieved 90.62% (cross-validated) and in this case the single best result was 92.91% for the second session used for training, and the worst was 88.93% for the fourth session in training.

[1] http://sites.google.com/site/bosaristoolkit

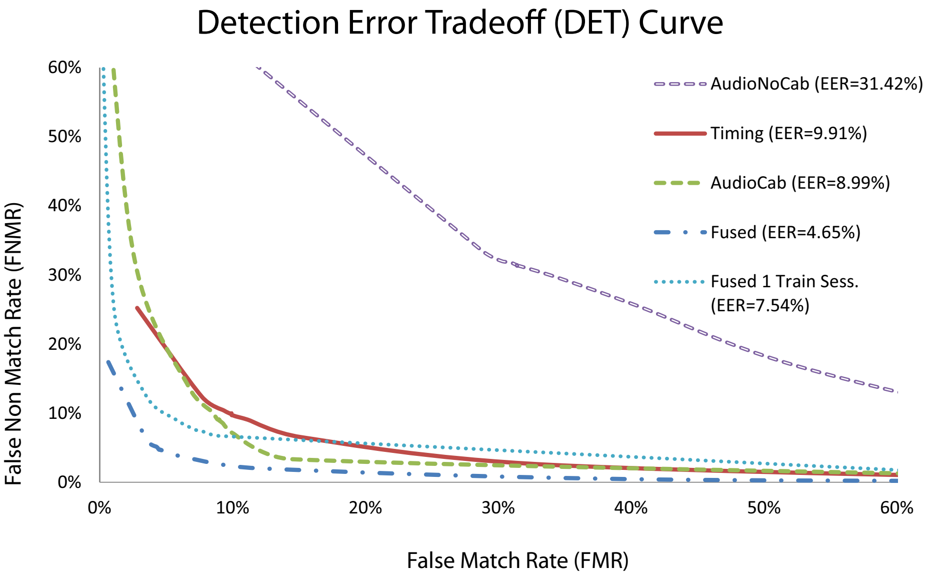


Fig. 7 Authentication Results Detection Error Trade-off (DET) Curve for 3 training sessions using Audio analysis with no calibration (AudioNoCab), using only Timing analysis (Timing), Calibrated audio results (AudioCab), Fused calibrated audio & timing results (Fused), and lastly the same for only 1 training session and 3 testing sessions (Fused 1 Train Sess.)

From the above analysis we see that acoustic information obtained from typing a password does not provide high quality data for authentication purposes without calibration - which means one user must calibrate the keyboard before the authenticated user. When

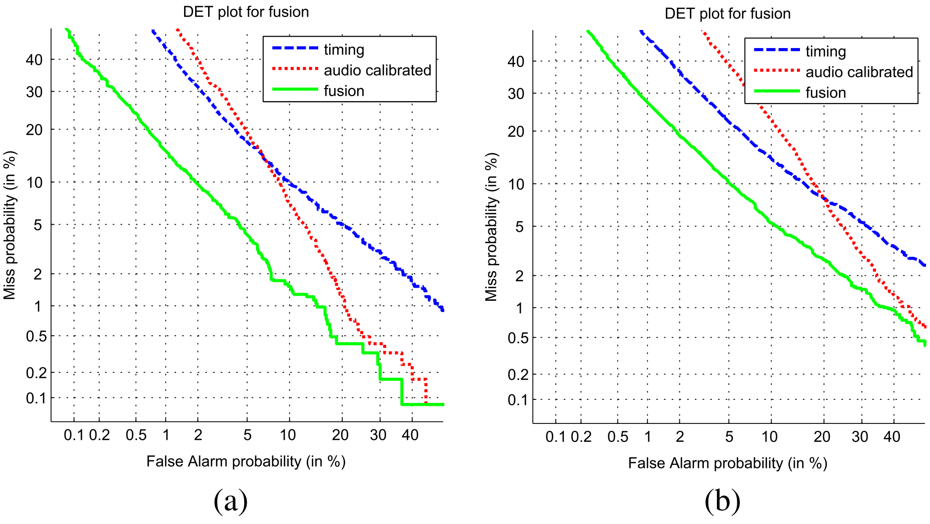


Fig. 8 Bosaris Toolkit fusion algorithm Authentication Results Detection Error Trade-off (DET) Curve for (**a**) 3 training sessions and (**b**) 1 training session

using the timing information obtained at the same time, the performance is higher (11.7% compared to 19.1% Equal Error Rate). For calibrated audio data analysis the results of timing and audio analysis are comparable. It is interesting that, for identification purposes, this conclusion is reverse, as the rank-1 accuracy based on the timing data is only 64.6% while it is 97.3% for audio based data.

It should be taken into consideration that no timing features were taken into account for audio identification analysis model in this phase and the best results are for 1 state HMM model for 8 character password. We can say the color of the sound (fingers touching the keys) play the most important role for audio analysis. Of course in the future we plan to add also timing data into the feature vectors or train 8 - 16 state HMM (close to the number of clicks in the recording), but that will require more data or training of combined (timing & audio) HMM models, which is not a trivial task.

In this paper we have used a restricted setting, where a fixed keyboard and microphone were used and background noise was reduced as much as possible. This setting is in line with similar research on acoustic keystroke dynamics. Any variation of such setting is worth investigating, as well as an extension to a continuously processing setting, where audio produced while typing on a keyboard can be used to confirm that no change of the user has occurred.

Different keyboards have been tested with regard to timing information and it has been shown that change of hardware does not have a major impact on the performance. As far as the audio information is concerned, there would be a major influence, but this has not yet been tested and shall be part of future research. Future research will also include introducing background noise while typing and combination with mouse dynamics [14].

7 Conclusions

We have conducted the experiment to investigate whether timing-based KD analysis or acoustic-based KD analysis of typing a password would give a better performance, both for authentication and for identification task. Next, we made a fusion of the timing and audio analysis for authentication setup and achieved a significant improvement of the Equal Error Rate (EER). In the proposed solution there is no extra effort expected from a user, all data is collected in one action while the performance is clearly better.

The best cross-validated identification accuracy of 97.03% and authentication EER of calibrated models of 11.6% was achieved using audio-only analysis.

The fusion of the timing and calibrated audio analysis results for authentication task leads to 4.65% EER with 3 training sessions and 7.54% EER with 1 training session (25 utterances).

Other fusion techniques [17] and feature analysis could lead to better results in our future work and to a reliable authentication application. We plan to evaluate the fusion of timing features and MFCC features and use only HMM models for testing. Another approach is investigating different score fusion techniques and normalizations/calibrations together with adaptive audio modeling and testing environment, and keyboard adaptation to increase the robustness of the audio analysis.

In the future we want to test the system on bigger databases or capture new recordings with different setups to prove the methodology in more real conditions. For example, to have a special audio setup for every end-user (plus one for calibration - the technician performing laptop installation) for the scenario where the user wants to be authenticated on his own laptop (using internal microphone and built in keyboard).

8 Thanks

We would like to thank all the anonymous participants in the experiment who spent their time so that we could obtain the data that was used in the analysis described in this paper. We are also thankful to COST IC1106 "Integrating Biometrics and Forensics for the Digital Age" Action which brought us together on IWBF workshop organized by the Action and partially by the INDECT IP FP7 Project (Project ID: 218086). Finally we want to thank Pauly Dutko (To Pang Ching) for English proofreading and corrections.

Acknowledgments This publication was supported partially by the Ministry of Education, Science, Research and Sport of the Slovak Republic under the projects VEGA 1/0075/15 & partially by the Slovak Research and Development Agency under the contracts No. APVV-15-0517 & APPV-15-0731.

References

1. Asonov D, Agrawal R (2004) Keyboard acoustic emanations. In: 2004 IEEE Symposium on Security and Privacy, 2004. Proceedings, pp 3–11
2. Banerjee SP, Woodard DL (2012) Biometric authentication and identification using keystroke dynamics: A survey. J Pattern Recognition Research 7(1):116–139
3. Barisani A, Bianco D (2009) Sniffing keystrokes with lasers/voltmeters. Proceedings of Black Hat USA
4. Bours P (2012) Continuous keystroke dynamics: A different perspective towards biometric evaluation. Inf Secur Tech Rep 17:36–43
5. Bours P, Kiktová E, Pleva M (2015) Static Audio Keystroke Dynamics. In: Dziech A, Leszczuk M, Baran R (eds) Multimedia Communications, Services and Security. Communications in Computer and Information Science, vol 566. Springer, Cham pp 159–169, doi:10.1007/978-3-319-26404-2_13
6. Brümmer N, de Villiers E (2013) The Bosaris toolkit: Theory, algorithms and code for surviving the new DCF. arXiv preprint arXiv:1304.2865
7. Dozono H, Itou S, Nakakuni M (2007) Comparison of the adaptive authentication systems for behavior biometrics using the variations of self organizing maps. Int J Comput Commun 1(4):108–116
8. Imperva Consumer password worst practices (2010) Technical report, The Imperva Application Defense Center (ADC)
9. Kiktova E, Lojka M, Pleva M, Juhar J, Cizmar A (2013) Comparison of different feature types for acoustic event detection system. In: Dziech A, Czyewski A (eds) Multimedia Communications, Services and Security, volume 368 of Communications in Computer and Information Science, pp 288–297. Springer Berlin Heidelberg
10. Killourhy KS, Maxion RA (2009) Comparing anomaly-detection algorithms for keystroke dynamics. In: DSN'09. IEEE/IFIP International Conference on Dependable Systems & Networks, 2009, pp 125–134. IEEE
11. Liang W, Bours P (2014) Content reconstruction using keystroke dynamics: Preliminary results. In: 2014 5th International Conference on Emerging Security Technologies (EST), p. 13–18
12. Li Z, Zhou F, Tygar JD (2009) Keyboard acoustic emanations revisited. ACM Trans Inf Syst Secur (TISSEC) 13(1):3
13. Luo J-N, Yang M-H (2016) A mobile authentication system resists to shoulder-surfing attacks. Multimed Tools Appl 75(22):14075–14087
14. Mondal S, Bours P (2017) A study on continuous authentication using a combination of keystroke and mouse biometrics. Neurocomputing 230:1–22. doi:10.1016/j.neucom.2016.11.031
15. Nakakuni M, Dozono H, Itou S (2008) Adaptive authentication system for behavior biometrics using supervised pareto self organizing maps. In: 10th WSEAS International Conference on Mathematical

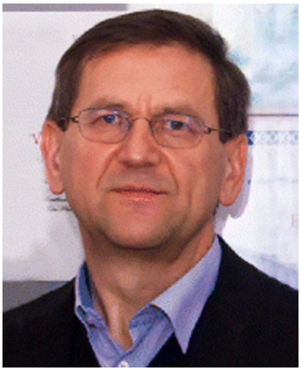
Methods, Computational Techniques and Intelligent Systems, MAMECTIS'08, pages 277–282, Stevens Point, Wisconsin, USA, 2008. World Scientific and Engineering Academy and Society (WSEAS)

16. Novotný O, Matejka P, Plchot O, Glembek O, Burget L (2016) Analysis of speaker recognition systems in realistic scenarios of the SITW 2016 challenge. Interspeech 2016:828–832
17. Peng J, Li Q, El-Latif AAA, Niu X (2015) Linear discriminant multi-set canonical correlations analysis (LDMCCA): an efficient approach for feature fusion of finger biometrics. Multimed Tools Appl 74(13):4469–4486
18. Rao KR, Anne VPK, Chand US, Alakananda V, Rachana KN (2014) Inclination and pressure based authentication for touch devices. In: ICT and Critical Infrastructure: Proceedings of the 48th Annual Convention of Computer Society of India-Vol I, pages 781–788. Springer
19. Rodriguez-Fuentes LJ, Varona A, Penagarikano M, Bordel G, Diez M (2013) GTTS systems for the SWS task at MediaEval 2013. In: Working Notes Proceedings of the MediaEval2013 Workshop, Barcelona, Spain, October 18-19, CEUR-WS. org, ISSN 1613–0073
20. Roth J, Liu X, Ross A, Metaxas D (2013) Biometric authentication via keystroke sound. In: 2013 international conference on biometrics (ICB), pages 1–8. IEEE
21. Roth J, Liu X, Ross A, Metaxas D (2015) Investigating the discriminative power of keystroke sound. IEEE Trans Inf Forensics Secur 10(2):333–345
22. Tasi C-J, Chang T-Y, Cheng P-C, Lin J-H (2013) Two novel biometric features in keystroke dynamics authentication systems for touch screen devices. Security and Communication Networks 7(4):750–758
23. Teh PS, Teoh BJ, Andrew, Yue S (2013) A survey of keystroke dynamics biometrics . Sci World J:1–24. Article ID 408280
24. Traore I, Woungang I, Obaidat MS, Nakkabi Y, Lai I (2014) Online risk-based authentication using behavioral biometrics. Multimed Tools Appl 71(2):575–605
25. Young S, Evermann G, Kershaw D, Moore G, Odell J, Ollason D, Povey D, Valtchev V, Woodland P (2006) The HTK Book, version 3.4. Cambridge University Engineering Department. Cambridge University Press, UK

Matúš Pleva was born in Kosice, Slovakia in 1977. In 2010 he graduated PhD. in study program Telecommunications at the Department of Electronics and Multimedia Communications of the Faculty of Electrical Engineering and Informatics at the Technical University of Kosice. He works as a researcher in the field of the acoustic modeling, acoustic event detection, speaker recognition, speech processing, human-machine interaction, security & biometrics, networking, etc. He was MC member in IC1106 COST action "Integrating Biometrics and Forensics for the Digital Age". He also participated in more than 30 national and international projects and COST actions. He has published over 90 technical papers in journals and conference proceedings.

Patrick Bours was born in Sittard in the Netherlands. He got his MSc and PhD in Discrete Mathematics from Eindhoven University of Technology in the Netherlands with a specialization in Coding Theory. He worked for 10 years for the Dutch Government in the area of asymmetric crypto and since 2005 he is working at Norwegian University of Science and Technology, Gjøvik, first as a PostDoc, since 2008 as an Associate Professor and since 2012 as Professor. His research focus is on behavioral biometrics with a special interest in keystroke dynamics and continuous authentication. His research also includes finding innovative manners to identify persons based on their daily behavior. He has published over 90 technical papers in journals and conference proceedings.

Stanislav Ondáš was born in Prešov, Slovakia in 1981. In 2004 he graduated M.Sc. (Ing.) at the Department of Electronics and Multimedia Communications of the Faculty of Electrical Engineering and Informatics at the Technical University of Košice. He received his Ph.D. degree at the same department in the field of Telecommunications in 2008. He is currently working as an Assistant Professor in the Laboratory of Speech and Mobile Technologies at the same department. He is a specialist in the field of human-machine interaction, dialogue modelling and management, natural language processing and semantic analysis.



Jozef Juhár was born in Poproč, Slovakia in 1956. He graduated from the Technical University of Kosice in 1980. He received Ph.D. degree in Radioelectronics from Technical University of Kosice in 1991, where he works as full Professor at the Department of Electronics and Multimedia Communications. He is author and coauthor of more than 200 scientific papers. His research interests include digital speech and audio processing, speech/speaker identification, speech synthesis, development in spoken dialogue and speech recognition systems in telecommunication networks. Prof. Juhár is a member of ISCA, AES and IEEE. He is a member of the editorial boards and reviewer of several international scientific journals.