Using Chimeric Users to Construct Fusion Classifiers in Biometric Authentication Tasks: An Investigation

Norman Poh [a] Samy Bengio [a]

IDIAP–RR 05-59

October 2005

submitted for publication

[a] IDIAP, CP 592, 1920 Martigny, Switzerland

IDIAP Research Report 05-59

Using Chimeric Users to Construct Fusion Classifiers in Biometric Authentication Tasks: An Investigation

Norman Poh Samy Bengio

October 2005

submitted for publication

Abstract. Chimeric users have recently been proposed in the field of biometric person authentication as a way to overcome the problem of lack of real multimodal biometric databases as well as an important privacy issue – the fact that too many biometric modalities of a same person stored in a single location can present a *higher* risk of identity theft. While the privacy problem is indeed solved using chimeric users, it is still an open question of how such chimeric database can be efficiently used. For instance, the following two questions arise: i) Is the performance measured on a chimeric database a good predictor of that measured on a real-user database?, and, ii) can a chimeric database be exploited to *improve* the generalization performance of a fusion operator on a real-user database?. Based on a considerable amount of empirical biometric person authentication experiments (21 real-user data sets and up to 21 × 1000 chimeric data sets and two fusion operators), our previous study [16] answers **no** to the first question. The current study aims to answer the second question. Having tested on four classifiers and as many as 3380 face and speech bimodal fusion tasks (over 4 different protocols) on the BANCA database and four different fusion operators, this study shows that generating multiple chimeric databases *does not degrade nor improve* the performance of a fusion operator when tested on a real-user database with respect to using only a real-user database. Considering the possibly expensive cost involved in collecting the real-user multimodal data, our proposed approach is thus *useful* to construct a trainable fusion classifier while at the same time being able to overcome the problem of small size training data.

1 Introduction

Biometric authentication is a problem of verifying an identity claim using a person's behavioral and physiological characteristics. While this can be achieved based on a single modality (voice or face prints for instance), the current literature provides several approaches towards studying fusion of such modalities for better performance and robustness. One practice is to construct a large database containing several biometric traits for each user. This, however, can be very time-consuming, expensive, and of ethical concern. Another practice is to combine biometric modalities of a database with biometric modalities of another biometric database. Since both databases do not necessarily contain the *same* users, such combination results in *chimeric users*. From the experimental point of view, these biometric modalities belong to the same person. While this practice is commonly used in the multimodal literature, e.g., [17, 7] among others, it was questioned whether this was a right thing to do or not during the 2003 Workshop on Multimodal User Authentication [5].

There are at least two arguments that justify the use of chimeric users, i.e., i) *modality independence assumption* – that two or more biometric traits of a single person are often assumed independent of each other; and ii) *privacy issue* – participants in the multimodal biometric experiments are often not ready to let institutes keep record of too much of their personal information (raw biometric data) at the same place. If such information is misused, it could be dangerous, e.g., identity theft. It is for this same reason that processed biometric features are preferred for storage to raw biometric data. Note that the first argument is *technical* while the second one is *ethical*. Although both arguments are equally important, the second one is beyond an experimenter's control and is related to the usage policy of the database. For instance the policy should address who can use the database and how it should be used. When a database is carefully designed to protect the participants' privacy right, this issue should be resolved. For this reason, this paper focuses on the first argument.

In our previous study [16], we addressed the question: "Is the performance measured on a chimeric database a good estimator of that measured on a real-user database?". Having conducted a considerable amount of empirical experiments (21 real-user data sets and up to 21 × 1000 chimeric data sets on two fusion operators, the answer is no. In other words, the performance based on a chimeric database can *possibly be biased*. This means that, for instance, one cannot claim that novel algorithm A is better than state-of-the-art algorithm B on a real multimodal biometric authentication task if the comparison was conducted on a chimeric database. A similar investigation was reported in [9] with the conclusion that favors the use of chimeric users. It should be pointed out that these studies were undertaken with the following differences: (i) the former was tested on 21 fusion tasks whereas the latter was tested on two fusion tasks (clean and noisy); (ii) the former engaged in a standard hypothesis test whereas the latter did not – only the mean DET curves derived from both real and chimeric databases were visually compared; (iii) the former is based on a threshold dependent assessment – whereby a threshold is optimized *a priori* on a development (training) set and a performance is measured on an evaluation (test) set using the chosen threshold; a similar assessment as the yearly NIST evaluation protocols [12]) – whereas the latter is based on a threshold free assessment via a DET curve; (iv) two fusion operators are considered in the former and only one considered in the latter; and (v) a bootstrap procedure was used in the latter to estimate the distribution of performance on the real-user database and the former did not[1]. While most differences are methodological, it should be remarked that our findings show that only approximately a third of 21 fusion data sets, independent of the fusion classifier used, reports inconsistency of performance between the real-user and chimeric databases. Hence, the inconsistency may very well not be visible with only 2 experiments.

This paper addresses another issue with respect to chimeric users: "Can a chimeric database be exploited to *improve* the generalization performance of a fusion operator on a real-user database?". Very often, due to lack of training data, a fusion operator has very limited amount of data for training. Hence, by using chimeric database, one can generate much more data to train the fusion classifier that would then be assessed on real multimodal user scores. If this is the case, then, even if the performance

[1] Recognizing that this issue is important, our on-going work takes into account of such information but the experimental outcome does not change the conclusion reported in [16].

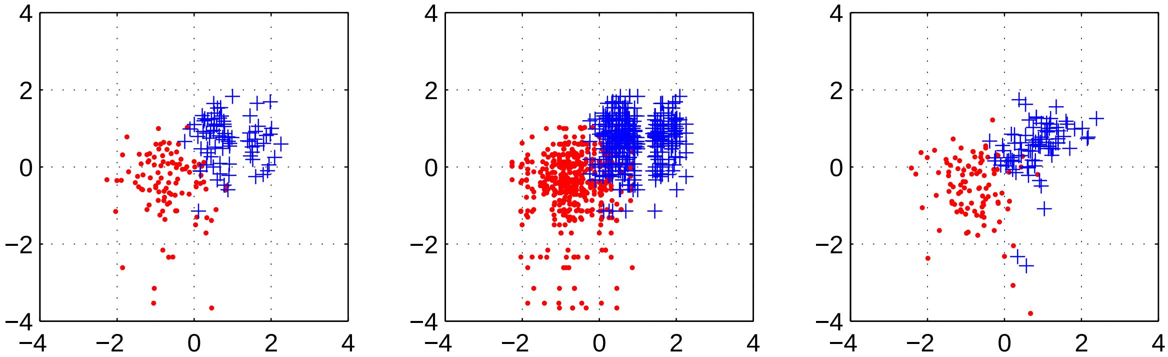


Figure 1: Left: an original bimodal fusion training data set whose x-axis is a speech expert score and y-axis is a face expert score. Center: a bimodal fusion training data set generated using chimeric users. Right: a bimodal real-user test data set. This is a typical example among the 3380 fusion tasks taken from the BANCA database. In each figure, crosses (upper right cluster) denote client accesses and dots (lower left cluster) denote impostor accesses.

measured on a chimeric database is biased as in [16], a chimeric database is still *at least* useful for other purposes such as to help construct a fusion classifier. To verify this hypothesis, we limit our scope to studying such effect to bimodal as generalization to more than two modalities is direct.

This paper is organized as follows: Section 2 contains a description on the general methodology used; Section 3 describes the BANCA database used; Section 4 presents the four fusion classifiers used; and Section 5 presents the experimental outcomes.

2 Methodology

To illustrate the idea, we first plot a bimodal fusion training and test sets in the left and right panels of Figure 1, respectively. By random mix-and-match of (scores of) modalities according to different identities, we obtained a *much larger* training set as shown in the middle panel of Figure 1.

Although this methodology is rather simple, there is still a fundamental question of how many chimeric users are necessary. Suppose that there are N real users for which we recorded 2 modalities. Then, in theory, in order to construct a bimodal chimeric database, one can generate up to $N \times (N-1)$ chimeric users (by excluding the N real users). Our initial experiments with $N, 2N, 3N, \ldots, (N-1) \times N$ (as a multiple of the user size) show that the number of users has not much effect on the performance. We thus fixed this multiple factor to 10 so that the fusion constructed on chimeric users had 10 times more data than that trained on real users.

3 Database

We used the real bimodal face and speech BANCA database. Some of the scores were obtained from [11][2] while the rest of the data, based on face systems, are taken from [6]. These systems contain experimental as well as the state-of-the-art systems based on Principal Component Analysis, Linear Discriminant Analysis, Gaussian Mixture Models, Hidden Markov Models, Support Vector Machines, Normalized Correlation, etc, to name a few. In the BANCA database, there are 7 different protocols, of which we chose four: Mc, Ua, Ud and P. The first three represent matched controlled, unmatched adversed, unmatched degraded scenarios, respectively. The last one is a pooled scenario containing the first three. A *matched* scenario implies that the mismatch between a training and a test set is

[2] Available at “ftp://ftp.idiap.ch/pub/bengio/banca/banca_scores”

minimal (due to using the same type of microphone, video camera and data acquired in similar and clean conditions). There are two *unmatched* scenarios: adversed and degraded. The former refers to the mismatch due to different acquisition environment whereas the latter refers to using a degraded acquisition device (by simulation). There are five language subsets but only the English subset is used in this study. By combining a speech-based biometric system with a face-based biometric system, the first three protocols contain 840 fusion tasks whereas the last one contains 860. In the BANCA protocols, two groups of users are distinguished and are labeled by g1 and g2. We used g1 as a development (training) set and g2 as an evaluation (test) set; hence, while g1 was modified to create chimeric users, g2 was kept with real users only, in order to be able to assess performance on real users.

4 Fusion Classifiers and Threshold Estimation

Four classifiers are used, namely Logistic Regression (LR) [14], Gaussian Mixture Model (GMM) with dependent assumption [3], GMM with independent assumption and the mean operator. Note that the LR classifier used here is more general than the one used in [14] (which assumes common covariance of both client and impostor distributions) but rather the *standard* approach as described in [10]. Let $\boldsymbol{y} \equiv [y_1, \ldots, y_M]^T$ be a vector of scores consisting of M biometric modalities. The LR classifier has the following form:

$$y_{LR} \equiv P(C|\boldsymbol{y}) = \frac{1}{1+\exp(-g(\boldsymbol{y}))},$$

where

$$g(\boldsymbol{y}) = \sum_{i=1}^{M} \beta_i y_i + \beta_0.$$

We used an implementation described in [4]. The classical approach of using GMM in classification [3, Chap. 2] is to establish a Log-likelihood ratio (LLR) test between the client and impostor classes, i.e., $k = \{C, I\}$. The LLR takes the following forms:

$$y_{dep} \equiv \log \frac{p(\boldsymbol{y}|C)}{p(\boldsymbol{y}|I)}, \tag{1}$$

for the dependent assumption and

$$y_{indep} = \log \frac{\prod_i p(y_i|C)}{\prod_i p(y_i|I)}, \tag{2}$$

for the independent assumption. The approximations to Eqn. (1) and Eqn. (2) using GMM can be written as follow:

$$\hat{p}(\boldsymbol{y}|k) = \sum_{c}^{N_c} w_c^k \mathcal{N}(\boldsymbol{y}|\boldsymbol{\mu}_c^k, \boldsymbol{\Sigma}_c^k), \tag{3}$$

$$\hat{p}(y|k) = \sum_{c}^{N_c} w_c^k \mathcal{N}\left(y|\mu_c^k, (\sigma_c^k)^2\right), \tag{4}$$

for any $y \in \{y_i | i = 1, \ldots, M\}$, respectively, where, the c-th component of the class conditional (denoted by k) mean vector is $\boldsymbol{\mu}^k = [\mu_1^k, \ldots, \mu_M^k]^T$ and its covariance matrix of dimension $M \times M$ is $\boldsymbol{\Sigma}_c^k$. The mean and variance of $p(y|k)$ are defined similarly except that it is single dimensional. The GMM parameters can be optimized using the Expectation-Maximization algorithm [3] for instance and the number of components can be tuned by validation or optimization of a criterion, e.g., minimum description length [8]. Finally, the fused score using the mean operator has the following form:

$$y_{mean} = \frac{1}{M} \sum_{i=1}^{M} \frac{y_i - B_i}{A_i},$$

where B_i and A_i are called a bias and a scaling factor, respectively. In our implementation, both parameters are estimates of mean and standard deviation from the training scores, respectively. The resultant normalized y_i score is sometimes called a z-score.

Note that the above fusion classifiers do not include a threshold[3]. The complete model has the following decision function:

$$\text{decision}(y) = \begin{cases} \text{accept} & \text{if } y > \Delta \\ \text{reject} & \text{otherwise,} \end{cases} \tag{5}$$

where Δ is a global decision threshold and y in our context is any of the combined scores $y \in \{y_{LR}, y_{dep}, y_{indep}, y_{mean}\}$ discussed before. In a *threshold-dependent* assessment based on Expected Performance Curve [2], the Δ is chosen to minimize, on a separate development set, the following criterion, known as Weighted Error Rate (WER),

$$\Delta_* = \arg\min_{\Delta} \text{WER}_\alpha(\Delta) \tag{6}$$

where

$$\text{WER}_\alpha(\Delta) \equiv \alpha\text{FAR}(\Delta) + (1-\alpha)\text{FRR}(\Delta),$$

and α ranges from 0 to 1. This parameter balances between the *costs* between FAR and FRR estimated from a *development* set. Note that although not having the exact same formulation, similar criteria were employed in the yearly NIST speaker evaluation plans [12] and the BANCA protocols [1]. Using this threshold, we can then evaluate WER for several values of α on the *evaluation* set. This enables us to obtain unbiased estimates of performance since all hyper-parameters of the fusion operator, *including the threshold*, are selected on the development or a separate validation set. Note that only *a priori* performances are reported here.

5 Experimental Results

Figure 2 shows pooled EPC curves of four fusion classifiers trained on a real-user development set and an *augmented* chimeric-user development set having 10 times more data than the former development set. This gives $4 \times 2 = 8$ modes of fusion. The total statistics to be analyzed can be summarized by $WER_{pooled}(\alpha, p, COM, data)$ for

- the performance cost $\alpha \in [0, 1]$,
- on the protocol $p = \{\text{Mc, Ua, Ud, P}\}$,
- using any fusion operator $COM \in \{LR, dep, indep, mean\}$ and
- trained on the $data \in \{real, chim\}$ (real or chimeric).

One can see from Figure 2 that in most cases, the generalization performance on real users was similar whether we used real users for training (thin continuous lines) or chimeric users (thick dashed lines).

We then pooled all measures coming from different fusion operators in order to compare the relative performance between chimeric-based fusion models and real user-based fusion models. This relative performance is calculated as

$$(WER_{chim} - WER_{real})/WER_{real}.$$

Hence, a negative change implies that the fusion operator derived from chimeric users improves over its real-user counterpart.

[3] The LR classifier has a bias but it is not used since the algorithm does not explicitly optimize Equal Error Rate or any authentication-related performance.

The results, shown in Figure 3, suggest that the generalization performance using chimeric users is not very different from the one using real users (the average relative change is near 0), across different fusion operators. However, for protocol Mc (clean conditions) and Ud (degraded conditions), the fusion operator trained on chimeric users has a higher chance (never less than 50% and 38% of the time, respectively) of being *consistently* better than its real-user counterparts across all cost of α values. Finally, a mixed performance is observed for protocols Ua (adversed) and P (pooled over all three scenarios). In summary, a chimeric database can have a higher chance of improving generalization performance of a fusion operator over not using such information, especially under matched (clean) conditions. It does not make a fusion operator more robust because it is not designed to do so – suggesting that other prior knowledge such as quality information is necessary.

6 Conclusions

Although the use of virtual users is somewhat novel, it should be mentioned that training using *virtual samples* in machine learning is not new, e.g., [13, 15]. However, different from them, this paper explores *how* a model can be built using a chimeric database, an approach which to the best of our knowledge, has not been investigated before. One important conclusion from this preliminary study is that a fusion operator derived from a chimeric-user database does not improve nor degrade the generalization performance (on real users) with respect to training it on real users. The advantage, however, is that much more training data can be artificially generated thus in this way it can overcome the lack of training data. This is especially useful when using trainable fusion operators. Note however that, as explained in [16], while chimeric data can be useful to train good fusion operators, the obtained fusion models can only be evaluated on real multimodal biometric data, and not on chimeric data.

Acknowledgement

The authors thank the following people who helped brainstorming the subject: Johnny Mariéthoz, and many participants of MLMI'05 and IDIAP's TAM. This work was supported in part by the IST Program of the European Community, under the PASCAL Network of Excellence, IST-2002-506778, funded in part by the Swiss Federal Office for Education and Science (OFES) and the Swiss NSF through the NCCR on IM2. This publication only reflects the authors' view.

References

[1] E. Bailly-Baillière, S. Bengio, F. Bimbot, M. Hamouz, J. Kittler, J. Mariéthoz, J. Matas, K. Messer, V. Popovici, F. Porée, B. Ruiz, and J.-P. Thiran. The BANCA Database and Evaluation Protocol. In *LNCS 2688, 4th Int. Conf. Audio- and Video-Based Biometric Person Authentication, AVBPA'03*. Springer-Verlag, 2003.

[2] S. Bengio and J. Mariéthoz. The Expected Performance Curve: a New Assessment Measure for Person Authentication. In *The Speaker and Language Recognition Workshop (Odyssey)*, pages 279–284, Toledo, 2004.

[3] C. Bishop. *Neural Networks for Pattern Recognition.* Oxford University Press, 1999.

[4] A. J. Dobson. *An Introduction to Generalized Linear Models.* CRC Press, 1990.

[5] J-L. Dugelay, J-C. Junqua, K. Rose, and M. Turk. *Workshop on Multimodal User Authentication (MMUA 2003).* no publisher, Santa Barbara, CA, 11–12 December, 2003.

[6] Fabien Cardinaux, Conrad Sanderson, and Samy Bengio. User Authentication via Adapted Statistical Models of Face Images. IDIAP Research Report 38, IDIAP, 2004. accepted for publication in *IEEE Trans. Signal Processing*, 2005.

[7] J. Fierrez-Aguilar, J. Ortega-Garcia, J. Gonzalez-Rodriguez, and J. Bigun. Kernel-Based Multimodal Biometric Verification Using Quality Signals. In *Defense and Security Symposium, Workshop on Biometric Technology for Human Identification, Proc. of SPIE*, volume 5404, pages 544–554, 2004.

[8] M.A.T. Figueiredo and A.K. Jain. Unsupervised learning on finite mixture models. *Pattern Analysis and Machine Intelligence*, 24(3), March 2002.

[9] S. Garcia-Salicetti, M. A. Mellakh, L. Allano, and B. Dorizzi. A Generic Protocol for Multibiometric Systems Evaluation on Virtual and Real Subjects. In *LNCS 3546, 5th Int'l. Conf. Audio- and Video-Based Biometric Person Authentication (AVBPA'05)*, pages 494–502, New York, 2005.

[10] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning.* Springer-Verlag, 2001.

[11] Christine Marcel. Multimodal Identity Verification at IDIAP. Communication Report 03-04, IDIAP, Martigny, Switzerland, 2003.

[12] A. Martin. NIST Year 2001 Speaker Recognition Evaluation Plan, 2001.

[13] P. Niyogi, F. Girosi, and T. Poggio. Incorporating Prior Information in Machine Learning by Creating Virtual Examples, 1998.

[14] S. Pigeon, P. Druyts, and P. Verlinde. Applying Logistic Regression to the Fusion of the NIST'99 1-Speaker Submissions. *Digital Signal Processing*, 10(1–3):237–248, 2000.

[15] N. Poh, S. Marcel, and S. Bengio. Improving Face Authetication Using Virtual Samples. In *IEEE Int'l Conf. Acoustics, Speech, and Signal Processing*, pages 233–236 (Vol. 3), Hong Kong, 2003.

[16] Norman Poh and Samy Bengio. Can Chimeric Persons Be Used in Multimodal Biometric Authentication Experiments? Research Report 20, IDIAP, Martigny, Switzerland, 2005. To appear in *MLMI 2005.*

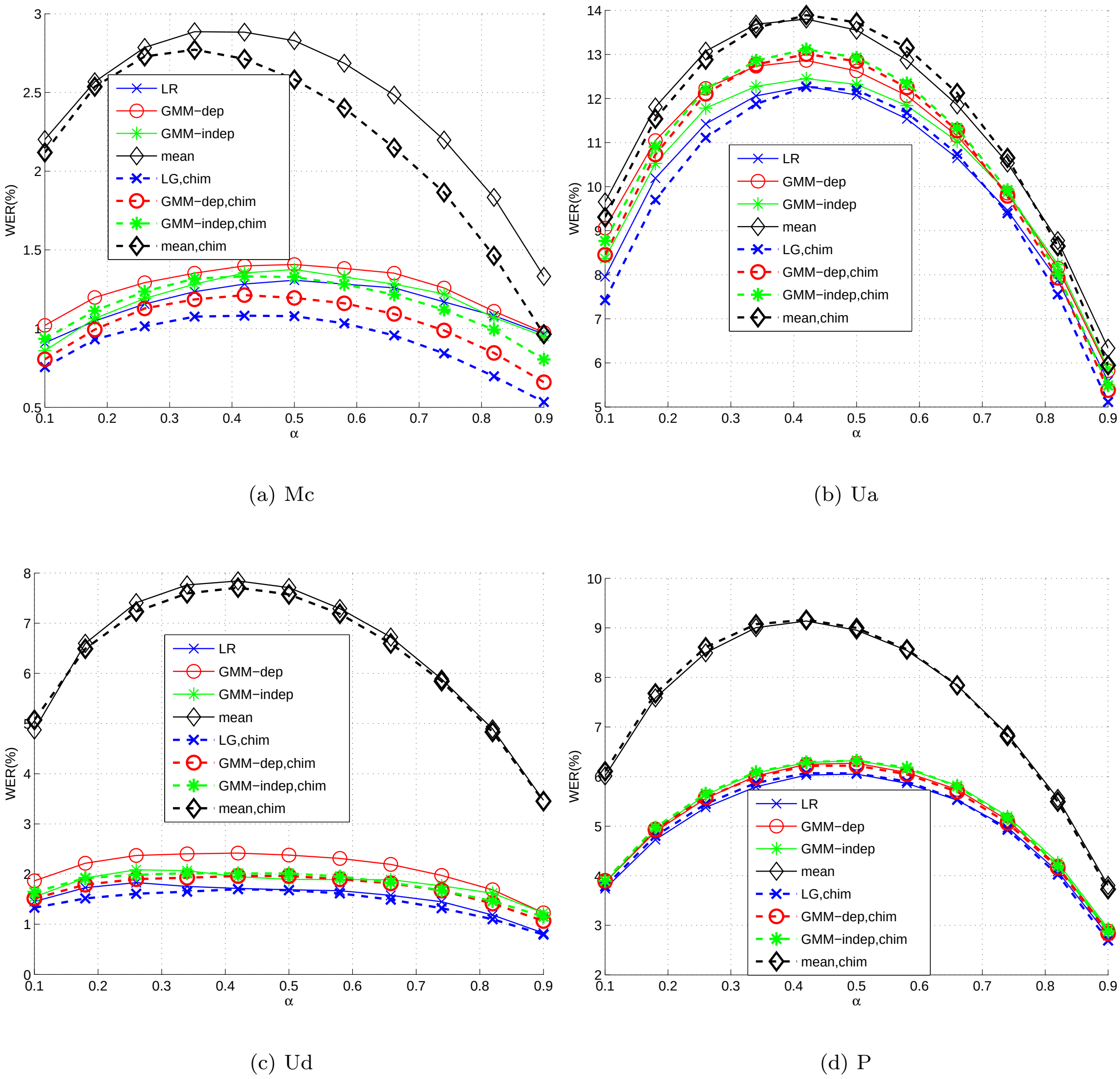
[17] A. Ross, A. Jain, and J-Z. Qian. Information Fusion in Biometrics. *Pattern Recognition Letter*, 24(13):2115–2125, September 2003.

A Additional Evaluations

This section contains additional experiments whereby, instead of just using *a priori* WER, *a priori* Half Total Error Rate (HTER) is also used. It is defined as $\frac{1}{2}\big(\mathrm{FAR} + \mathrm{FRR}\big)$. The figures are summarized in Table 1. HTER versus α are plotted to be consistent with other reports based on EPC. Although both performance measures are different, the conclusion presented in Section 6 does not change.

Table 1: Summary of figures

Statistics	Protocols			
	Mc	Ua	Ud	P
WER	Fig. 2 (a)	Fig. 2 (b)	Fig. 2 (c)	Fig. 2 (d)
Relative change of WER	Fig. 3			
HTER	Fig. 4 (a)	Fig. 4 (b)	Fig. 4 (c)	Fig. 4 (d)
Relative change of HTER	Fig. 5			
Relative change of (a) HTER and (b) WER	Fig. 6	Fig. 7	Fig. 8	Fig. 9



(a) Mc

(b) Ua

(c) Ud

(d) P

Figure 2: WER versus α (the lower the better) on (a) Mc (840 fusion sets ×4 fusion operators), (b) Ua (840 × 4), (c) Ud (840 × 4) and (d) P (860 × 4) protocols. The four fusion operators are: logistic regression (cross), GMM with dependent assumption (circle), GMM with independent assumption (asterisk) and the mean operator (diamond). Comparison should be made between a *thin continuous* line and a *thick dashed* line.

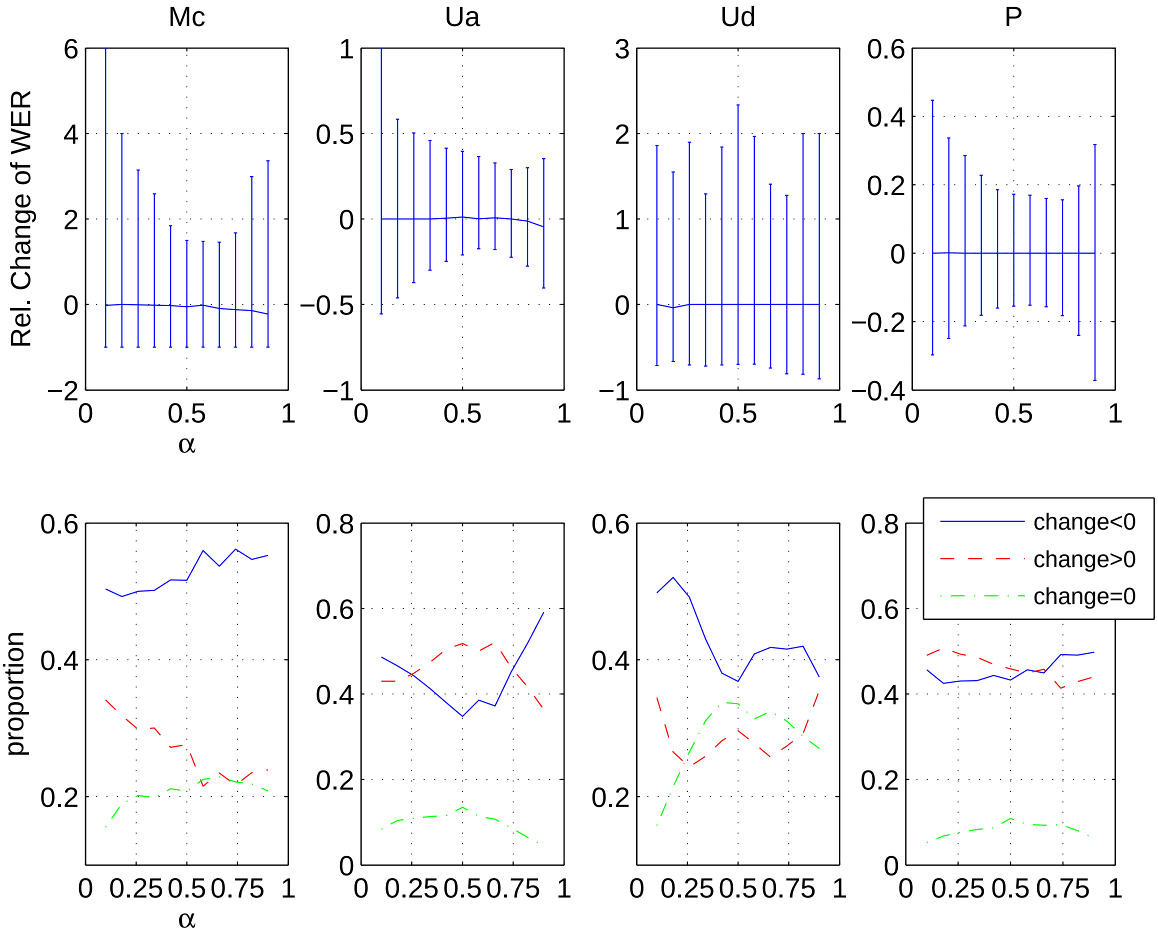
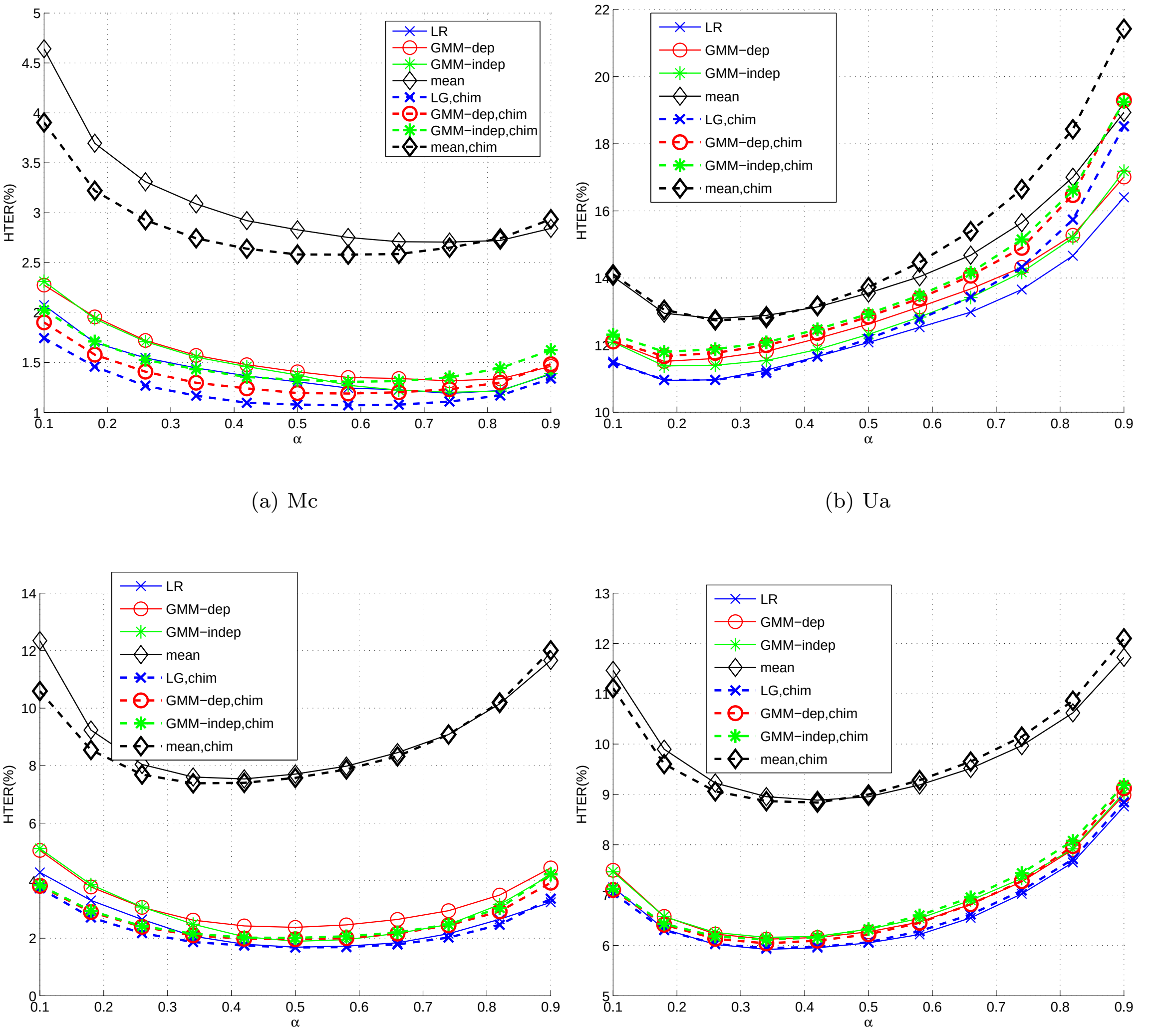


Figure 3: Upper rows: Relative change of *pooled* WER on Mc, Ua, Ud and P protocols depicted as error bars. Each bar indicates the 2.5th and the 97.5th percentiles and is linked to each other via their respective median. The corresponding lower rows show the proportion of change < 0 (favors the operators due to chimeric users), > 0 (favors that due to real users) or $= 0$ (i.e., both give *exactly* the same value).



(a) Mc

(b) Ua

(c) Ud

(d) P

Figure 4: As per Figure 2, except that the HTER criterion is used instead of WER. The four fusion operators are: logistic regression (cross), GMM with dependent assumption (circle), GMM with independent assumption (asterisk) and the mean operator (diamond). Comparison should be made between a *thin continuous* line and a *thick dashed* line.

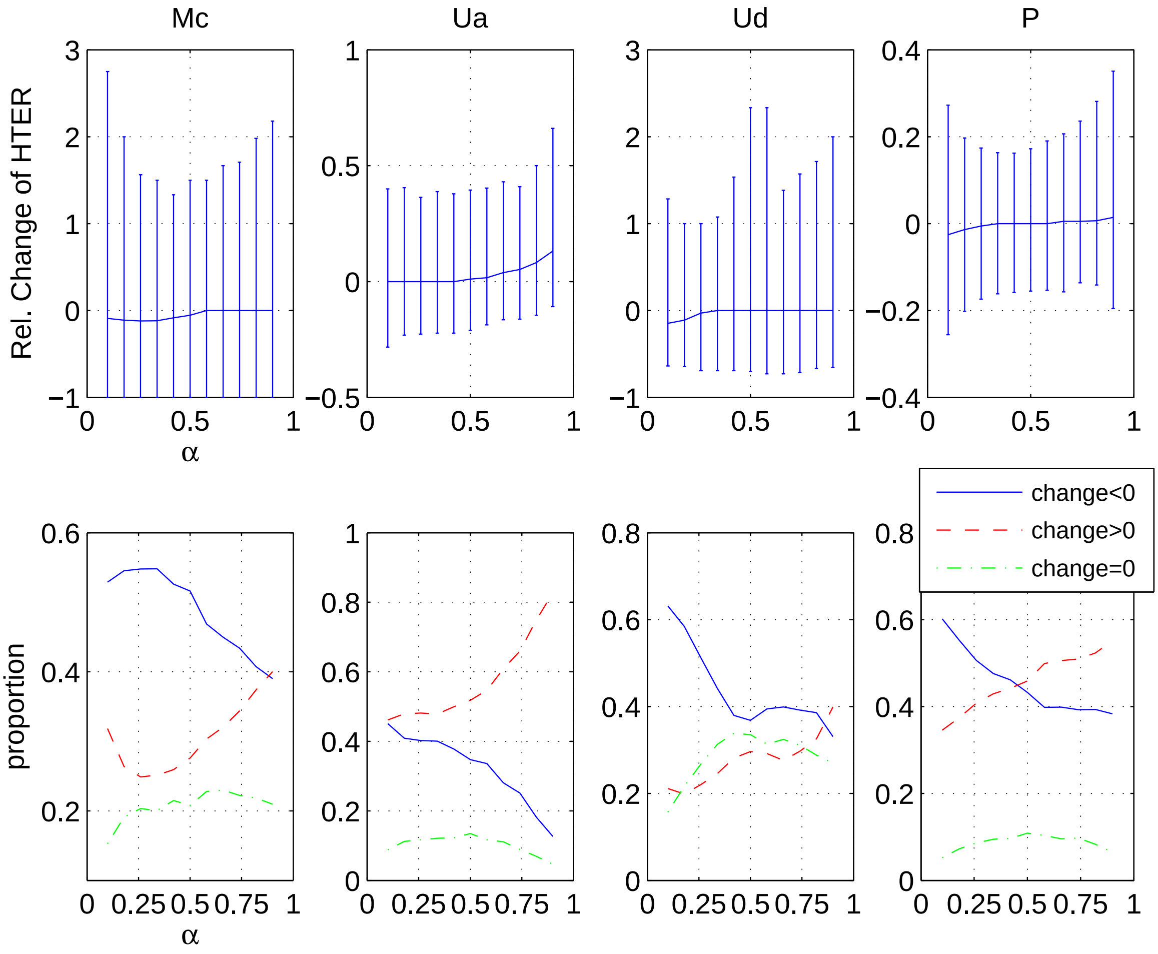
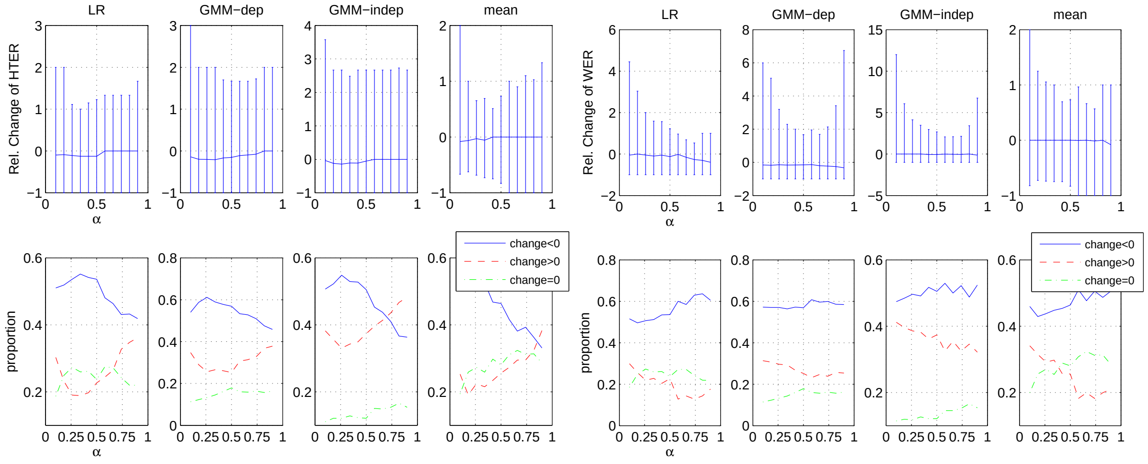


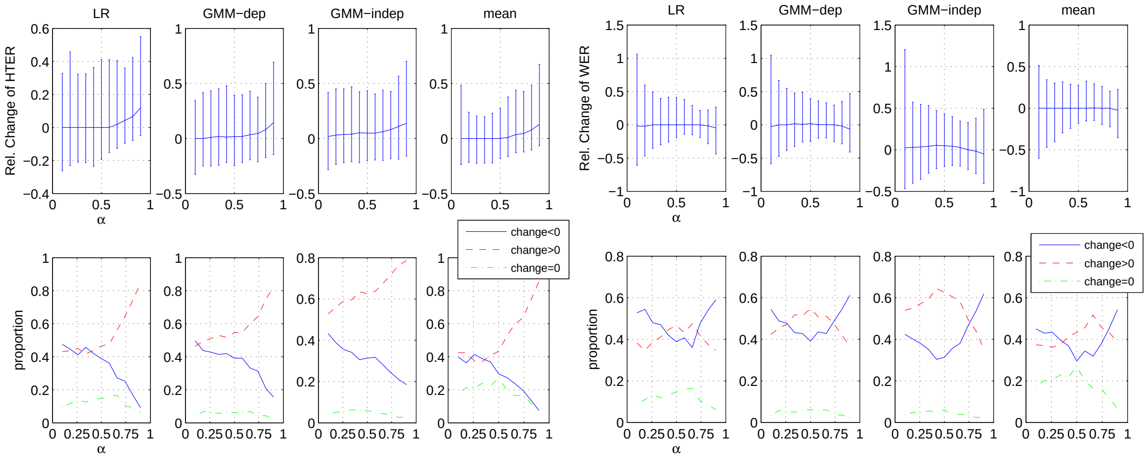
Figure 5: As per Figure 5, except that the relative performance is calculated based on HTER instead of WER. Upper rows: Relative change of *pooled* WER on Mc, Ua, Ud and P protocols depicted as error bars. Each bar indicates the 2.5th and the 97.5th percentiles and is linked to each other via their respective median. The corresponding lower rows show the proportion of change < 0 (favors the operators due to chimeric users), > 0 (favors that due to real users) or $= 0$ (i.e., both give *exactly* the same value).



(a) HTER

(b) WER

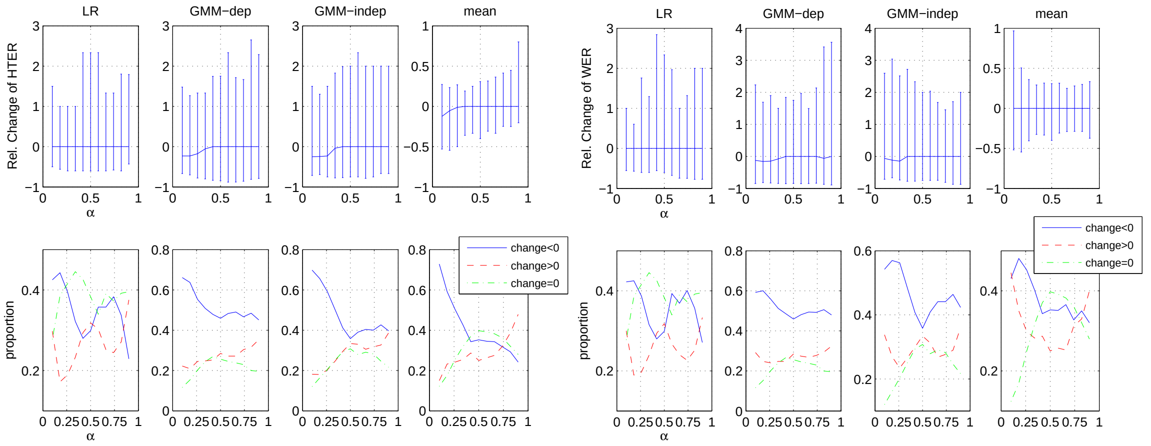
Figure 6: Relative change of (a) HTER and (b) WER and their corresponding proportions, summarized over 840 fusion sets on the protocol Mc. The four fusion operators are: logistic regression, GMM with dependent assumption, GMM with independent assumption and the mean operator (after applying z-score normalization). "Change" calculates the number of proportion is negative (thus favoring the use of chimeric user to compute the fusion model), positive (the opposite) or zero (no effect).



(a) HTER

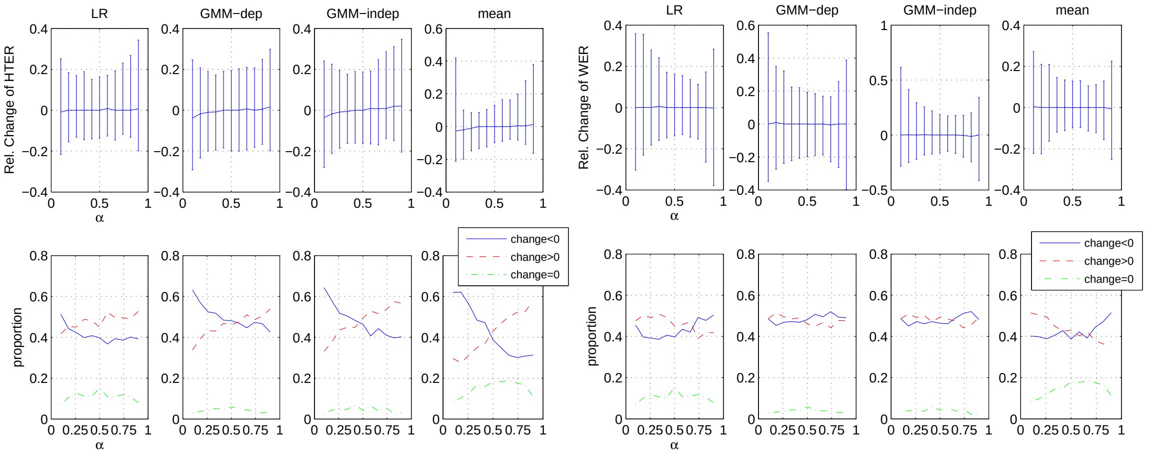
(b) WER

Figure 7: As per Figure 6, except that the protocol Ua is used. These figures are summarized over 840 experiments.



(a) HTER (b) WER

Figure 8: As per Figure 6, except that the protocol Ud is used. These figures are summarized over 840 experiments.



(a) HTER (b) WER

Figure 9: As per Figure 6, except that the protocol P is used. Different from the rest, these figures are summarized over 860 experiments instead of 840.