Person Recognition Using Acoustic and Visual Cues

Daniele Falavigna, Roberto Brunelli

Abstract— **A person recognition system that makes use of acoustic and visual features is described. The system combines features at a score level and is capable of performing either an identification or a rejection. The improved performance of the integrated system with respect to the separate subsystems (acoustic and visual) is quantified.**

Keywords— **Normalization, Integration, Rejection.**

1. Introduction

Person identification and verification is a crucial task performed by both people and automatic systems in every day life. Since in many situations mistakes must be avoided (for example, banking transactions, database inquiry, access control to reserved areas or buildings and so on) it is necessary to adopt secure and reliable techniques to assure the desired degree of safety.

The approach chosen at IRST to cope with this problem [1] is the integration of different types of distinctive cues of the person. A system that uses both acoustic features, extracted from the user's speech signal, and visual ones, derived from an image of the face, was developed. The features are processed by independent subsystems that provide similarity measures with respect to models (acoustic or visual) of reference users. The measures, represented by lists of scores, are combined into a unique final list on which the decision about the identity of the user is made. Clearly if the acoustic or visual cues are not available the system behaves exactly as a face recognizer or a speaker recognizer respectively; furthermore, the method proposed for the integration of the acoustic and visual information is quite general, thus allowing the use of other types of cues (e.g. fingerprints or other biometric data) to increase the overall reliability of the system. This paper will focus on the speaker recognition system but results of the face recognizer and the integrated system also will be given.

The speaker recognition system is based on the Vector Quantization (VQ) technique and is similar to that described in [2]. Each speaker is represented by two codebooks (instantaneous and transitional) containing acoustic prototype vectors; during the recognition phase the distances among the acoustic features (instantaneous and transitional), derived from the spectral analysis of the input speech signal, and a corresponding set of reference codebook couples, are evaluated. In this way two lists of scores, having as many entries as the number of people in the reference database, are obtained. These lists are first normalized and then combined, as explained in Section 3, into a unique list. When this list is used to decide the identity of the person, a correct identification percentage equal to 88% is obtained[1].

The face recognition system makes use of a crosscorrelation technique to compare the input image of the face with the images in the reference database [3]. Three regions in particular, related to eyes, nose and mouth, are located within the face image and compared with the corresponding regions of the images contained in the reference database. This results in the evaluation of three lists of scores, one for each region, which are normalized and combined in the same way as that proposed for the acoustic data. The face recognition system has a correct identification percentage of 91%.

Finally the integrated system combines all of five lists of normalized scores, two acoustic and three visual, into a unique final list. When this list is used to decide the identity of the person, performance rises to 98%.

An important capability of the system is the rejection of an identification if it is not made with a sufficient degree of confidence. The method proposed for this task makes use of a linear classifier; the classifier domain contains both the integrated score of the best candidate and the corresponding ranks in the various normalized lists. In the following the speaker recognition system will be described in detail while an accurate report about the face recognizer can be found in [3]. The integration and rejection procedures will be described and the results for the separate acoustic and visual subsystems and for the integrated system will be given.

2. System description

A block diagram of the person recognition system is given in Figure 1. The input to the system consists of a randomly generated string of digits, that the user utter in a continuous way, and of the image of the user's face. The acoustic feature extraction block performs spectral analysis on the digitized speech signal, sampled at 16 kHz, every 10 ms. The analysis is carried out on overlapping Hamming windows having length equal to 20 ms. For each analysis window the following parameters are computed:

1. 8 Mel scaled cepstral coefficients (instantaneous parameters), derived from a bank of 24 triangular filters;
2. the corresponding first order time derivatives (transitional parameters);
3. a binary variable that allows the window to be marked as 'speech' or 'background noise'; this parameter is computed using the algorithm described in [4].

D. Falavigna, R. Brunelli: IRST, Povo, 38100 Trento, Italy. E-mail:falavi@irst.it, brunelli@irst.it.

[1]The performance has been evaluated on 164 test interactions for a reference database composed of 89 people.

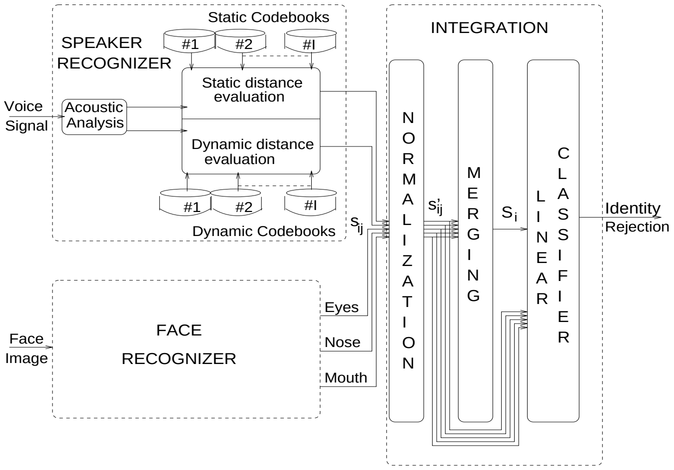

Fig. 1. Block diagram of the Person Recognition System.

The parameter vectors related to windows marked 'background noise' are removed from the two output sequences of the feature extraction block and the total distances from the corresponding speaker codebooks are evaluated. If $\boldsymbol{\Theta}$ is the instantaneous/transitional parameter sequence and $\boldsymbol{\Psi}_i$ is the i-th instantaneous/transitional codebook, then the total instantaneous/transitional distance is:

$$D(\boldsymbol{\Theta}, \boldsymbol{\Psi}_i) = \frac{1}{T} \sum_{k=1}^{T} \min_{j=1}^{M} (\boldsymbol{\theta}_k - \boldsymbol{\psi}_{ij})^t \boldsymbol{W}^{-1} (\boldsymbol{\theta}_k - \boldsymbol{\psi}_{ij}) \quad (1)$$

where $\boldsymbol{\Theta} = \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_T$, $\boldsymbol{\Psi}_i = \boldsymbol{\psi}_{i1}, \ldots, \boldsymbol{\psi}_{iM}$ $(1 \leq i \leq I)$, T is the number of analyzed windows, I is the number of people in the reference database, M is the codebook size, t denotes transpose and $\boldsymbol{W}$ is the covariance matrix of the feature vector components, which is supposed to be diagonal. If T, the number of analyzed windows, is lower than 200 (2 seconds) the system requires the user to repeat the digit string. The evaluated distances (instantaneous and transitional) are organized in two sorted lists of scores and sent to an integration module. This one combines them with three analogous lists produced by the visual subsystem and gives, as outputs, an identification or a rejection.

To design the speaker codebooks, 200 isolated utterances of Italian digits (from 0 to 9) for 89 users were used. For each user a frontal image of the face was acquired as well. The speech recordings were realized in an acoustically isolated room during five different sessions separated by a one week interval on average. A Digital Audio Tape (DAT) was used: the speech signal, sampled at 48 kHz, was downsampled to 16 kHz, manually end-pointed and stored on a computer disk. The training vectors were clustered using the LBG algorithm [5] and the centroids of the instantaneous and transitional training spaces were stored in a couple of codebooks. As demonstrated in [2] system performance depends on both the input utterance length and the codebook size; results of preliminary experiments have suggested that the speaker, in order to be identified, should utter a string of at least seven digits. The codebook size was set to $M = 64$, because higher values did not improve performance.

Finally, since background noise and acquisition conditions during the test phase were different from those used in training, the speaker codebooks were adapted to the test environment. In fact the test sessions were carried out in an office environment using an ARIEL board as acquisition channel. Furthermore the test phase was carried out about five months after the training recordings. The adaptation method is similar to the one described in [2]; in this method the acoustic prototypes of a given talker, represented by the centroids of the corresponding codebook, are linearly combined with acoustic vectors extracted from the test utterances. Therefore, in order to update the i-th codebook, a set of acoustic parameter vectors (adaptation vectors), derived from one interaction of the i-th user with the system, was used. The adaptation phase can be summarized as follows:

1. the centroid $\boldsymbol{\mu}_i$ of the adaptation vectors and the centroid $\boldsymbol{\nu}_i$ of the i-th codebook are evaluated;
2. the deviation $\boldsymbol{\Delta}_i = \boldsymbol{\mu}_i - \boldsymbol{\nu}_i$ is evaluated;
3. the vectors $\{\boldsymbol{c}_{i1}, \ldots, \boldsymbol{c}_{iM}\}$ of the codebook to adapt are shifted by the deviation vector $\boldsymbol{\Delta}_i$ obtaining the set $\{\boldsymbol{c}'_{ij} = \boldsymbol{c}_{ij} + \boldsymbol{\Delta}_i\}, 1 \leq j \leq M$;
4. the adaptation vectors are clustered using the set $\{\boldsymbol{c}'_{ij}\}$ so that a set of centroids $\{\boldsymbol{o}_{ij}\}$ and the corresponding cell occupancy $\{n_{ij}\}$ are evaluated;
5. The adapted codebook $\{\boldsymbol{\psi}_{ij}\}$ is obtained according to the following equation:

$$\boldsymbol{\psi}_{ij} = \boldsymbol{c}'_{ij} + (1 - e^{-\alpha n_{ij}})(\boldsymbol{c}'_{ij} - \boldsymbol{o}_{ij}) \qquad 1 \leq j \leq M \quad (2)$$

where α was chosen to be equal to 0.1.

3. Score integration and rejection

As depicted in Figure 1 the sorted lists of scores produced by the acoustic and visual classifiers supply an integration module, which can be regarded as a multiple classifier. This module can work using only a subset of the five lists (e.g. only the acoustic or the visual ones) and has the ability to reject an identification if it is not classified with a sufficient degree of confidence. The module operates essentially in three steps:

1. the scores are mapped into the interval (0,1) (normalization step);
2. the weighted geometric average of the input lists is computed (merging step);
3. a rejection decision is taken based on the output of a linear classifier whose domain contains both score and ranking information (rejection step).

Normalization. The signs of the distances produced by the acoustic classifiers are inverted in order to render the corresponding lists homogeneous with the visual ones (which contain correlation values). The mapping within the same interval (0,1), before merging the lists, is necessary due to the different locations and scales of the score distributions in the various lists. Furthermore, the mapping reduces intersession variability (e.g. the variability of the scores due to the variations in the input utterance length or in the illumination conditions) making the rejection decision more robust. The centers and scales of each score list are estimated by evaluating their medians and Median Absolute Deviations (MAD) respectively. Therefore if $\{s_{ij}\}_{1 \le i \le I}$ is the list of scores from classifier j, I being the number of entries in the list (i.e. the number of reference users), the corresponding normalized list is derived as:

$$s'_{ij} = \frac{1}{2}\left[\tanh\left(0.01\,\frac{(s_{ij} - \mathrm{median}_j)}{\mathrm{MAD}_j}\right) + 1\right] \in (0,1) \quad (3)$$

The median and MAD are preferable to the usual arithmetic average and standard deviation because they are less sensitive to outlying points (tails) in the score distributions, which could give erroneous estimates of the corresponding centers and scales; the hyperbolic tangent forces the scores within the (0,1) interval.

Merging. The scores $\{s'_{ij}\}$ are combined using the following formula:

$$S_i = \left(\prod_j (s'_{h_j(i)j})^{w_j}\right)^{1/\sum_j w_j} \qquad 1 \le i \le I \qquad (4)$$

where $h_j(i)$ is a function which looks, in list j, for the entry labelled as the i-th one in the integrated list $\{S_i\}$. The weights w_j are adaptively evaluated from the distributions $\{s'_{ij}\}$:

$$w_j = \frac{s'_{1j} - \mathrm{median}_j}{s'_{2j} - \mathrm{median}_j}. \qquad (5)$$

In this way the higher the ratio of the first to the second best score, the higher the weight attributed to j-th classifier.

Rejection. When a person is not identified with sufficient accuracy the system performs a rejection. This type of capability is fundamental for systems that have to control the access to security areas or services. The method we propose uses two types of information: that provided by the absolute scores contained in the final integrated list and that given by a measure of the *degree of agreement* of all the single feature classifiers. This last type of information can be quantified by means of the ranks of the entries across all the classifier outputs. In fact, if all the feature classifiers could be considered independent an unknown user would rank randomly from list to list while a known user should rank in the first positions. By looking, in a 3D space, to the score and rank distributions of acceptable and non acceptable best candidates, the use of a linear classifier to separate known people from impostors appears convenient. The linear classifier equations are as follows:

$$\sum_{i=1}^{l} w_i d_i + w_{l+1} > 0 \quad \text{accept}$$

$$\sum_{i=1}^{l} w_i d_i + w_{l+1} \le 0 \quad \text{reject} \qquad (6)$$

where $\boldsymbol{d} \in \mathcal{R}^l$ is the input vector to classify (accept or reject) and $\boldsymbol{w}$ are the coefficients of the classifier to be estimated. The vector $\boldsymbol{d}$ was chosen to be of cardinality 18, hence $l = 18$, and has the following components:

1. the integrated score, S_1, of the best candidate;
2. the normalized ratio of the first to the second best integrated score:

$$R = \frac{S_1 - S_{I/2}}{S_2 - S_{I/2}}; \qquad (7)$$

3. the minimum and maximum ranks of the first and second final best candidates (4 entries);
4. the rank standard deviation of the first and second final best candidates (2 entries);
5. the individual ranks of the first and second final best candidates (10 entries).

The vector $\boldsymbol{w}$, whose cardinality is 19, is found by minimizing the following error:

$$E = \alpha \sum_i \left(1 - \frac{1 - e^{-(\sum_{k=1}^{l} w_k p_{ik} + w_{l+1})}}{1 + e^{-(\sum_{k=1}^{l} w_k p_{ik} + w_{l+1})}}\right)^2 +$$

$$\beta \sum_j \left(1 + \frac{1 - e^{-(\sum_{k=1}^{l} w_k n_{jk} + w_{l+1})}}{1 + e^{-(\sum_{k=1}^{l} w_k n_{jk} + w_{l+1})}}\right)^2 \qquad (8)$$

where $\{\boldsymbol{p}_i\}, \{\boldsymbol{n}_j\} \in \mathcal{R}^l$ are positive and negative example vectors and α, β are the weights attributed to the corresponding error. If the system must work in a strict mode, that is no impostors may be accepted, $\beta >> \alpha$ should be used. If $\alpha = \beta = 1$, E represents the error of a linear perceptron with a symmetric sigmoidal function. The method used to derive the positive and negative examples will be explained in the following Section.

4. Experiments and results

System performance was evaluated using data acquired during three different sessions. Of the 89 people stored in the reference database, 87 have participated in at least 1 test session (no real impostors were tested). The time interval between adjacent test sessions is one week. The acquisition of the speech signal was carried out by means of an ARIEL board: sampling frequency is 16 kHz while sample accuracy is 16 bit. The acoustic distances were evaluated by means of an HEURICON SURFBOARD S/N 20 which contains six DSP ATT32C; actually one DSP transfers samples from the A/D output to the board memory, one DSP performs acoustic analysis and one DSP evaluates distances. The SURFBOARD communicates, via VME bus, with a Sun4/330 workstation. The face recognizer uses a COHU camera and communicates with the Sun4 via an FG100 board mounted on a VME slot of the workstation.

The average response time of the system is 30 seconds. A new release, which uses an Hp-735 workstation, is under way.

One of the three test sessions was used to adapt the acoustic and visual databases (in the last case the images of the session were simply added to those available); therefore, session 1 was used to adapt session 2 and session 2 to adapt session 3. As the number of interactions for each adapted session is 82 the total number of test interactions was 164. Table I reports the correct identification percentages for the acoustic instantaneous and transitional features and for their integration. Note the effectiveness of codebook adaptation. As a comparison, the performance, given in terms of correct identification, obtained on the training database for randomly generated digit strings of length 3, is 100 %. Table II reports the correct identification percentages for the three separate visual features and for their integration. In Table III the performance for the separate features (acoustic and visual) and for their integration is reported. As can be seen, the use of all the features allows a consistent increase in performance to be achieved over the separate subsystems.

TABLE I

IDENTIFICATION PERFORMANCE OF THE ACOUSTIC FEATURES.

	Ist.	Trans.	Ist. + Trans.
Not adapted codebooks	65.5 %	60.0 %	72.5 %
Adapted codebooks	82.4 %	72.8 %	88.0 %

TABLE II

IDENTIFICATION PERFORMANCE OF THE VISUAL FEATURES.

Eyes	Nose	Mouth	Eyes+Nose+Mouth
80 %	77 %	83 %	91 %

TABLE III

IDENTIFICATION PERFORMANCE OF THE SYSTEM.

Voice	Vision	Voice+Vision
88 %	91 %	98 %

To evaluate the performance of the previously described rejection strategy only 155 interactions, out of 164, containing both acoustic and visual data have been considered. As pointed out in Section 3 a set of positive examples, $\{\boldsymbol{p}_i\}$, and a set of negative ones, $\{\boldsymbol{n}_j\}$, have been used to train a linear classifier. The components of the vectors $\{\boldsymbol{p}_i\}$, $\{\boldsymbol{n}_j\}$ have been also defined in Section 3. The set $\{\boldsymbol{p}_i\}$ has been derived from interactions that provided correct classifications (correct interactions) while the set $\{\boldsymbol{n}_j\}$ was built using data belonging to wrong interactions. Furthermore, for each correct interaction, a negative example was obtained by removing the data of the interacting user from the reference database, thereby simulating an interaction with an impostor. In order to maximize the data available for the training phase a leave-one-out strategy was used. The data from all 155 interactions were grouped per user, resulting in 87 sets: one of these was used for testing the linear classifier and the remaining 86 for training. The results of the verification strategy are reported in Table IV for the separate features and for their integration. Also in this case note the effectiveness of combining the acoustic and visual data.

TABLE IV

VERIFICATION PERFORMANCE OF THE SYSTEM.

	Face	Voice	Face+Voice
Impostor accepted	4.0 %	14.0 %	0.5 %
Familiar rejected	8.0 %	27.0 %	1.5 %
Familiar misrecog.	0.5 %	1.0 %	0.0 %

5. CONCLUSIONS

A person recognition system which uses acoustic and visual features has been described. The features are processed independently by separate subsystems which provide sorted lists of scores. The lists are sent to a multiple classifier system which first normalizes and then combines them by evaluating a weighted geometric average. The normalization step produces score distributions ranging in the same interval. In this way intersession variability (i.e. the variability of the scores from trial to trial) is reduced and the successive acceptance or rejection of the best candidate identity is more robust. The acceptance/rejection decision makes use of a linear classifier.

The classification of two acoustic features is carried out by means of the Vector Quantization technique. Codebook adaptation has been also used to reduce the effects of variations between training and test conditions. Three visual features, related to regions (eyes, nose and mouth) of the face image, have been adopted. The classification of the visual features is based on crosscorrelation. The method to integrate all of five features is quite general: it can be used for a subset of them or further features (e.g. fingerprints, or other biometric data) could be added as well. The superior performance of the integrated system over the separate subsystems (acoustic or visual) has been demonstrated.

REFERENCES

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

[2] A. E. Rosenberg, F. K. Soong, *Evaluation of a Vector Quantization Talker Recognition System in Text Independent and Text Dependent Modes*, Computer Speech and Language, Vol.2, no.3,4 (1987): 143-157.

[3] R. Brunelli, T. Poggio, *Face Recognition: Features versus Templates*, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.15, no.10 (1993): 1042-1052.

[4] G.Carli, R. Gretter, *A Start-end Point Detection Algorithm for a Real-time Acoustic Front-end based on DSP32c VME Board*, ICSPAT, Boston, November 1992. pages 1011-1017.

[5] J. Makhoul, S. Roucos, H. Gish, *Vector Quantization in Speech Coding*, Proceedings of IEEE, Vol.73, no.11 (1985): 1551-1588.

[6] L. R. Rabiner, *A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition*, Proceedings of IEEE, Vol.77, no.2 (1989): 267-295.