Automatic Person Recognition by Using Acoustic and Geometric Features

R. Brunelli[1], **D. Falavigna**[1] **L. Stringa**[1], **T. Poggio**[2,1]
[1] Istituto per la Ricerca Scientifica e Tecnologica
I-38050 Povo, Trento, ITALY
[2]Artificial Intelligence Laboratory
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139, USA

Abstract—**The paper describes a multisensorial person identification system: visual and acoustic cues are used jointly for person identification. A simple approach, based on the fusion of the lists of scores produced independently by a speaker recognition system and a face recognition system, is presented. Experiments are reported which show that integration of visual and acoustic information enhances both performance and reliability of the separate systems. Finally two network architectures, based on radial basis function theory, are proposed to describe integration at different levels of abstraction.**

Keywords: face recognition, speaker identification, classification

1. Introduction

This paper describes an automatic person recognition system[1] which uses both acoustic features, derived from the analysis of a given speech signal, and visual ones, related to distinctive parameters of the face of the person who uttered that speech signal. Visual and acoustic cues are used jointly for person identification: several methods to combine, at different levels, the acoustic and visual information will be described. Experiments will be presented which show the superior performance of the whole system with respect to both the Speaker Recognition System (SRS) and the Face Recognition System (FRS) considered separately. The system can be used for either identification (e.g. as an electronic concierge to recognize people in small organizations) or verification applications (e.g. as a smart door key to control the entrance of a house) even if, in our experiments, we have stressed only the identification function.

The SRS utilizes acoustic parameters computed from the spectrum of short time windows of the speech signal while the FRS is based on geometric data represented by a vector describing discriminant facial features such as: position and width of nose, chin shape and so on. We have considered different ways to combine the SRS and the FRS results:

1. at the level of the outputs of the single classifiers: the list of scores (probability estimates or distances) produced independently by the two systems can be used in an integrated classifier system (i.e. the two sets of scores are optimally weighted and summed in order to produce a unique final list);
2. at the level of the measurements carried out on both the speech and visual signals: acoustic and geometric parameters are considered as a unique vector, lying in the cartesian product of the acoustic and visual spaces, which will be successively classified by a speaker specific classifier. Examples of such a classifier could be: a Bayesian classifier, a MultiLayer Perceptron (MLP) classifier or a Radial Basis Function (RBF) classifier.

The experiments reported in this paper are based on the first integration strategy. Some HyperBF network architectures supporting the second integration strategy are also described. The paper is organized as follows:

- Section 2 describes the overall system architecture;
- Section 3 presents the acoustic and visual databases;
- Section 4 analyzes the SRS and FRS utilized for this work;
- Section 5 describes the method used in this work to integrate the two systems. An alternative strategy, currently under investigation, is also presented;
- Section 6 reports the results of the experiments.

2. System overview

The basic structure of the identification system is depicted in Figure 1. An attention module detects changes in the image captured by a CCD camera by using background subtraction and thresholding. Whenever a significant change in the scene is detected, a snapping module is activated[2]. This module waits until a stable scene is detected (a still person staring into the camera) and check for the satisfaction of simple constraints (mainly on the changed area) before unlocking the separate action of the two recognition systems. The snapped image is directly fed to the face recognition system while the system asks the person to utter a sequence of digits in a continuous way. A boundary detection module, applied to the input speech signal, separates the voice signal from the background noise determining both the starting and ending times of each voice segment and the corresponding duration. If the total duration of the voice segments, detected by the start-end point detection module, is not long enough the system asks the user to read again the requested digits.

The outputs of the SRS and FRS are fed into an integration module whose architecture is discussed in the following sections. The output of the integration module is sent to the last module that identifies the user.

3. Test database description

The acoustic and the visual databases consist of 42 and 47 people respectively; the database used in the experiments is composed of 33 people, without shaves and mous-

[1] This research was done within MAIA, the integrated Artificial Intelligence project under development at IRST [20], [18]

[2] We thank Dr. S. Messelodi for kindly sharing the C language implementation of the attention and snapping module.

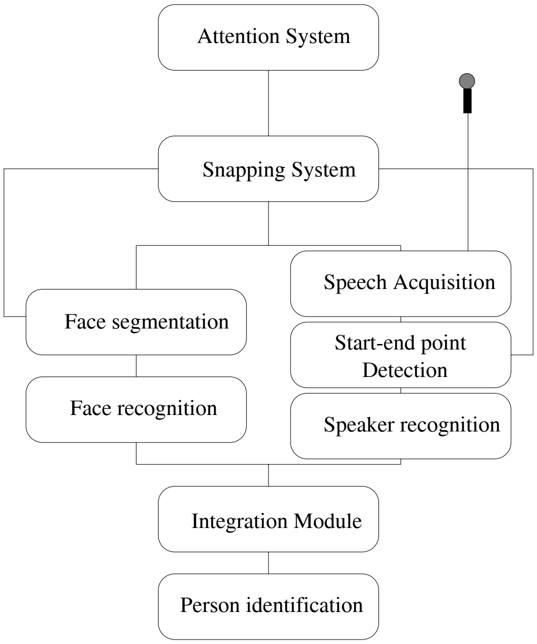


Fig. 1. The flow of activation of the different modules belonging to the multisensorial identification system.

tache, belonging to both databases. Training of FRS and SRS was performed separately for each person in the two databases while testing (see Section 5) was led on the common 33 people database.

3.1 The acoustic database

The database consists of 8400 isolated Italian digits, from zero to nine. Each speaker uttered 200 digits in five, different, recording sessions: during each session the speaker provided four repetitions of each digit. The entire database was collected over a period of about two months: each recording session is separated from the previous one by an interval varying from three to 10 days. The recordings were realized in a room of our Institute by means of a Digital Audio Tape (DAT) equipment (SONY TCD-D10). The digital signal on the DAT tape (sampled at 48 kHz) was downsampled to 16 kHz, manually endpointed and stored on a computer disk.

3.2 The visual database

The database is composed of 188 images, four per person. Of the four pictures available, the first two were taken in the same session (at a time interval of a few minutes) while the other pictures were taken at intervals of some weeks (two to four). The pictures were acquired by a CCD camera at a resolution of 512×512 pixels as frontal views. The subjects were asked to look into the camera but no particular efforts were made to ensure perfectly frontal images. The illumination was partially controlled: the same powerful light was used but the environment where the pictures were acquired was exposed to sun light through windows.

The pictures were taken randomly during the day time. The distance of the subject from the camera was fixed only approximately, so that scale variations of as much as 30% were possible.

4. The recognition systems

In this section the SRS and the FRS used in the experiments will be described. Both systems operate in two steps:

1. a sequence of parameter vectors or a single parameter vector are derived from the speech signal or the visual signal respectively;
2. a matching procedure, based on distance measures with respect to reference models, is applied to the input parameter vectors so that each system produces a list of scores labeled with the identifier of the corresponding reference model.

4.1 The speaker recognition system

Automatic speaker recognition is a topic which has been widely investigated in the past years [2], [11], [9]; recently some methods based on the same processing and modeling techniques used also for speech recognition have been studied and tested [1], [24].

Speaker recognition systems can be divided, according to the application area, into *speaker verification* or *speaker identification* systems. A speaker verification system simply controls (confirms or not) the identity claimed by a person (e.g. before she/he accesses a reserved place or service). A speaker identification system has to determine who (within a known group of people) uttered the input speech signal. Speaker recognition systems can also be *text dependent* (i.e. the user must utter a predefined key sentence) or *text independent*. The major problem in speaker recognition is represented by intersession variability (variability of the speech signal with time) due to both different recording or transmission conditions and to intrinsic variability in the voice of people.

The SRS used for this work is based on Vector Quantization (VQ) and is similar to the one described in [1]. A block diagram of this system is depicted in Figure 2.

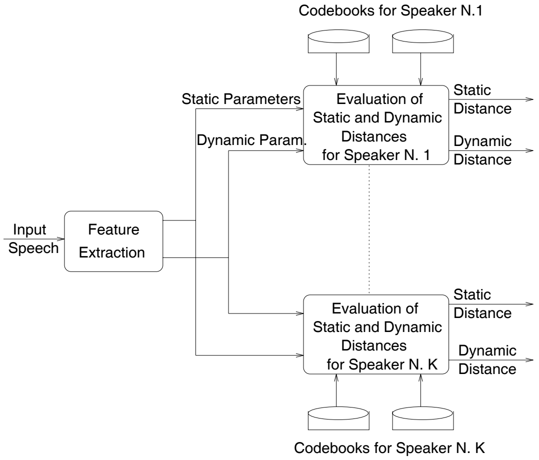

Codebooks for Speaker N.1
Static Parameters
Dynamic Param.
Evaluation of Static and Dynamic Distances for Speaker N. 1
Static Distance
Dynamic Distance
Input Speech
Feature Extraction
Evaluation of Static and Dynamic Distances for Speaker N. K
Static Distance
Dynamic Distance
Codebooks for Speaker N. K

Fig. 2. Block diagram of the VQ based speaker recognition system.

The system measures the distances between two distinct sets of acoustic parameters (static and dynamic) and corresponding prototypes, *codebooks*, derived, during a training phase, from speaker specific speech material.

4.1.1 Speech signal analysis

The speech signal, sampled at 16 kHz and quantized over 16 bit, is first preemphasized by using a digital filter having transfer function: $H(z) = 1 - 0.95 \times z^{-1}$. The preemphasized signal is analyzed every 10ms using a 20ms Hamming window and, for each window position, the following parameters are computed:

1. 8 Mel scale cepstral coefficients [19] computed from the log-energy outputs of a 24 triangular band-pass filter bank applied to the power spectrum of the given window; these parameters are called *static*;
2. the corresponding first order time derivatives, evaluated by means of a linear fit over nine frames of static parameters, centered on the given window (these parameters are called *dynamic*).

4.1.2 The VQ based speaker recognition system

As seen previously the reference models of each speaker (see Figure 2) consist of two codebooks: one corresponding to the static parameters and one to the dynamic ones. The codebooks were generated by applying the Linde-Buzo-Gray algorithm [12] to a set of short time spectral vectors (static or dynamic) derived from 100 digits belonging to the first two and half recording sessions of each speaker. This algorithm searches for prototype vectors which minimize a global distortion measure defined on the given training set.

A weighted Euclidean distance was used for codebook design and recognition; the weights correspond to the inverse of the pooled variances of the components of the training vectors averaged over all training utterances and speakers. Therefore if θ_i, ψ_i are two parameter vectors, their distance is defined as:

$$d(\theta_i, \psi_i) = \sum_{k=1}^{p} \frac{1}{\sigma_k^2} (\theta_{i_k} - \psi_{i_k})^2 \qquad (1)$$

where σ_k^2 represents the average pooled variance of the k^{th} component of the parameter vector. As previously seen p is equal to 8.

During the recognition phase, the distances (see Figure 2) between the sequences of static and dynamic vectors, obtained from the input speech signal, and the corresponding codebooks of each speaker are evaluated. Therefore, if $\mathbf{\Theta} = \theta_1, \ldots, \theta_T$ is the static (or dynamic) input sequence and $\mathbf{\Psi_j} = \psi_{j_1}, \ldots, \psi_{j_M}$ are the vectors of the j^{th} static (or dynamic) codebook, then the total static (or dynamic) distortion is defined as:

$$D(\mathbf{\Theta}, \mathbf{\Psi_j}) = \frac{1}{T} \sum_{t=1}^{T} \min_{i=1}^{M} d(\theta_t, \psi_{j_i}) \qquad (2)$$

The static and dynamic distances are then normalized with respect to their average values, computed on the training set, and summed. Performance of the system depends on both the acoustic resolution (i.e. the number M of vectors in each speaker codebook) and the number L of digits contained in the input signal. The average identification error evaluated on a test set composed of 100 digits per speaker (belonging to the last two and half recording sessions) is 49% for $M = 4$, $L = 1$ and 0% for $M = 64$, $L = 3$. In Figure 3 results for spectral resolutions varying from 4 to 64 and utterance lengths varying from 1 to 10 digits are reported.

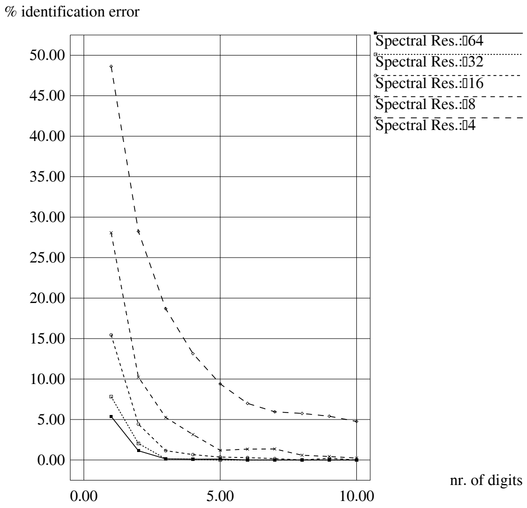


Fig. 3. Error rate of the SRS, evaluated on the whole 42 speaker test set, as a function of the number of digits in the iput signal: different curves correspond to different spectral resolution values.

4.2 The face recognition system

Recognition of people by visual cues is a basic task humans perform every day without apparent effort. The ease with which we recognize familiar people from their face has led to underestimate the difficulty of the problem. Some extended psychophysical experiments unrevealed the fact that even for people the recognition process requires a lot of processing and is by no means an innate ability.

There are two main strategies for automatic face recognition, both of which can be said to mimic some of the processes used by people:

iconic: it is based on the comparison of suitably preprocessed image patches; recognition is effected by comparing (e.g. through the value of cross-correlation or some other suitable distance measure) an unknown image with stored templates of distinctive facial regions [3], [25], [21], [22], [6];

geometric: a set of geometric features, describing the size and the layout of the different features in the faces, is computed and recognition proceeds by comparing the unknown descriptive vector with a set of reference vectors (known people) stored in a data base [13], [7], [8], [14], [4], [5], [6].

Several approaches can be classified within this simple taxonomy and a comparison of the two basic strategies can be found in [6]. In this paper we focus on the geometric strategy. The reason is twofold:

1. it gives a more compact representation and guarantees a high speed in the recognition process;
2. it gives an example of how good performance results can be obtained by integrating simple (and fast) recognition modules which do not have a very high performance when considered separately.

As will be apparent from the discussion of the integration strategies, the use of a *template matching* strategy fits in a natural way. The typical way *template matching* proceeds is by comparison of suitably preprocessed images (or patches of images) representing the salient features of the face pattern. Whether the comparison is done by a suitable distance measure or by a correlation coefficient the result

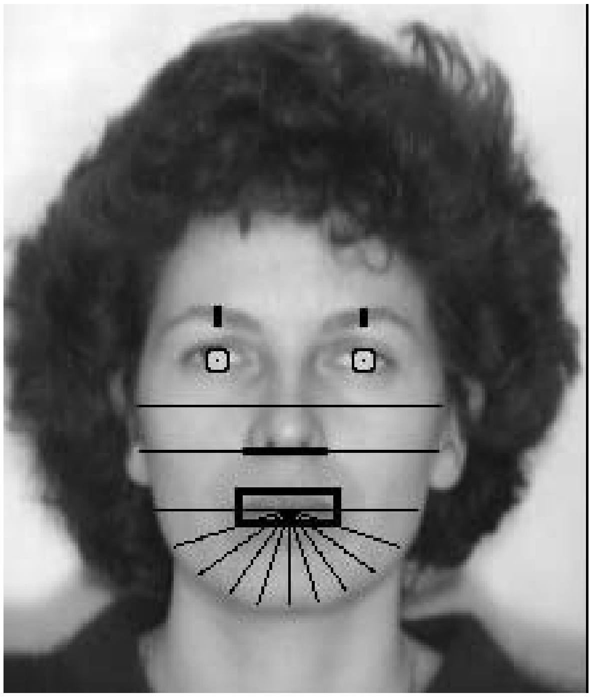


Fig. 4. Geometrical features (black) used in the face recognition experiments (information on the eyebrow arch is not reported).

can be readily incorporated into the proposed system both in alternative or conjunction with the explicitly used information.

A set of geometrical features describing a frontal view of a face is computed automatically through the following steps [6]:

1. eyes are located so that the image can be normalized both in size and rotation in the image plane [23];
2. an average face model is used to progressively focus the system, in sequential way, on the different regions of the face so that relevant feature points can be computed;
3. a descriptive vector is built from the relative positions of the feature points.

This results in the computation of 35 geometrical features (see Figure 4) which can be used for recognition:

1. eyebrows thickness and vertical position at the eye center position;
2. a coarse description of the left eyebrows arches (8 measurements);
3. nose vertical position and width;
4. mouth vertical position, width (upper and lower lips) and height;
5. eleven radii describing the chin shape;
6. bigonial breadth (face width at nose position);
7. zygomatic breadth (face width halfway between nose tip and eyes).

A detailed description of the employed algorithms can be found in [6]. Classification can then be based on a Bayes classifier. We make the simplifying assumption that the measurements of the different features share the same Gaussian distribution for all the people apart from their average value [4]. The covariance matrix Σ can then be estimated and classification can be based [10] on the following distance which is related to the probability of the given measurement:

$$d(\mathbf{x}, \mathbf{m}_j) = (\mathbf{x} - \mathbf{m}_j)^T \Sigma^{-1} (\mathbf{x} - \mathbf{m}_j) \tag{3}$$

The identification of an unknown vector is taken to be that of the nearest vector in the database of known people. The database used for the reported experiments consists of 132 vectors, 4 per person, representing the complete set of 35 geometrical features.

The performance is estimated on a round robin recognition experiment where three of the available vectors are used to compute an average representative vector while the remaining one is used for testing. The average performance on this database was found to be 92%

5. Integration of Vision and Voice

In this section we briefly describe methods for integrating the two classifiers into a more robust recognition system. There are two general classes of possible architectures: the first is based on the integration of the output of two independent classifiers, one for voice and one for face; the second one is based on the direct combination of voice and face features. The first approach is simpler and is used in the reported integration experiments. The second method is more speculative and could be applied in a larger scale application.

5.1 Score integration

Both classifiers are essentially nearest neighbor with a suitably defined metric. As said above, when given an unknown input they generate a list of possible labels, marked with their distance from the input vector. The list is sorted with increasing distance so that the first element of the list should be taken as the correct matching label. The main difficulty in combining the results of the two classifiers is given by the inhomogeneous distances they produce. The simplest way to fix this is by inverse standard deviation normalization. Given the two lists, if d_{v_i} and d_{s_i} represent the distances computed respectively by the face recognizer and the speaker recognizer for the i^{th} reference and σ_v^2 and σ_s^2 are the corresponding variances, a combined distance can be defined as

$$D_i = \frac{d_{v_i}}{\sigma_v} + \frac{d_{s_i}}{\sigma_s} \tag{4}$$

A natural way to look at the answer of a nearest neighbor classifier is to map it into a list of scores rather than a list of distances. A possible mapping is:

$$s_{x_i} = e^{-\frac{d_{x_i}}{\sigma_x}} \tag{5}$$

where x stays alternatively for *V*ision or *S*peech. This mapping associates to a distance a value in the interval $(0, 1]$. In some sense, the higher the score the more likely correct the correspondence is. Each list could be further normalized by imposing the following condition:

$$\sum_i s_{x_i} = 1 \tag{6}$$

The resulting lists could be given a Bayes interpretation, suggesting the following integration strategy in the hypothesis that the two systems are independent:

$$S_i = s_{v_i} \times s_{s_i} \tag{7}$$

As the two recognition systems do not have the same performance, it is natural to introduce a weighted merged score:

$$S(w)_i = s_{v_i}^{w} \times s_{s_i}^{(1-w)} \tag{8}$$

The optimal weight w can be found by maximizing the performance of the integrated system on one of the available test sets.

5.2 A more integrated system

A more close integration of the two recognition systems could be made with a network which learns the relative reliability and discriminating power of the different features. We propose two such architectures based on HyperBF Networks. Before introducing the architectures, let us briefly recall the basic theory of these networks in the general framework of solving learning problems through multivariate function reconstruction [17].

Whenever the examples in a learning task are given in numeric form learning can be seen as a problem of surface reconstruction from sparse data (the examples). It is fairly evident that, as it is, the task of reconstructing a surface given its value at some points is ill posed. To have a well behaved problem some additional constraints must be imposed. The most important, as far as learning is concerned, is *smoothness*. If the reconstructed function is required to be smooth we are assured that generalization from available examples is indeed possible. The assumption of smoothness allows us to formulate a variational problem whose solution is the surface we are looking for:

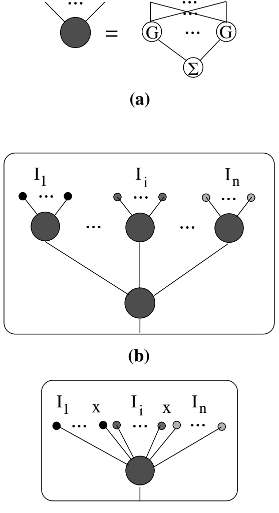
$$H[f] = \sum_{i=1}^{N} (y_i - f(\mathbf{x}_i))^2 + \lambda \phi[f] \tag{9}$$

where the first term measures the distance between the data and the desired solution f; the second term is a functional reflecting the cost associated with the deviation from some constraint reflecting some *a priori* knowledge. The unknown function can be considered as a linear combination of the following form:

$$f(\mathbf{x}) = \sum_j c_j G(|\mathbf{x} - \mathbf{t}_j|_W) \tag{10}$$

where $\mathbf{t}_j$ are called expansion centers and W is a square matrix used to compute the norm of a vector $(\mathbf{x} - \mathbf{t}_j)^T W^T W (\mathbf{x} - \mathbf{t}_j)$. Function G could be, among others, a Gaussian or a multiquadric (for a more complete introduction to the theory of HyperBF Network see [17]). The variational problem can be solved finding the coefficients, the centers and the metric minimizing the reconstruction error on the available examples plus the smoothness enforcing term.

The network architectures we propose for the integrated recognition system are shown in Figure 5. The underlying idea is to build a set of HyperBF modules, one per person to be identified, trained to reconstruct the characteristic function of that person. The module is supposed to output 1 on inputs corresponding to the given person and 0 otherwise. Each module is trained using both positive and negative examples (*competitive learning*) so that both optimal example selection and adaptive metric can be profitably used. A typical choice for the basis function in this type of network would be the Gaussian.



(a)

(b)

(c)

Fig. 5. (a): The basic HyperBF module; (b): A score-level integration network; (c): A feature-level integration network.

The first network, see Figure 5b, is closer to the simple approach outlined above. The modules in the first layer represent the outputs of recognizers based on different types of data (such as visual, static or dynamic acoustic information). The modules can be trained separately, taking advantage of the reduced input dimensionality. The integration is operated by the module in the second layer which is trained, again on a per person basis, to maximize its output on the vectors of the corresponding person. The resulting network could be called a score-integration network. The approach detailed in the current paper can be seen as a simplified version of this type of network where the first layer is removed.

The second network, see Figure 5c, architecture we propose is a feature-integration network. The input to the HyperBF modules is taken to be the cartesian product of the different inputs and the module takes advantage of the simultaneous knowledge of the available information. This network presents a more difficult training due to the extended input dimensionality.

6. Experiments and results

The strategy we choose to integrate the two recognition systems has been that of score fusion. The distances given by each of the systems have been transformed into score lists using exponential mapping and inverse variance normalization, as explained in 5.1.

As said previously the database used for the integration experiments consists of 33 people. For each face four lists of scores were available while for each speaker and for each utterance length, varying from one to six digits, we randomly selected eight lists of scores from those corresponding to codebook sizes equal to four. In this way a total of 32 integration experiments could be performed for each person and for each input utterance length (from one to six digits). One of these 32 sets of score lists was used to estimate the optimal integration weight (see Figure 8) while performance was measured on the remaining sets. A useful data on the robustness of the classification is given by an estimate of the intra-class variability as opposed to the inter-class variability. This can be done using the so called Min/Max ratio (see [15], [16]) which is defined as the minimum distance on a wrong correspondence over the maximum distance from the correct correspondence. To estimate the optimal weighting factor the interval $[0, 1]$, representing the weight of the FRS, was evenly sampled and the value maximizing the performance was chosen (ties were resolved by maximizing the Min/Max ratio). The performance of the SRS, FRS and integrated systems are quantified by:

1. the average performance (Figure 6);
2. the Min/Max ratio (Figure 7).

The analysis of this integrated performance (see Figure 6) shows that, even with the simple integration scheme, nearly perfect recognition is achieved. The performance of the FRS is 92% while that of the SRS, using different speaker models, varies from 51% to 100%: the performance of the integrated system, using the less complex speaker model is already up to 95% and achieves 100% with a speaker model of low complexity. Further benefits of integration are evident in Figure 7, where the Min/Max ratio, which represents an estimate of the average separation among classes, is plotted. The increased class separation of the integrated system, as measured by the Min/Max ratio, suggests that rejection can be introduced with a more limited impact on performance if compared to the two independent systems. The curves in Figure 8 represent the performance of the integrated system as a function of the weight used to merge the scores; as can be seen the performance is a smooth function of the weight: this means that the the system is not very sensitive to the weighting factor. As it could be easily anticipated from the performance of the two different systems, the optimal combining weight shifts from a vision dominance at a low number of digits toward a voice dominance at an high number of digits. The same result would be reached passing from lower to higher spectral resolutions of the SRS.

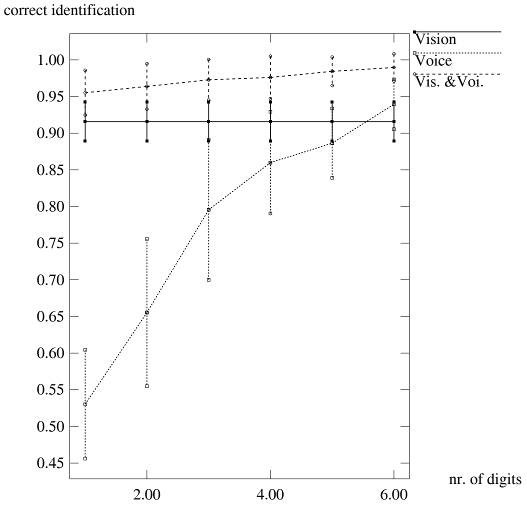


Fig. 6. Performance of the FRS (horizontal line) and of both the SRS and the integrated system versus the number of digits in the input speech signal; the spectral resolution of the SRS is 4; the vertical bars represent the standard deviation of the measured quantity.

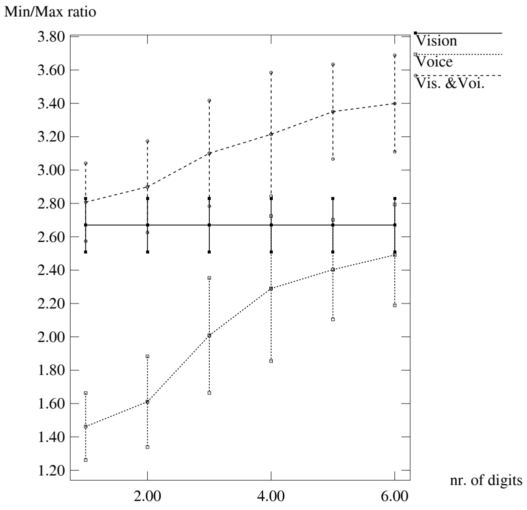


Fig. 7. Min/Max ratio of the FRS, SRS and integrated system (these last two are functions of the number of digits in the input speech signal); the spectral resolution of the SRS is 4; the vertical bars represent the standard deviation of the measured quantity.

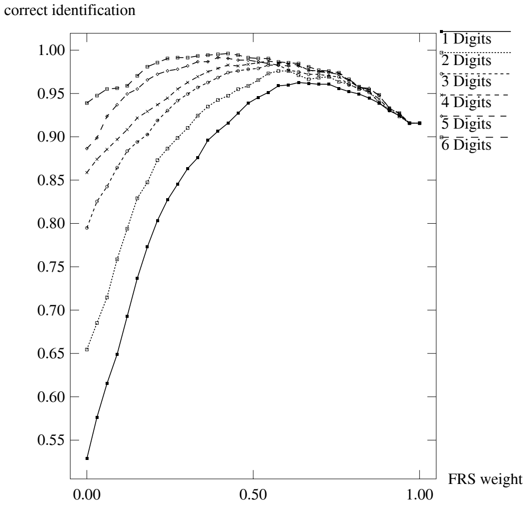


Fig. 8. Performance of the integrated system as a function of the weight assigned to the FRS using a score level integration method; different curves correspond to different number of digits in the input speech signal. The spectral resolution of the SRS is 4.

7. Conclusions

In this paper the superior performance which can be attained by using multisensorial input has been demonstrated. Integration of two recognizing systems, based on speech and vision respectively, greatly increases the performance reaching *state of the art* with a low to moderate complexity in the constituent systems. Finally some speculative network architectures for integration have been proposed for use on larger databases which actually requires computationally expensive recognition systems. A system based on the described integration strategy and working in an uncontrolled environment is currently under evaluation: preliminary results suggest that the use of multiple identification cues has a major impact not only on absolute performance but also on the ability of the system to reject an unknown user.

References

[1] F. K. Soong A. E. Rosenberg. Evaluation of a Vector Quantization Talker Recognition System in Text Independent and Text Dependent Modes. *Computer Speech and Language*, 2(3–4):143–157, 1987.

[2] B. Atal. Automatic Recognition of Speakers from Their Voices. *Proceedings of IEEE*, 64:460–475, 1976.

[3] R. J. Baron. Mechanisms of human facial recognition. *International Journal of Man Machine Studies*, 15:137–178, 1981.

[4] Martin Bichsel. *Strategies of Robust Object Recognition for the Identification of Human Faces*. PhD thesis, Eidgenossischen Technischen Hochschule, Zurich, 1991.

[5] R. Brunelli and T. Poggio. Face Recognition through Geometrical Features. In G. Sandini, editor, *ECCV'92, Santa Margherita Ligure*, pages 792–800. Springer-Verlag, 1992.

[6] R. Brunelli and T. Poggio. Face Recognition: Features versus Templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(10):1042–1052, 1993.

[7] G. Cottrell and M. Fleming. Face recognition using unsupervised feature extraction. In *Proceedings of the International Neural Network Conference*, 1990.

[8] I. Craw, H. Ellis, and J.R. Lishman. Automatic extraction of face features. *Pattern Recognition Letters*, 5:183–187, 1987.

[9] G. R. Doddington. Speaker Recognition, Identifying People by Their Voices. *Proceedings of IEEE*, 73(11), 1985.

[10] R. O. Duda and P. E. Hart. *Pattern Recognition and Scene Analysis*. Wiley, New York, 1973.

[11] S. Furui. Cepstrum Analysis Technique for Automatic Speaker Verification. *IEEE Transactions on Acoustic, Speech and Signal Processing*, 29(1):254–272, 1981.

[12] H. Gish J. Makhoul, S.Roucos. Vector Quantization in Speech Coding. *Proceedings of IEEE*, 73(11):1551–1588, 1985.

[13] T. Kanade. Picture processing by computer complex and recognition of human faces. Technical report, Kyoto University, Dept. of Information Science, 1973.

[14] O. Nakamura, S. Mathur, and T. Minami. Identification of human faces based on isodensity maps. *Pattern Recognition*, 24(3):263–272, 1991.

[15] T. Poggio and S. Edelman. A Network that Learns to Recognize Three-Dimensional objects. *Nature*, 343(6225):1–3, 1990.

[16] T. Poggio and F. Girosi. A theory of networks for approximation and learning. A.I. Memo No. 1140, Massachusetts Institute of Technology, 1989.

[17] T. Poggio and F. Girosi. Networks for Approximation and Learning. In *Proc. of the IEEE, Vol. 78*, pages 1481–1497, 1990.

[18] T. Poggio and L. Stringa. A Project for an Intelligent System: Vision and Learning. *International Journal of Quantum Chemistry*, 42:727–739, 1992.

[19] P. Melmerstein S. B. Davis. Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuosly Spoken Sentences. *IEEE Transactions on Acoustic, Speech and Signal Processing*, 28(4):357–366, 1980.

[20] L. Stringa. An Integrated Approach to Artificial Intelligence: the MAIA Project. Technical Report 9110-26, I.R.S.T, 1991.

[21] L. Stringa. Automatic Face Recognition using Directional Derivatives. Technical Report 9205-04, I.R.S.T, 1991.

[22] L. Stringa. S_Net Implementation of a Face Recognizer Based on Directional Derivatives. In E. R. Caianiello, editor, *Proceddings Fifth Italian Workshop on Neural Nets, Vietri*. World Scientific, 1992.

[23] L. Stringa. Eyes Detection for Face Recognition. *Applied Artificial Intelligence*, 7:365–382, 1993.

[24] N. Z. Tishby. On the Application of Mixture AR Hidden Markov Models to Text Independent Speaker Recognition. *IEEE Transactions on Signal Processing*, 39(3):563–570, 1991.

[25] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, 1991.