Multimodal Biometric Fusion — Joint Typist (Keystroke) and Speaker Verification

Jugurta R. Montalvão Filho and Eduardo O. Freire

***Abstract*— Identity verification through fusion of features from keystroke dynamics and speech is addressed in this paper. We experimentally compare performances from identity verification algorithms based on fusion of median pitch (prosodic level information), mel-frequency cepstral coefficients (short-time spectral level information) and keystroke dynamics (down-down intervals). These experimental results are also compared to the corresponding unimodal performances. All experiments are carried out on a small but true multimodal public database. Three fusion strategies are experimentally compared. A nonnegligeable performance improvement is then observed, mainly with a simple linear fusion strategy based on fusion of estimates.**

***Index Terms*— Multimodal Biometry, Keystroke Dynamics, Speaker Verification, Typist Verification, Biometrics Fusion.**

I. Introduction

In biometrc-based systems for identity verification, static and/or dynamic biometric measures may be used as personal "passwords". Consequently, most security systems based on biometric signals demand specific data acquisition hardware.

Nevertheless, there are some possible exceptions to this rule. One of them is typing biometrics, more commonly referred to as keystroke dynamics. Indeed, keystroke dynamics looks at the way a person types or pushes keys on a keyboard.

Furthermore, thanks to the widespread use of voice communication over the Internet, headset devices and soundcards (with analogic/digital converter inside) became almost as common as the keyboard itself, in conventional personal computer setups.

In this work, we address identity verification with both biometric signals: speech and keystroke dynamics, first through unimodal approaches, with straightforward algorithms, and finally through fusion of those biometrics, at the matching score level[1], [2].

Fusion of two or more biometrics in automatic identity verification systems can provide more reliable systems and, recently, this issue has received an increasing attention (see [3], [4], [1], [2], for instance).

Nevertheless, to the best of our knowledge, there is no previous work in the literature that addresses keystroke dynamics and speech biometrics fusion, though it seems to be potentially useful, mainly if we consider Internet applications.

Furthermore, in spite of the weak discrimination between users that keystroke usually provides (excepting some experiments with long typed texts, as in [12]), it is almost immune to background noise, if compared to speech signals. As a result, we may expect a bimodal system better than individual ones, and more robust to noise.

The authors are with the Universidade Federal de Sergipe (UFS), São Cristóvão, CEP. 49100-000. E-mail:jmontalvao@ufs.br, efreire@ufs.br.

This paper is organized as follows: in Section II, the data acquisition procedure is presented. Then, in Section III, it is explained how features are extracted from speech and keystroke dynamics, along with a brief description of the algorithms used for features comparison. In Section IV, fusion strategies used in this work are presented and, finally, both unimodal and multimodal experimental results on identity verification are presented in Section V.

II. Data Acquisition

Both speech and keystroke samples correspond to signals recorded during the uttering/typing of a single set of four words in Portuguese — "*chocolate, zebra, banana, táxi.*" — equally spelled in English, apart from the accent in "táxi".

Each subject uttered/typed this set of four words 10 times, 5 (five samples) during a first session, and 5 more samples during a second session, about a month later.

All subjects, men and women not necessarily familiar to a computer keyboard, were invited to type on the very same conventional keyboard (standard 101/102 keys, Brazilian layout - similar to the EUA layout), in our laboratory, in both sessions. Sequences of down-down (DD) time intervals from two keystrokes were thus recorded.

Likewise, during each session, each subject was asked to utter the same four words five times, using a conventional headset (electret microphone plus headphone), whose microphone was plugged to an Analogic/Digital converter. Speech signals were thus digitalized and recorded with 16 bits per sample, at 22050 samples per second. All recordings were made in our laboratory, under low background noise. For a while, only 10 subjects were invited to take part in the experiment.

Several approaches toward studying fusion, presented in the literature (see [5] and references therein), use virtual identities, also known as "chimeric" users, where a biometric modality from one person is paired with the biometric modality of another person. By contrast, in our experiments, the database is a true multimodal set of samples, publicly available (see Section VI).

In order to simplify explanation, we denote samples as follows: each subject S_i, $i = 1, 2, \ldots, 10$, provided 10 bimodal samples $s_{i,m}$, $m = 1, 2, \ldots, 10$, were samples from $m = 1$ to $m = 5$ were recorded during the first session, whereas samples from $m = 6$ to $m = 10$ were recorded during the second session, about one month later.

III. Features Extraction and Comparison

From each recorded sample, $s_{i,m}$, three kind of features are extracted, namely:

(a) Long-term spectral features: median pitch from structured utterances (duration of each utterance: 3 s);
(b) Short-term spectral features: sequences of 13 Mel Frequency Cepstral Coefficients (MFCC) vectors, from the same utterances of (a);
(c) Keystroke based features: sequences of DD time intervals from the typing of structured texts (31 keystrokes).

It is worth noting that, for the speaker recognition task, two levels of information are taken into account [6]: spectral level, through MFCC vectors; and prosodic level, through the median of the pitch variation in each utterance.

Since typed texts are structured (i.e. the very same text was typed by every subject), keystroke samples are compared according to the simplest algorithm proposed in the seminal work on keystroke dynamics by Bleha et al. [7]. However, according to [8], we know that algorithm performance can be greatly improved if the nonlinear memoryless mapping

$$g(\Delta t) = \frac{1}{1 + \exp\left(-\frac{K(\log_e(\Delta t) - \mu_y)}{\sigma_y}\right)}$$

where $K = 1.7$, $\mu_y = -1.56$ and $\sigma_y = 0.65$ is applied to each DD interval, Δt, prior to comparison.

Indeed, it is shown, in [8], that the random variable that models DD intervals is approximately log-normal and, consequently, the memoryless mapping $g(\cdot)$ significantly improves performance of verification algorithms that do not compensate for the unbalanced probability density functions of this random variable.

The resulting comparison between keystroke timing samples from each pair of (multimodal) samples $s(i,m)$ and $s(j,n)$ — i.e. samples m and n, from subjects i and j, respectively — is a distance denoted by $d_K(s(i,m), s(j,n))$.

On the other hand, each utterance (3 s of digital audio) is pre-processed as follows:

1. The signal is split into nonoverlapping time frames of 45 ms;
2. Pitch is estimated from each frame;
3. From the estimated pitch and the frame power, each frame is classified as voiced or unvoiced;
4. Unvoiced frames are discarded;
5. Median pitch is estimated from the remaining frames (only voiced signals);
6. Finally, from each remaining window, a sub-window of 10 ms is taken and then mapped into 13-MFCC column vectors.

Consequently, from step 5, a median pitch (scalar) is estimated from each utterance, whereas, from step 6, a sequence of 13-dimensional vectors (a matrix, where each column is a MFCC from a frame) is obtained.

Median pitchs from (multimodal) samples $s(i,m)$ and $s(j,n)$, are compared in a straightforward manner: $d_P(s(i,m), s(j,n))$ is just the absolute difference between median pitchs from $s(i,m)$ and $s(j,n)$.

Finally, concerning the comparison of two utterances through its short-term spectral features, i.e. their respective sequences of MFCC, vectors are aligned through time warping [11] and Euclidean (l_2-norm) distances between corresponding vectors are summed up to provide the distance $d_M(s(i,m), s(j,n))$.

Therefore, from each pair of bimodal samples $s(i,m)$ and $s(j,n)$, three unimodal distances are obtained, namely: d_K, d_P and d_M, that can be used separately, for monomodal identity verification, or all together, for multimodal verification.

IV. Data Fusion Strategy

Our approach is based on the assumption that metric fusion, or fusion at the match score level, according to [2], can outperform decision fusion, or fusion at the decision level, in terms of false alarme rate (FAR) and false rejection rate (FRR). It is a quite straightforward reasoning because, clearly, decision fusion can be regarded as a specific case of metric fusion, but the contrary is not true.

Accordingly, we regard each column vector $\mathbf{d} = [d_P, d_M, d_K]^T$ as a point, in the 3D space, to be classified true (Class 1, corresponding to distances between samples from the same subject) or false (Class 2, otherwise). That is to say that the fusion problem become a standard classification problem, where two classes are to be considered.

From this point of view, we dispose of two main approaches to implement distances fusion: (a) Classification after linear projection of incoming vector — linear classification boundary — or (b) Classification of points $\mathbf{d}$ with a nonlinear boundary between classes.

A. Linear Data Fusion with Fisher's Linear Discriminant

A well-known approach to obtain a linear discriminant boundary is the Fisher's Linear Discriminant (FLD)[10]. In few words, the Fisher's solution projects points onto a specific vector $\mathbf{v}_F$, i.e.:

$$y_F = \mathbf{d}^T \mathbf{v}_F$$

Clearly, for $\mathbf{d} = [d_P, d_M, d_K]^T$ and $\mathbf{v}_F = [v_P, v_M, v_K]^T$, for instance, y_F is just a weighted sum of distances:

$$y_F = v_P d_P + v_M d_M + v_K d_K$$

The particularity of the vector $\mathbf{v}_F$ comes from the way it is calculated, i.e.:

$$\mathbf{v}_F = \left((\mathbf{R}_1 + \mathbf{R}_2)^{-1} (\mu_1 - \mu_2)^t \right)$$

where μ_i and $\mathbf{R}_i$ stand for the mean and the covariance matrix of class i. In this case, we have just two classes:

- Class 1 - where samples are vectors, $\mathbf{d}$, of distances between features from the same subject.
- Class 2 - samples are vectors of distances between features from two different subjects;

This projection vector, $\mathbf{v}_F$, is optimal in the sense that it finds a tread-off between within-class dispersion minimization and between-class distance maximization. Consequently, classification accuracy is improved when decision is taken

by comparing the scalar projection y to a preset decision threshold.

Unfortunately, it is not optimized in terms of classification performance, since the decision error minimization is a more complex problem than squared error (dispersion) minimization.

B. Linear Data Fusion based on Optimal Estimation

Another interesting linear fusion approach comes from the optimal fusion of estimates theory, commonly applied to fusion of signal from different sensors (in robotics [9], for instance).

Let $\hat{x}_1$, $\hat{x}_2$ and $\hat{x}_3$ be independent measurements of x (a constant, for instance), disturbed by additive unbiased measurement errors n_1, n_2 and n_3, respectively, modeled as independent random variables. Then, the best linear fusion of estimates, i.e.:

$$y_S = a_1\hat{x}_1 + a_2\hat{x}_2 + a_3\hat{x}_3 \tag{1}$$

that minimizes the variance of y_S, is obtained when [13]:

$$a_i = \frac{\sigma_j^2\sigma_k^2}{\sigma_i^2\sigma_j^2 + \sigma_i^2\sigma_k^2 + \sigma_j^2\sigma_k^2} \tag{2}$$

$i, j, k \in 1, 2, 3$ and $i \neq j$, $i \neq k$, $j \neq k$.

In order to apply this approach to distances between features d_P, d_M and d_K, we should first "centralize" them so that, given that features come from the same subject, all distances are around zero, otherwise, they are closer to one. Thanks to this centralization, we can now see all distances as noisy measurements of a single distance that equals 0 when features come from a single subject — Class 1 —, and equals 1 otherwise — Class 2.

This centralization is obtained as follows:

$$\breve{d}_x = \frac{d_x - \mu_{1,x}}{\mu_{2,x} - \mu_{1,x}}, \quad \mu_{2,x} \neq \mu_{1,x}$$

where $x \in \{P, M, K\}$, and $\mu_{1,x}$, $\mu_{2,x}$ stand for one-dimensional means of classes 1 and 2, respectively.

Finally, the Simplified Fusion of Estimates (SFE) is given by:

$$y_S = a_P\breve{d}_P + a_M\breve{d}_M + a_K\breve{d}_K$$

where indices P, M and K play the role of 1, 2 and 3, in Equation 2, respectively. It is worth noting that, in this approach, a conditional fusion of estimates is provided. More precisely, conditionally to the fact that biometric multimodal features come from the same subject or not, centralized distances $\breve{d}_P$, $\breve{d}_M$ and $\breve{d}_K$ are noisy estimates of constants 0 or 1, respectively.

Figure 1 illustrates the centralization procedure.

C. Non-Linear Data Fusion

In Figure 2, a 2D plot of samples of d_M *versus* samples of d_K illustrates that, though a linear classification boundary seems to provide a good solution (in terms of classification error), it is also clear that a non-linear boundary can, potentially, outperform the linear one.

A quite simple but useful nonlinear classification strategy is the Bayesian Classification for Normal Distribution (BCND),

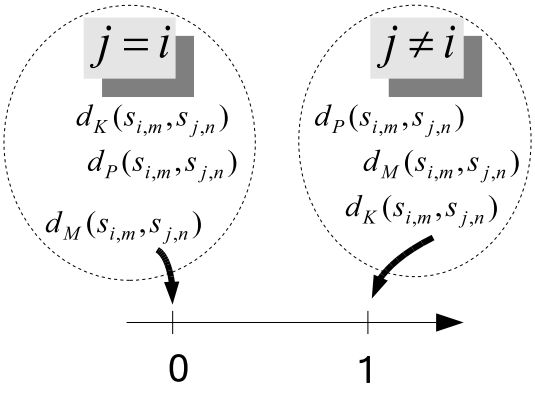

Fig. 1. Distances centralization procedure.

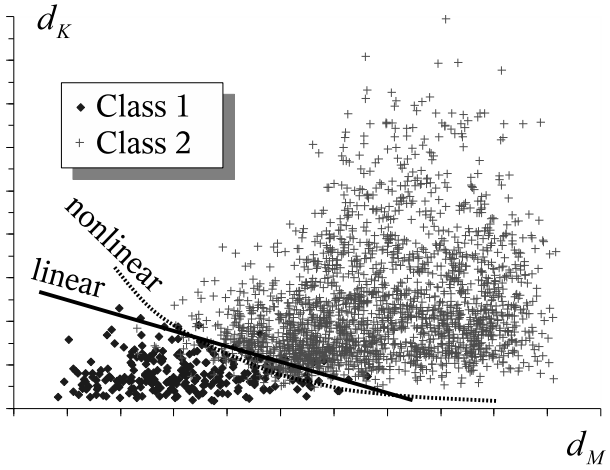

Fig. 2. A 2D projection: d_M versus d_M.

where it is assumed that each class corresponds to a random variable Ω, and that both variables are Normal. Accordingly, each class has a 3-dimensional multivariate normal density probability given by:

$$p(\mathbf{d}|\omega_k) = \frac{1}{\sqrt{(2\pi)^3|\mathbf{R}_k|}} \exp\left(-0.5(\mathbf{d} - \mu_\mathbf{k})^t\mathbf{R}_k^{-1}(\mathbf{d} - \mu_\mathbf{k})\right)$$

Assuming that *a priori* probabilities of classes are the same, $\mathbf{d}$ is then classified by comparing $p(\mathbf{d}|\omega_1)$ to $p(\mathbf{f}|\omega_2)$. Furthermore, in order to provide a flexible balance between FAR and FRR, a parameter λ may be included, yielding the new decision rule: $\mathbf{d}$ is in class k if

$$p(\mathbf{d}|\omega_k) > \lambda p(\mathbf{f}|\omega_l) \;\; \forall k \neq l$$

This rule provides a quadratic (thus nonlinear) decision boundary [10].

Figure 3 illustrates the use of single Gaussians to model classes. Note that mixtures of Gaussians or even Multi-Layer Perceptrons may be alternatively used, at the price of increased computational burden.

V. Experimental Setup and Results

Since our database was built up in two sessions, providing 5 samples per subject per sec session, we do simulate an

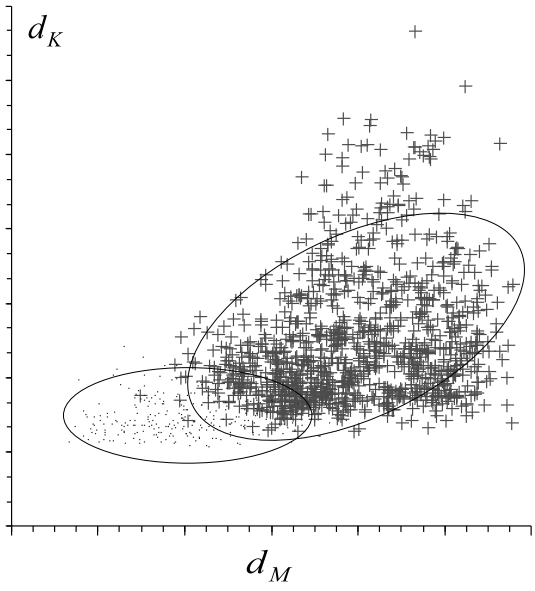

Fig. 3. Modeling classes with single Gaussians.

enrollment procedure by using only samples from the first session as user models (profiles), and performing user verification exclusively with samples from the second session.

Only one sample from each subject, sampled during the first session, is to be used as subject model (profile), and each model is compared to each single sample from the second session, thus 2500 (50 possible models $\times$ 50 samples $\mathbf{x}_i$) comparisons are carried out, being

(a) 250 with template and test signals from the same subject, and
(b) 2250 with template and test signals from different subjects.

For simulations whose results are presented in this section, we refer to the 250 comparisons in (a) as true verification attempts, while the 2250 comparisons in (b) are seen as false ones.

In all experiments, half randomly chosen comparison, i.e. 125 true attempts and 1125 false ones, are used to estimate the classifiers parameters, such as means, covariances and EER thresholds, while the remaining comparisons are used for test. Hence, experimental results presented in this section were obtained from 50% of the 2500 comparisons distances.

A. Unimodal Experiments

When simple unimodal verification experiments are independently performed with distances d_K, d_P and d_M, the corresponding results, in terms of false alarme rate (FAR), false rejection rate (FRR) and equal error rate (EER, the operational point for which FAR equals FRR), are presented in Table I, and Figure 4 illustrates the dependence of FAR and FRR on the threshold choice.

B. Multimodal Experiments

For the multimodal experiments, distances were fused by using (see Section IV):

TABLE I
UNIMODAL VERIFICATION PERFORMANCE — 1 ENTRY PER ENROLLMENT, 1 ENTRY PER VERIFICATION.

Mode	EER
Keystroke	14.8%
Pitch	19%
MFCC	9.7%

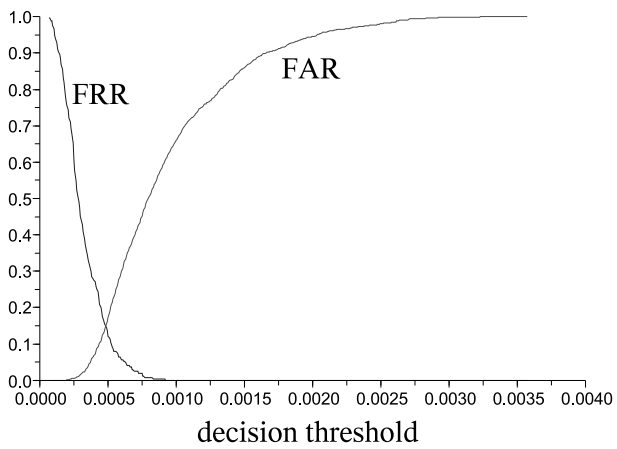

Fig. 4. FAR and FRR from keystroke samples only.

- The Fisher's Linear Discriminant (FLD);
- Simplified Fusion of Estimates (SFE);
- Bayesian Classification for Normal Distribution (BCND).

Figures 5 and 6 illustrate performance results from fusion of median pitch distance (d_P), MFCC distance d_M and keystroke distance (d_K), with FLD and SFE, respectively.

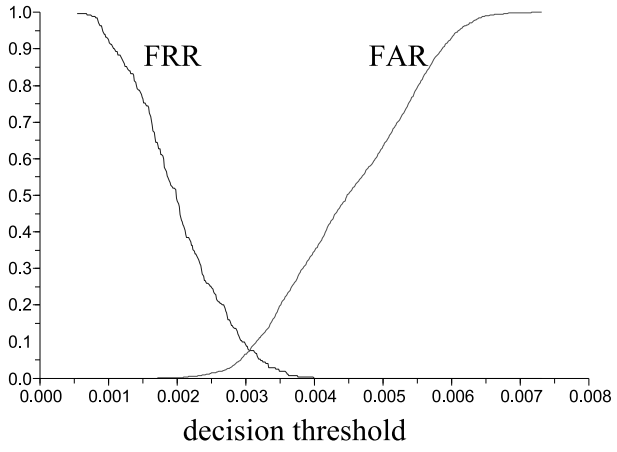

Fig. 5. FAR and FRR from fusion of d_P, d_M and d_K, with FLD.

Alternatively, Figures 7 and 8 are Detection Error Tradeoff (DET) curves, corresponding to the same results shown in Figures 5 and 6.

In order to allow a closer comparison between the two linear fusion strategies, namely FLD and SFE, Figure 9 shows the two DET curves together.

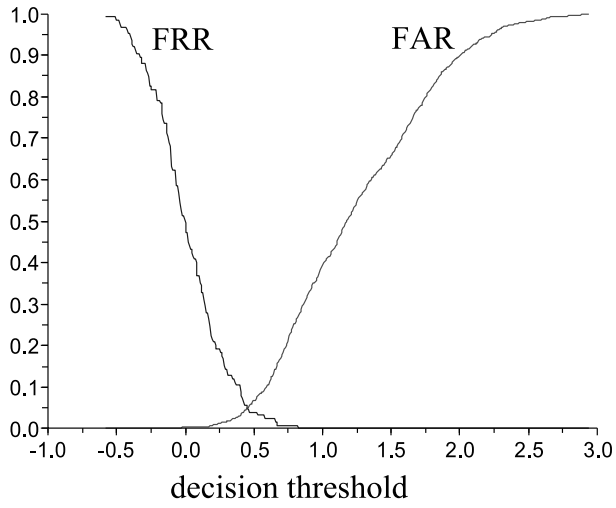

Fig. 6. FAR and FRR from fusion of d_P, d_M and d_K, with SFE.

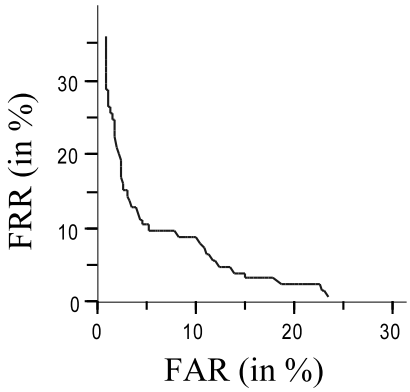

Fig. 7. DET curve from fusion of d_P, d_M and d_K, with FLD.

Tables II, III and IV present some results, in terms of EER, with the three fusion strategies: FLD, SFE and BCND, respectively.

Note that for fusion of just two distances with SFE, we should replace Equations 1 and 2 with

$$y_S = a_1\hat{x}_1 + a_2\hat{x}_2$$

and

$$a_1 = \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2} \qquad a_2 = \frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2}$$

respectively.

TABLE II
MULTIMODAL VERIFICATION PERFORMANCE WITH FLD — 1 ENTRY PER ENROLLMENT, 1 ENTRY PER VERIFICATION.

FLD Fusion	EER
Keystroke and Pitch	8.3%
MFCC and Pitch	9.6%
Keystroke and MFCC	8.4%
Pitch, MFCC and Keystroke	8.0%

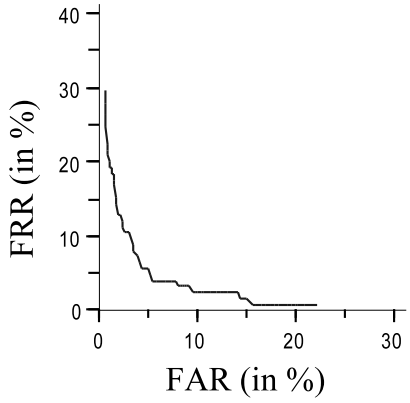

Fig. 8. DET curve from fusion of d_P, d_M and d_K, with SFE.

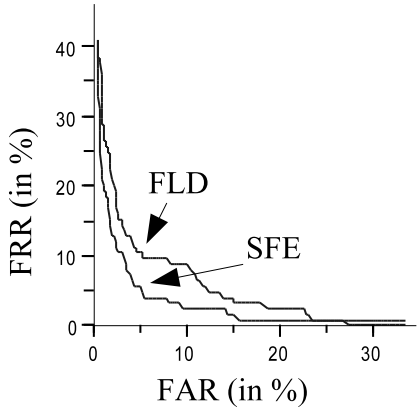

Fig. 9. DET curve from fusion of d_P, d_M and d_K, with FLD.

VI. DISCUSSION AND CONCLUSIONS

According to the results presented in Section V, SFE approach seems to outperform both FLD and BCND, what is quite a surprising result. Indeed, we can barely raise some hypothesis to explain the superiority of this simple linear fusion approach, but we believe that it is related to the "shape" (dispersion) of Class 2 (see, for instance, the dashed contour in Figure 10), whose average diameter is much greater then the length of a suitable classification boundary (between-classes interface).

According to this first attempt to explain SFE superiority, we believe that most data in Class 2 are far from any suitable decision boundary. Nevertheless, given its strong contribution to the covariance matrix of Class 2, both boundaries provided

TABLE III
MULTIMODAL VERIFICATION PERFORMANCE WITH SFE — 1 ENTRY PER ENROLLMENT, 1 ENTRY PER VERIFICATION.

SFE Fusion	EER
Keystroke and Pitch	7.9%
MFCC and Pitch	8.4%
Keystroke and MFCC	6.7%
Pitch, MFCC and Keystroke	5.0%

TABLE IV

MULTIMODAL VERIFICATION PERFORMANCE WITH BCND — 1 ENTRY PER ENROLLMENT, 1 ENTRY PER VERIFICATION.

BCND Fusion	EER
Keystroke and Pitch	10.5%
MFCC and Pitch	8.0%
Keystroke and MFCC	9.5%
Pitch, MFCC and Keystroke	5.8%

by FLD and BCND are deviated by them from what could be a better solution, in terms of EER.

In other words, even if the FLD does minimize the Rayleigh quotient [10], which is a quadratic criterion, it deviates from the minimization of the EER criterion.

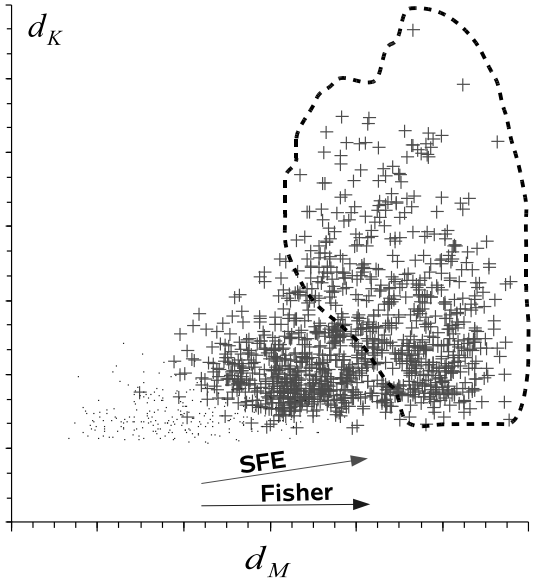

Fig. 10. Comparison between FLD and SFE projection vectors.

Fortunately, the SFE approach seems to be less sensitive to this bad influence of Class 2 dispersion and decision boundary, probably because of the centralization procedure, and also because it does not take into account covariance matrices of classes, as FLD and BCND do.

In fact, those hypothesis are to be tested in future works.

On the other hand, identity verification through fusion of features from keystroke dynamics and speech was addressed, and we experimentally studied the performance of identity verification algorithms based on fusion of median pitch (prosodic level information), mel-frequency cepstral coefficients (short-time spectral level information) and keystroke dynamics (down-down intervals).

Experimental results, from multimodal biometrics, were provided along with the corresponding unimodal based performances, in terms of EER.

All experiments were done with a small but true multimodal public database. Three fusion strategies are applied during experiments and in almost all of them, speech and keystroke fusion showed a non-negligible performance improvement.

Our experiments were realized with single samples as references (3s of speach and 31 keystrokes per subject prototype, and the same amount of data per test sample). Though it leads to quite a limited performance, possibly, it may be useful in some applications where fast enrollment is suitable and low security levels are acceptable.

Note, however, that performance can be improved by the use of more samples (or longer samples). Nonetheless, improvements seem to be statistically representative, even with results from such a small database.

On the other hand, in spite of the smallness of the databases, it is a true multimodal one, and to allow further comparisons between the results reported here and performances of other approaches, with the same database, multimodal samples used in this work are available to download at www.ufs.br/biochaves[1]. (Internet web site).

ACKNOWLEDGMENTS

This work was partially granted by both the *Fundação de Amparo à Pesquisa de Sergipe* (FAP-SE) and the *Conselho Nacional de Desenvolvimento Científico e Tecnológico* (CNPq). We also thank a lot all students and fellows whose keystrokes dynamics and recorded voices were used as samples in this work.

REFERENCES

[1] A. Ross, A.K. Jain, “Information Fusion in Biometrics”, Pattern Recognition Letters, Vol. 24, Issue 13, pp. 2115-2125, September, 2003.

[2] A. Ross, A.K. Jain, “Multimodal Biometrics: An Overview,” Proc. of 12th European Signal Processing Conference (EUSIPCO), (Vienna, Austria), pp. 1221-1224, September 2004.

[3] R. Brunelli, D. Falavigna, “Person Identification Using Multiple Cues,” IEEE Trans. on Pattern Analysis and Machine Intelligence, 17 (10) (1995) 955-966.

[4] K.-A. Toh, W.-Y. Yau, Fingerprint and Speaker Verification Decisions Fusion Using a Functional Link Network, IEEE Trans. on Systems, Man, and Cybernetics 35 (3) (2005) 357–369.

[5] N. Poh, S. Bengio, Database, Protocols and Tools for Evaluating Score-level Fusion Algorithms in Biometric Authentication, Pattern Recognition (Elsevier) 39 (2006) 223–233.

[6] M. Faundez-Zanuy, E. Monte-Moreno, “State-of-the-art in Speaker Recognition,” IEEE Aerospace and Electronic Systems Magazine, 20 (5) (2005) 7-12.

[7] S. Bleha, C. Slivinsky, B. Hussien, Computer-access security systems using keystroke dynamics, IEEE Trans. on Pattern Analysis and Machine Intelligence 12 (12) (1990) 1217–1222.

[8] J. Montalvão, E. O. Freire, On the Equalization of Keystroke Timing Histograms, Pattern Recognition Letters (to appear).

[9] N. Bergman, “Recursive bayesian estimation: Navigation and tracking applications,” Ph.D. dissertation, May 1999.

[10] R. Duda, P. Hart, Pattern Classification and Scene Analysis, Wiley-Interscience, New York, 1973.

[11] L.R. Rabiner, R.W. Schafer, Digital Processing of Speech Signals, Prentice-Hall, Englewood Cliffs, NJ, 1978.

[12] D. Gunetti, C. Picardi, Keystroke analysis of free text, ACM Trans. Inf. Syst. Secur.8 (3), (2005) 312–347.

[13] A. Gelb, J. Kasper, R. Nash, C. Price, and A. Sutherland, Applied Optimal Estimation, The M.I.T. Press, 1974.

[1] Website’s mirror at http://www.infonet.com.br/biochaves