Agatha: Multimodal Biometric Authentication Platform in Large-Scale Databases

David Hernando[1] · David Gómez[1] · Javier Rodríguez Saeta[1]
Pascual Ejarque[2] · Javier Hernando[2]

[1]Biometric Technologies S.L., Barcelona, Spain
{d.hernando | d.gomez | j.rodriguez}@biometco.com

[2]TALP Research Center
Universitat Politècnica de Catalunya, Spain
{pascual | javier}@gps.tsc.upc.edu

Abstract

Biometric technologies are each time more demanded for security applications. In this sense, systems for identifying people are gaining popularity, especially in governmental sectors, and forensic applications have climbed to the top of the list when talking about biometrics. However, some problems still remain as cornerstones in identification processes, all of them linked to the length of the databases in which the individual is supposed to be. The speed and the error are parameters that depend on the number of users in the database and measure the quality of the whole system.

In this paper, two different biometric technologies are used in order to increase speed and shorten error rates. Face recognition –normally faster than speaker recognition – is used to select a group of individuals and speaker recognition provides a finer adjustment. Multimodality plays an important role not only reducing the search time but also providing lower error rates.

1 Introduction

The number of biometric applications has increased a lot in the last few years, especially from the 11th of September of 2001. Concerns about security have been raising and each time more, biometric systems are playing an important role in order to protect networks or buildings. The automatic person recognition by some physical traits like fingerprints, face, voice or iris, has a very high demand around the world and the technology is already mature.

Speaker and face recognition technologies have increased their popularity in a market dominated by fingerprint technologies. Speaker recognition is not the most used technology but it is expected that will be much more important in the future, especially for voice portals in the Internet. In the case of face recognition, its use is growing every day because of low intrusiveness and the facility of capturing images.

There are other kind of biometric applications in where speaker recognition has experienced a high increase: forensic applications. In this type of applications, speaker recognition is used to prove if the evidence belongs to the suspect or not. Forensic speaker recognition is also used to identify speakers

N. Pohlmann, H. Reimer, W. Schneider (Editors): Securing Electronic Business Processes, Vieweg (2008), 1-8

looking for a certain voice in a database of suspects. Police forces in lots of European countries and also the FBI in the USA have speaker databases for this purpose.

Main characteristics –common to this type of applications- are the great amount of recorded data, because recordings are usually acquired without the consciousness of the person being recorded; the text-independence, although there is not any dependent text but also natural speaking; and finally the fact that data usually comes from the telephone lines because we are dealing basically with recorded conversations.

With regard to face recognition, its forensic view is very clear. The identification of suspects in a public place like an airport or a train station is an application every time more demanded. Police forces around the world have databases with photographs of criminals. When they get an image of an individual, they wish to identify as soon as possible if this person is in the database. They want to know who the unknown suspect is.

The aim of this paper is the implementation of an identification platform by means of speaker recognition, face recognition or the combination of both of them (multimodality). The system will be used in criminalistic or security environments. The present project intends to provide a solution to the problem of identifying an individual in large-scale databases from biometric characteristics by taking them individually or by using a combination of voice and face. The system will return the N most probable users ordered by probability. It is not necessary to make the identification in real time but it is important that results will be provided in a reasonable time depending on the technology available.

Speaker identification is a very complex task that it does not normally occur very fast with large databases. On the other hand, face identification can be much faster, although error identification rates strongly depend on the way of acquiring images.

This project wants to create a tool for speaker recognition, a technology in where much more investigation can be done, and face recognition to be used together. Both biometrics will be fused in the case that we have both kinds of data from the user. The high identification speed of face identification can be a perfect complementary technology to increase speed in speaker identification. Face identification system can provide, in an initial identification, the N most probable candidates. In this case, the combination of both biometric technologies is not only used to improve error rates but also to increase speed in speaker recognition, a biometric technology in where traditionally it had been impossible to identify speakers in real time.

Finally, it is worth noting that one of the main interesting points of the platform is that it gathers in a unique system both types of identification at the same time, giving the possibility of using one of them individually, speaker or face recognition, or in parallel.

2 Search strategies

The speed of biometric identification algorithms can be a big issue for large population applications which require a short delay. If no search strategy is used, a full search approach entails a linear increase of the identification delay with the number of clients registered in the system. Therefore, the goal of search strategies is to achieve reasonable identification delays for the target application while maintaining the system performance, with a minor degradation. In this section, we review some approaches that have been proposed in the literature and then we introduce a method to reduce the identification delay in a multimodal framework.

Based on a recognition algorithm using HMMs (typically for text-dependent recognition) or GMMs (text-independent recognition), usually adapted to an individual speaker using MAP from a reference universal background model (UBM) [Reynolds95], some methods have been proposed in speaker recognition to speed-up the identification process and to reduce the computational cost.

A simple strategy reported in [McLaughin99] studies the system degradation just by reducing the number of components in the speaker model as well as decimating the test sample. For instance, reducing from 2048 to 512 components leads to less than a 1% loss in EER. Regarding the decimation, the paper shows how discarding 90% of frames, i.e., with a decimation factor of 10, the EER only increases by 1%.

In [Reynolds95], for each speech frame, only the mixtures with the highest scores against the UBM are used to match the test feature vector with each speaker model in the identification process. Other methods build a hierarchical set of speaker models. In [Beigi99], the GMM models are merged in pairs in an iterative way, building a tree structure with two models on the top. Similarly, the ISODATA clustering algorithm is applied in [Sun03] to this task achieving speed-up factors from 3:1 to 6:1 with almost no degradation respect to the full-search strategy.

Another approaches [Auckenthaler01] compute a hash model from a large GMM model which consists of a reduced number of mixtures indexing a list of the best expected scoring components in the large model. For instance, given a model of 512 mixtures, a hash model of 32 mixtures indexing at least 16 components of the large model results in the scoring of just 48 mixtures per frame instead of the original 512. With a minor degradation, [Aunckenthaler01] reports a speed-up factor of 10:1.

A speaker pruning can be done with the model proposed in [Pellom98]. The input sequence is processed as usual but a reduced selection of nonadjacent frames is first scored against the speaker models. Speakers with lower scores are discarded before repeating the selection with a higher number of frames and updating the accumulated probability of each speaker model. This process is repeated until no speakers are pruned out or the complete input signal is evaluated. The authors presented a time reduction by a factor of 140 over the full search with this method.

In [Kinnunen06] an extensive summary of speed-up approaches is presented and some of them are applied to a vector quantization (VQ) based speaker identification system. In this work, the input frames are pre-quantized in order to reduce the number of feature vectors used to score the input signal against the set of speakers. Four different pre-quantization techniques are used: random subsampling, averaging, decimation and clustering. Together with a speaker pruning method they achieve a speed-up factor of 16:1 with minor degradation in the identification performance.

In general, face identification is faster than voice. The algorithm used in this work reaches delays as low as 0.5 seconds for a two hundred clients database. However, the speaker identification process with a full search needs about 19 seconds for an average speech signal duration of 6.7 seconds.

A multimodal identification platform that combines speech and face can exploit the high speed of face recognition to make a search in a reduced set of speakers. We explore in this work a simple method to reduce the total identification delay by a factor of 8:1 with a slight increase in error identification rates.

Our approach starts with a full search with the face recognition system followed by a selection of the N clients with the highest confidence scores. Then, with the speaker recognition system we search only among the reduced set of N clients. Finally, we fuse both modalities results and if the highest score is over a previously estimated threshold we determine that we have a positive identification.

We will discuss in section 4 how this method gives a higher weight to the modality where the first pruning is done. In this case, it is the face algorithm. Therefore, for small values of N, the performance of each algorithm can lead to an increase or to a decrease of the error rates.

3 Multimodal fusion

A multimodal biometric system involves the combination of two or more human characteristics (voice, face, fingerprints, iris, hand geometry, etc.) in order to achieve better results than using unimodal recognition systems [Bolle04]. Furthermore, the use of several biometrics makes the system more robust to noise or spoof attacks.

When several biometric traits are used in a multimodal recognition system, fusion is usually accomplished at three different levels: feature extraction level, matching score level or decision level.

Fusion at the matching score level is performed for the multimodal fusion of two unimodal experts: a speaker and a face recognition system. Matching score level fusion needs a previous score normalization step before the fusion itself [Fox03,Indovina03].

Since unimodal scores are usually non-homogeneous, the normalization process transforms the different scores of each unimodal system into a comparable range of values. The state-of-the-art Z-score technique, that normalizes the global mean and variance of the scores, has been used for the normalization of the unimodal biometrics. The normalized scores x_{ZS} are computed as

$$x_{ZS} = \frac{a - mean(a)}{std(a)}$$

where *mean(a)* is the statistical mean of the set of scores a, and *std(a)* is its standard deviation.

After normalization, the converted scores are combined in the fusion process in order to obtain a single multimodal score. Matcher weighting, one of the most conventional fusion techniques for the arithmetic combination of the scores, has been used for the fusion process. For the application of this technique, each unimodal score is weighted according to its accuracy, so that the weights for more accurate matchers are higher than for those of less accurate matchers. The weighting factor for every biometric is proportional to the inverse of its EER. Denoting x^m, w^m and e^m the set of scores, the weigthing factor and the EER for the mth biometric, and M the number of biometrics, the fused score u is computed as:

$$u = \sum_{m=1}^{M} w^m x^m ,$$

where

$$w^m = \frac{\frac{1}{e^m}}{\sum_{m=1}^{M} \frac{1}{e^m}} .$$

4 Experiments

4.1 Experimental setup

The XM2VTS database and the Lausanne evaluation protocol (Configuration I) [Luettin98] have been used in this work. The database contains speech recordings and face images from 295 users, 200 clients and 95 impostors. It is organized in 4 sessions with 2 shots per session. Furthermore, each shot is formed by 1 front face image and two speaking sequences of 10 digits each, which yields a total number of 8 face images and 16 speech signals per user. For our experiments, scores from both speech signals in each shot are averaged.

In the speaker recognition system [Saeta06], speech utterances are processed in 25 ms frames, Hamming windowed and pre-emphasized. The feature set is formed by 12th order Mel-Frequency Cepstral Coefficients (MFCC) and the normalized log energy. Delta and delta-delta parameters are computed to form a 39-dimensional vector for each frame. Cepstral Mean Subtraction (CMS) is also applied.

Left-to-right HMM models with 2 states per phoneme and 4 mixture components per state are obtained for each digit. Client and world models have the same topology. Gaussian Mixture Models (GMM) of 32 mixture components are employed to model silence.

The identification platform uses Neurotechnology's VeriLook for face recognition. This engine is composed of five main modules. *Face detector* searches any number of faces in a grayscale image with different scales and head rotation. *Facial feature detector* estimates eyes position before that the *feature extractor* module computes discriminating facial features by means of Gabor wavelets. When several templates of the same face are available, a more precise recognition can be achieved by means of the *features generalization* technique which combines them to deal with intra-class variability. Finally, *feature matcher* module compares two templates.

4.2 Results

According to the Configuration I of the Lausanne protocol, client models are trained with the first shot of sessions 1, 2 and 3. In the evaluation phase the second shot of the same client sessions and all the data from a set of 25 impostors are used to obtain a user independent threshold that gives the Equal Error Rate (EER) in this dataset. Finally, in the test phase, both shots from the last client session and 70 impostors are used to measure the Half Total Error Rate (HTER) given by the threshold previously estimated. EER is also given for the test set to show the increase of error due to the database partition as it is explained below.

Table 1 illustrates the system performance in terms of identification error for each individual biometric modality, speech and face, and the fusion of both.

Table 1: Identification errors

	FACE	VOICE	FUSION
EER – eval	1.56%	1.20%	0.29%
HTER – test	4.60%	4.75%	1.00%
EER – test	2.22%	4.17%	0.97%

Table 2: FP and FN in test set for EER-eval threshold

	FACE	VOICE	FUSION
FPt	7.71%	2.00%	0.74%
FNt	1.50%	7.50%	1.25%

Voice performs better than face for the evaluation set, with an accuracy almost a 25 % higher. However, whereas identification delay is about 0.5s for face recognition, it takes an average of 19s for speaker identification in a 200 client database. In addition, the EER degradation in the test set with regard to the evaluation set is more remarkable for voice than for face, probably due to the use of speech sequences from the same sessions for training and evaluation and a different session for the test set.

Figure 1 shows fusion error curves for both, evaluation and test sets. Figure 2 shows in detail the intersection of these curves, where the values in table 1 can be observed.

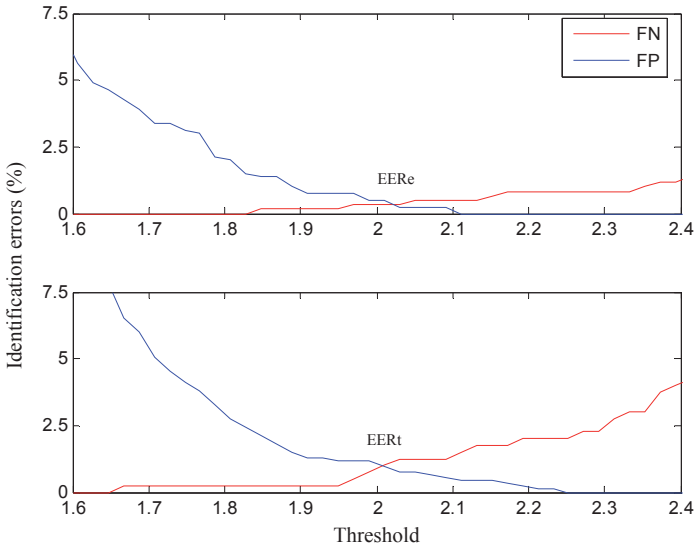

Figure 1: FP and FN for evaluation (top) and test (bottom) sets

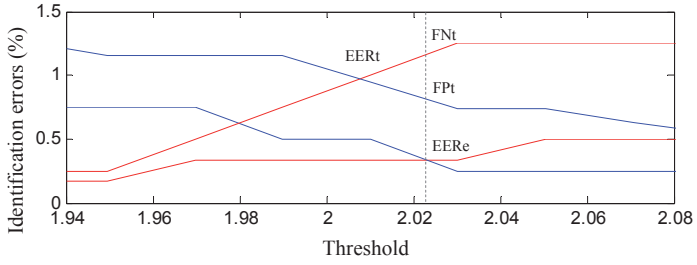

Figure 2: Evaluation and test identification errors from Tables 1 and 2

The search strategy described in section 2 is used in order to reduce recognition delay. Table 3 shows evaluation and test errors as the pre-selected number of faces, *N*, is reduced. Identification delay is also shown.

In Table 1 we observe that voice performs better than face. In contrast, for the test set face biometrics overcomes voice biometrics in terms of error rate. In our opinion, it can explain why the HTER increases

in the evaluation set when a smaller number of faces is pre-selected, and a minor weight is given to the voice recognition system. However, in the test set, giving a higher weight to face recognition discards potential errors in voice recognition, and reduces the HTER when *N* decreases.

Table 3: HTER for different values of N

N	200	100	50	20	10	5	3	1
eval	0.29	0.50	0.56	0.65	0.71	0.79	1.00	1.00
test	0.97	0.85	0.46	0.28	0.23	0.51	0.51	0.74
delay (s)	19	10	5.2	2.4	1.5	1.1	0.9	0.7

Figures 3 and 4 show graphically results from Table 3. We can see that the HTER remains practically unalterable from *N*=60 faces because the true identity is in most cases within the pre-selected faces. In addition, we can see that for *N*<10 faces the system performance is similar to the unimodal case with face recognition.

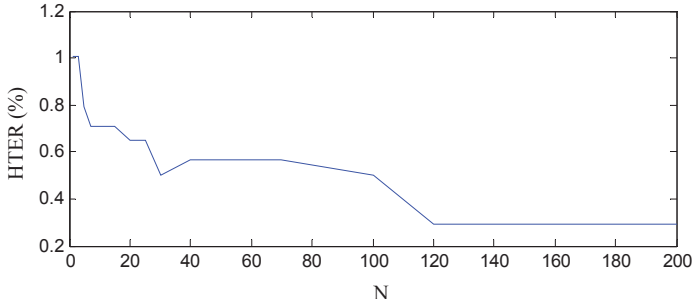

Figure 3: HTER in the evaluation set

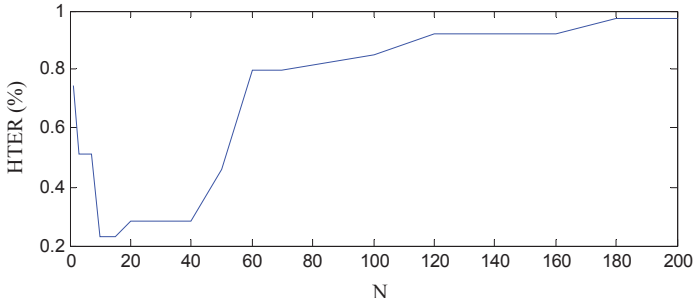

Figure 4: HTER in the test set

5 Conclusions

Biometric technologies are commonly used to enhance security and control the right to access to certain places. Forensic applications are not included in access control ones but they are becoming really important. They normally use large-scale databases and try to identify an individual among a group of previously enrolled users. Main challenges have to deal with the length of the databases, the identification delay and the performance in terms of EER. On the other hand, multimodal applications are also gaining popularity and can help to cope with the challenges mentioned before.

In this paper, we introduce a method to profit from the speed of face recognition with regard to speaker recognition to identify users in multimodal databases. Face recognition is used as the main engine to select a reduced number of speakers and speaker recognition provides an adjustment to improve speed as well as error rates. The final result can be seen as a trade-off between the identification delay and the error performance. The lower error is obtained for a selection of ten users through face recognition while lower delay is obviously provided for only one.

Acknowledgements

This work has been partially funded by the AGATHA project from the Spanish Ministry of Industry, Tourism and Trade.

References

[Auckenthaler01] Auckenthaler R., Mason J.S.: Gaussian selection applied to text-independent speaker verification. In Proc. Speaker Odyssey: The Speaker Recognition Workshop (Odyssey 2001), Crete, Greece, 2001, pp.83-88.

[Beigi99] Beigi H.S.M., Maes S.H., Sorensen J.S., Chaudhari U.V.: A hierarchical approach to large-scale speaker recognition. In Proc. 6th European Conf. Speech Communication and Technology (Eurospeech 1999). Budapest, 1999.

[Bolle04] Bolle R. M., Connell J. H., Pankanti S., Ratha N. K., and Senior A. W.: Guide to Biometrics. Editor: Springer, New York. 2004.

[Fox03] Fox N. A., Gross R., Chazal P., Cohn J. F., and Reilly R. B.: Person identification using automatic integration of speech, lip and face experts. In ACM SIGMM 2003 Multimedia Biometrics Methods and Applications Workshop, Berkeley, CA, 2003.

[Kinnunen06] Kinnunen T., Karpov E., Fränti P.: Real-Time Speaker Identification and Verification. In IEEE Transactions on Audio, Speech, and Language Processing, vol. 14, no. 1, January 2006.

[Indovina03] Indovina M., Uludag U., Snelik R., Mink A., and Jain A.: Multimodal Biometric Authentication Methods: A COTS Approach. In MMUA, Workshop on Multimodal User Authentication, Santa Barbara, CA, 2003.

[Luettin98] Luettin J., Maitre G.: Evaluation protocol for the XM2FDB database (Lausanne protocol). In Communication 98-05, IDIAP, Martigny, Switzerland, 1998

[McLaughin99] McLaughlin J., Reynolds D.A., Gleason T.: A study of computation speed-ups of the GMM-UBM speaker recognition system. In Proc. 6th European Conf. Speech Communication and Technology (Eurospeech 1999). Budapest, 1999.

[Pellom98] Pellom B.L., Hansen J.H.L.: An efficient scoring algorithm for gaussian mixture model based speaker identification. In IEEE Signal Process. Lett. vol. 5, no 11, pp. 281-284, 1998

[Reynolds95] Reynolds D.A., Rose R.C.: Robust text-independent speaker identification using gaussian mixture speaker models. In IEEE Trans. Speech Audio Process., vol. 3, no. 1, pp. 72-83. 1995.

[Saeta06] Saeta J.R., Hernando J.: Weighting scores to improve speaker-dependent threshold estimation in text-dependent speaker verification. In Lecture Notes in Computer Science. Editor: Springer-Verlag, vol. 3817, 2006, pp. 81-91.

[Sun03] Sun B., Liu W., Zhong Q.: Hierarchical speaker identification using speaker clustering. In Proc. Int. Conf. Natural Language Processing and Knowledge Engineering 2003, Beijin, China, 2003, pp. 299-304.