

The Speaker Recognition problem

Il problema del riconoscimento del parlatore per uso forense¹

Giovanna Jona Lasinio^{*}, Carla Rossi[°], Tommaso Bove[#]

^{*}DSPSA, University of Rome “La Sapienza”, [°]Mathematics Department University of Rome “Tor Vergata”, [#]Italian Scientific Police

Riassunto: Nell'applicazione di metodi di riconoscimento del parlatore si procede, ai fini dell'identificazione, all'estrazione di diverse caratteristiche del segnale vocale. Tale aspetto è comune ad innumerevoli altre aree applicative nelle quali si fa uso di metodi di riconoscimento basati su classificatori multipli con input diversificati. In questo lavoro si sviluppa una rassegna dei risultati ottenuti con diverse strategie. Inoltre si propone l'applicazione al problema del riconoscimento del parlatore (in un contesto text-independent) di diversi metodi basati su misture di esperti capaci di combinare classificatori di varia natura e diverse caratteristiche in input.

Keywords: Speaker identification, acoustic features, classification, mixture of experts.

1. Introduction

It is clear from everyday experience that the speech signal carries information about its producer. There are specific circumstances in which it is crucial to be able to determine the identity of a speaker from speech alone. For instance, a witness to a crime may have heard the masked perpetrators speaking, or tape recordings of telephone conversations may have to be compared with the voice of a criminal suspect.

Speaker recognition might be defined as any activity whereby a speech sample is attributed to a person on the basis of its phonetic, acoustic or perceptual properties. A distinction can be drawn between two methods: the naive speaker recognition and the technical (automatic or semi-automatic) speaker recognition. In the first case the recognition is performed by untrained observers (a human “expert”). The decision is based on what is heard and no special techniques are involved. In the second case a great deal of work on the comparison of recordings is required. To establish identity acoustic and spectral analysis of the vocal signal is required as well as the development of *ad hoc* probabilistic/statistical methodologies (see Figure 1 for a schematic representation of the problem). In speaker recognition we can also distinguish two categories: speaker identification and speaker verification problems. In the present paper we consider only the speaker identification problem.

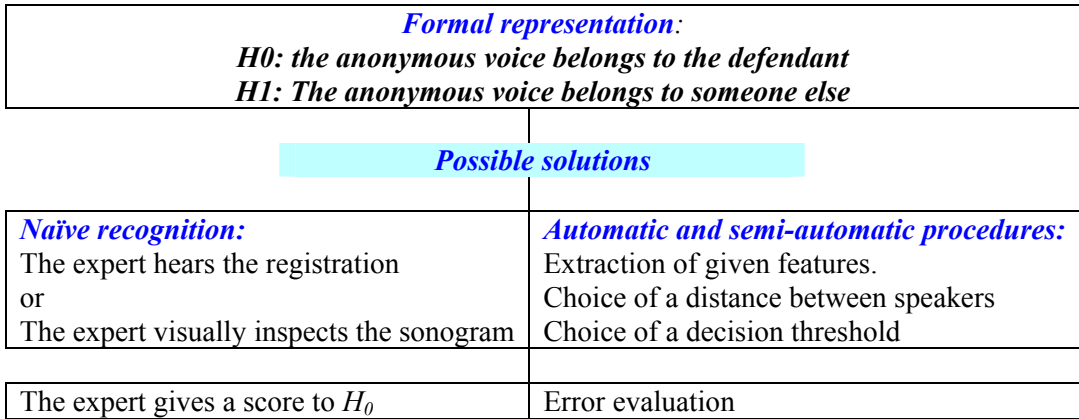
The forensic task is generally to compare two samples, one from a known speaker and one from an unknown source, and ultimately a criterion has to be applied to reach a decision as to whether the samples are “similar enough” to be from the same speaker.

The speech of an individual is not a constant, either in terms of those properties which result from the physical mechanism of speech or those which are a function of the linguistic system. If we imagine the many parameters which characterize a speaker as

¹ This work has been developed with the support of the OISIN European projects SMART1 and SMART2.

defining a location in a multiple dimensional space, we see that it is not a single point which characterizes the speaker but an area of variability, i.e. a probability distribution. There exists extensive literature dedicated to the identification problem. In the 1970's, there was a prevalent use of spectrographic analysis in voice signals; in the last few years, statistical analysis of properly defined acoustic and spectral parameters was developed, so as to take into account stochastic aspects. The almost all the statistical methodology used is of a parametric type for what concerns the estimation of the likelihoods, and it is based on classical statistical hypothesis testing for what concerns the classification problem. Several methods have been developed for identification purposes which do not relate to the forensic framework and for these techniques error rates of misidentification amounts (on average) to 20%. This rate is not acceptable for forensic purposes so the main task for all research projects developing new methodologies is to dramatically reduce the misidentification error.

Figure 1: *Schematic representation of the speaker recognition problem*



In the following we review the statistical framework in further detail and we report some proposals and results coming from two European projects on the speaker recognition problem developed in collaboration with the Italian Scientific Police. The main focus of these projects was to find automatic (and/or semiautomatic) techniques for speakers' identification. The interest in the automatic approach in forensic is mainly due to the reduced intervention of human "experts" in the recognition process and then, to a smaller degree of arbitrariness in the final result. However human intervention is still present, especially in the initial steps of most techniques when pre-treatment of the vocal signal is necessary (for instance noise filtering, signal fragmentation etc.). Here we do not go into any detail about signals pre-treatment, as our focus is on the statistical aspects of the problem.

2. The statistical-decision problem

Usually the technical speaker identification is based on the analysis of the frequency distributions of various features extracted from the vocal signal. Briefly the most commonly used are:

- *Vowels formants*: the acoustic frequencies corresponding to the local maximum intensity of the Fourier transform of the soundwave; depending on the position of these maximum points we can identify different vowels. Because of the effect of anatomical and socio-linguistic characteristics, the formants of a speaker have a particular probability distribution in a multidimensional space ‘characterising’ that speaker (Bove et al. 2003).
- *Cepstrum* (CEP) are the coefficients of the vocal signal Fourier transform.
- *Mel Cepstrum* (MCEP) are the Cepstrum coefficients on the log-scale.
- *Delta Cepstrum* (DCEP) are simply the first order differences of Cepstrum coefficients. They are a good measure of the dynamical properties of the signal.
- *Linear Prediction Coefficients* (LPC) are the coefficients of an autoregressive representation of the signal.
- *Wavelets* (complex and packets) computed on the signal. They seem to be the most promising feature for identification, however they cannot be easily used when phone calls recording have to be processed. This is due to reduced frequency band allowed by common telephones’ lines.

One of the fundamental issues in speaker identification is whether each individual in the population corresponds to a different and identifiable feature’s distribution or whether there is a wide overlap between speakers. It is well known that, unfortunately, the latter is the case. The acoustic parameters extracted should exhibit a large between speaker variability and a low within speaker variability, but for the actual data this is not the case. Several sources of variability are contributing to the within variation, such as the type and quality of recordings, the emotional condition of the speaker and his educational level etc. Furthermore we must take into account that the identification process, comprising the estimation of the individual feature distribution, has to be often based on a small sample of measures, especially when formants are used. These facts place a limitation, in principle, on the reliability of any act of speaker identification, a limitation which must always be acknowledged. It is clear, then, that a crucial point becomes the definition of a comparison term (a benchmark) for all identification procedures. Indeed the ideal validation protocol for identification procedures would include a database of vocal signals, which could be considered a “good” representation of the specific language population. It should include “enough” information on several variables:

1. Gender, age and educational level
2. Dialects and phonetic characteristics.
3. Type of recording (*recording channel*) like telephone, microphone etc.
4. Signal length and at least a few replications of the same speaker’s recording.

Such an ideal database has not yet been built (according to our knowledge) so all results that can be produced strongly depend on the type of data that are used in the checking of any identification technique’s performances.

In this work we report results obtained on a database of phone calls recorded by a radio station and some high quality phone call recordings produced by the Scientific Police. Only male voices are taken into account.

2.1. The classification problem

The main speaker recognition task needed in forensics is to decide how likely it is that two recordings are from the same person. This implies that there is the possibility for two error types. An innocent individual can be convicted, as a consequence of the similarity between the two recordings, or a guilty person can be released due to the dissimilarity.

The most important information for the decision makers is thus related to the measures of the risks related to the two errors' type. From a more formal point of view we can state the problem as follows:

- The unknown voice has to be classified in one of two alternative classes: the class H_0 of the defendant's vocal signals and the class H_1 of the vocal signals of people different from the defendant (*control set or reference population*).
- This is a hypothesis testing problem with two alternative hypotheses H_0 (null hypothesis) and H_1 (alternative hypothesis). Any procedure under this setting is built in two steps: Training (to build H_0) and Recognition (to build H_1).

In both steps the first issue to be addressed is the choice of suitable observable measures of the vocal signal (features), then the definition of a suitable test-statistics, denoted by E , (Mahalanobis distance and Mahalanobis norm, Euclidean distances etc.) is required. The last step is based on the estimation of the probability distribution functions (pdf) (possibly generalized density or probability) of the test statistics under the two hypotheses (likelihoods); $P(E/H_0)$ and $P(E/H_1)$.

In our work we developed several complementary techniques based on various features and statistical approaches. They are:

1. A distance based algorithm, in which vowels' formants and MCEP are, alternatively, analyzed (Bove et al. 2003).
2. A hierarchical clustering algorithm to be used to check how and if specific signal features are "good" measures for identification purposes (Fabi et al. 2003).
3. A Meta-Analysis approach (borrowed from pattern recognition problems) in which we combine several signal features and various classifiers (models) to obtain the final classification of a signal (Brutti et al. 2003). In what follows we are going to very briefly sketch the second method, while we devote more attention to the first and third approaches, the latter being the most promising for future developments. In all three approaches the two steps of training and recognition are performed:

■ **Training:** Given a speaker we train the system using a given set of features. The training procedure is necessary to build the class H_0 . If we use spectral type features we divide the vocal signal into frames of given length (20 - 25 ms). From each frame we extract the chosen features (F). Then, if the hierarchical clustering algorithm or the meta-analysis approach are used, for each F , we apply a classification algorithm building a frames' reference classification.

■ **Recognition:** This step is necessary to build class H_1 . We take another (or more than one) speaker and we apply the same technique chosen in the training step. We compute distances between the speaker under training and

the rest of our reference population (the reference database). The problem is the estimation of the pdf of distances. More precisely, we have to take into account that distances between vocal signals have to be evaluated considering on one hand a given speaker and the reference population, and on the other hand the speaker and himself in order to take into account the within speaker variability. Then we distinguish between *intra* and *inter* speakers distances and we have to evaluate both pdf. The final step of the recognition procedure involves the choice of a threshold to be used in the test. This also allows the computation of the two classification errors probabilities. In techniques 1 and 2 the threshold is chosen on the basis of intra and inter speakers distances pdf. In 3 an approach based on the number of “correctly” classified frames, proposed by Reynolds (1992) is adopted.

3.The Formants based algorithm

This technique is not fully automatic as formants are extracted from a vocal signal in a semi-automatic way. Formants’ extraction do not require any signal pre-treatment. The technique is based on the analysis and classification of the pitch and the first three formants of the four vowels: “a”, “e”, “i” and “o”. More precisely:

- From each recording four vowels are analysed, and four measures are extracted from each vowel, the so called formant frequencies, or simply formants (f_0 , f_1 , f_2 , f_3)
- The vowels are: “a”, “e”, “i” and “o”; the “u” has been excluded due to its low frequency in the Italian language.
- Then, each speaker is characterised, with reference to each vowel, by a matrix of four columns and a number of rows equal to the number of observations taken for that vowel.

Some formant can be missing. Most often f_0 or f_3 . The procedure imputes: (i) f_0 by averaging over the other values of f_0 for the same vowel and the same speaker (average over the first column); (ii) f_3 by multiple linear regression on the basis of f_0 , f_1 and f_2 for the same row. Using such measures, in the Training step estimates are provided for the pdf of the Mahalanobis distance of both the defendant from himself (intra-distance estimation by bootstrap resampling) and the defendant from the voices of the control set (inter-distance estimation), using the Kernel method (smoothing parameter estimated by direct plug-in) separately for each vowel. More specifically, as the problem of intra-speaker variability estimation cannot be addressed by obtaining several replications of the defendant voice signal, then it is solved, here, by bootstrapping intra-distance values. The strong correlation between formants of each vowel, suggest that the Mahalanobis distance is appropriate, since it takes into account the correlation by properly weighting the formants vector, by means of its covariance matrix. Then the univariate pdf of the Mahalanobis distance (jointly related to the four vowels) is calculated by Kernel (Gaussian) method, optimized by direct plug-in. This allows one to obtain a unique decision threshold and to calculate the prior probabilities of the two possible classification errors (false acceptance and false rejection). The prior probability of the two errors are evaluated by integrating under the densities of the intra and inter speaker distributions on the basis of the decision threshold (training phase). The posterior probability of the two errors are evaluated by integrating under the densities of

the intra and inter speaker distributions on the basis of the observed distance between the defendant and the unknown (Recognition step). In the Recognition step the sample under study (unknown speaker) is classified according to the Maximum Likelihood Classification Method and posterior probabilities (observed p-values) of the two possible classification errors are calculated. We calculate the likelihood ratio

$$\Lambda(H_0, H_1 / E) = \frac{P(E / H_0)}{P(E / H_1)}.$$

We choose the decision threshold E^* so that:

- $P(E/H_0) > P(E/H_1)$ for $E < E^*$
- $P(E/H_0) < P(E/H_1)$ for $E > E^*$

A package (written in S-plus) with user friendly interface (presently in Italian) has been implemented and made available to the Italian and Spanish Police Services. We applied the procedure to the radio-phone calls database obtaining false acceptance error rates ranging from 2 to 8% and false rejection rates from 2 to 7%. This algorithm can be used with other type of signal features and we are now testing its performances with MCEP.

4. Using other spectral features and approaches

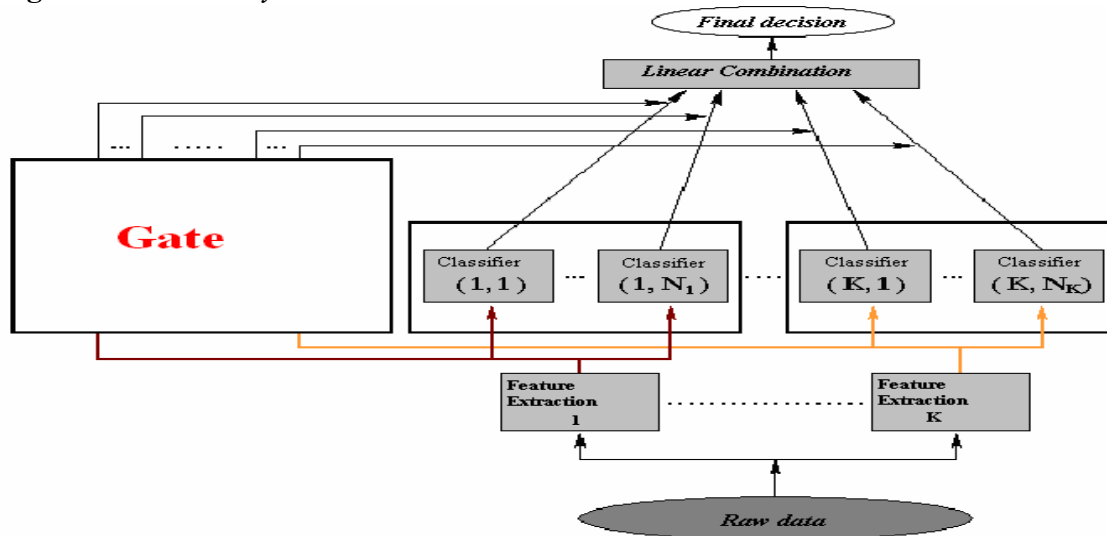
When Cepstrum coefficients or other spectral features (MCEP, DCEP, see section 2 for definitions) are used in the identification procedure we have to apply some pre-treatment to the vocal signal. We divide it into frames of a given length (ranging from 20 to 25 ms) and we use frames as the “individuals” in the classification algorithms. The frame length is chosen according to phonetic characteristics of a specific language. When using the simple clustering approach the Training step consists in building a classification of frames based on a given feature. The output of this step is the classes’ centroids vector. The Recognition step involves the evaluation of the distances between frames of signals in the reference database and centroids. In order to get “replications” of distance values we apply a procedure of sliding windows proposed by Reynolds (1992). Then we estimate non-parametrically (Gaussian kernel) the pdf of intra and inter speaker distances. Eventually, we perform the hypothesis testing as with formants. We applied the simple clustering approach to our phone-calls database, using the first 3 MCEP and DCEP, 4 LPC coefficients and LPC coefficients chosen adaptively for each frame. The most encouraging results are given by MCEP when “good” recordings are analysed (Type I error ranging from 0 to 2.6% and Type II errors ranging from 0.3% to 0.21) with a threshold chosen so to minimize Type II errors. However LPC gave interesting results when used adaptively, (Type I errors: 12% to 35%, Type II errors: 0 to 1.2%). DCEP returned very pure results as a stand-alone feature. When spectral features are used for identification purposes, the quality of the vocal signal’s recording becomes crucial. For instance if a speaker has the habit of almost yelling when talking over the phone and as the phone has a fixed passing band (from 0.3 to 3.4 kHz), the recorded signal will be *saturated*, i.e. its spectrum will cut out most of the peaks. This makes almost impossible to “properly” extract spectral features. In our experiment with the simple clustering approach, when a saturated signal was subjected to the identification process, we obtained very poor results (Type I error 78%, Type II error 0.11% with MCEP). However, as far as saturated recording are concerned, if in the signal there exist non-saturated parts, we can still identify the speaker using formants as

they can be extracted from these good elements of the spectrum. This consideration led us to consider a combination of identification techniques.

4.1 The Meta-analysis approach

Many classification approaches can be used and various features can be extracted from a signal for use in speaker identification. An idea that is currently being developed is to combine features and classifiers of different nature in one multilevel identification tool. This tool should be able to deal with poor quality recording, different type of recording, different signal length, etc. We call this procedure Comby. The general Comby scheme is illustrated in Fig. 2:

Figure 2: *The Comby Architecture*



The idea of a combination of classifier originated in pattern recognition and machine learning frameworks. Mixture of experts (ME) and their hierarchical version (HME) were introduced by Jacobs, Jordan, Nowlan and Hinton (1991) and by Jordan e Jacobs (1994), respectively. Their general aim was to find an automatic learning technique that could present a different structure (according to a given criterion) in different regions of the covariate space. For instance, in the ME architecture, experts are usually feed forward multilayers neural networks or one layer structures (for instance generalized linear models). These neural networks are trained simultaneously in such a way that each expert competes with the others in learning the input patterns and becoming “responsible” for a specific region of the covariate space. The competition is overlooked by the *gating network* which builds a sequence of weights evaluating the single expert performance on each region. This sequence of weights depends on the input’s variables. Our Comby is a generalization of ME and HME with respect to the type of architecture used in the experts combination and input structure. In order to establish the formal representation of this approach, let us illustrate in detail how the simplest architecture (ME) works. Each expert can be seen as a statistical process generating an output y according to a given probability measure $P(y|x, \Theta_e)$ depending on the input variables x and a parameters set Θ_e . The gating network can be represented as

a classifier (a parametric model) that, conditional on a parameters vector Θ_a , transforms the covariate vector $\mathbf{x}$ into a probability vector $\{g_j(\mathbf{x})\}_{j \in \{1, \dots, N(e)\}} = \{\mathbf{P}(e_j | \mathbf{x}, \Theta_a)\}_{j \in \{1, \dots, N(e)\}}$ ($N(e)$ being the number of experts and e_j being the j th expert) which reports the probability for each expert to return the “correct classification”. Then the gating network in assigning the probability vector $\mathbf{g}(\mathbf{x}) = [g_1(\mathbf{x}), \dots, g_{N(e)}(\mathbf{x})]^T$ to a point in the covariate space X generates a fuzzy partition of X into $N(e)$ regions $\{R_1, \dots, R_{N(e)}\}$. Each R_j is controlled by the expert e_j through the function $g_j(x)$. The probability that the j th expert returns the correct classification is given by

$$\mathbf{P}(e_j | \mathbf{x}, \Theta_a) \mathbf{P}(\mathbf{y} | \mathbf{x}, \Theta_{e_j}) = g_j(\mathbf{x}) \mathbf{P}(\mathbf{y} | \mathbf{x}, \Theta_{e_j}) \quad \forall j \in \{1, \dots, N(e)\},$$

the probability of observing a correct behaviour of the entire ME model is given by the following mixture (McLachlan and Peel, 2000)

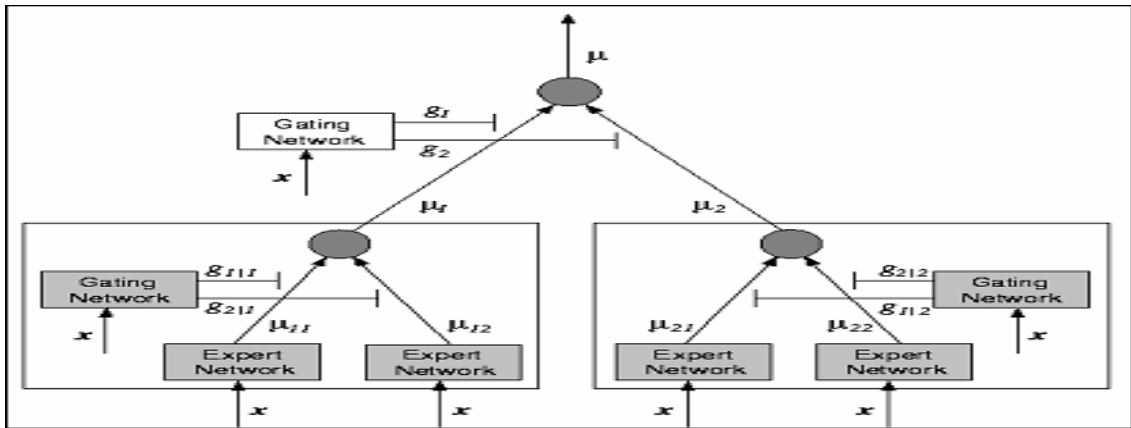
$$\mathbf{P}(\mathbf{y} | \mathbf{x}) = \sum_{j=1}^{N(e)} g_j(\mathbf{x}) \mathbf{P}(\mathbf{y} | \mathbf{x}, \Theta_{e_j}).$$

Once parameters $\{\Theta_{e_j}\}_{j=1, \dots, N(e)}$ and Θ_a are estimated we can summarize the above distribution by its mean vector:

$$\boldsymbol{\mu} = \sum_{j=1}^{N(e)} g_j(\mathbf{x}) \boldsymbol{\mu}_j.$$

The above architecture is completely general, we can specify any type of model for the experts and gating. When introducing a hierarchical part on ME we obtain the HME architecture. In this case the partitioning procedure is iterated recursively as well as the dynamical assignement of an expert (locally optimal) to a given region. This implies a hierarchical parametrization of $\mathbf{P}(e | \mathbf{x}, \Theta_a)$ through a nested partition sequence equivalent to a decision tree (Hastie *et al.*, 2001), each level corresponding to a tree's branch (Fig. 3)

Figure 3: HME architecture with two levels and branching factor 2 represented as a modular neural network (adapted from Jordan e Jacobs (1994))



The main drawback of an ME or HME architecture as described above is that we have to feed the system with a huge vector of input's variables (various acoustic features). Solutions based on standard dimensional reduction techniques (such as Principal Components Analysis) have been proposed, however, the strong correlation between frames is a severe counterindication to the use of these techniques. The following, more appropriate solution, to this dimensional problem has been proposed by Chen (1998b):

1. introduce a set of gating networks (the gate in Fig. 2) which is able to weight experts outputs on the basis of their performances on K vectors of features $\{\mathbf{x}_1, \dots, \mathbf{x}_k\}$, analysed separately.
2. To each feature vector a pool of N_j ($j=1, \dots, k$) experts is assigned.

The resulting model solves the dimensionality problem by giving the Gate the task of harmonizing the pairs, expert-feature outputs, through the evaluation of the weights system. Our proposal, though similar to Chen's, makes use of different type of models. Our Gate is a multivariate logistic model (Moerland, 2000) and the experts are of various nature (for instance Adaptive Gaussian Mixture Models, Mixture of Probabilistic Factor Analysis etc.). We carried out some experiments applying the following procedure.

As in the previous section we divide the defendant vocal signal into frames of 23 ms. From each frame we extract a set of features. We train our system by estimating parameters for each of the classifier-models introduced. These parameters are the output of the Training step. Then we perform a Recognition step in which for each signal in the reference dataset we divide it into frames and we extract the same features as in the training step obtaining, for each feature, a sequence of vectors $\{\mathbf{x}_1, \dots, \mathbf{x}_v, \dots\}$. From $\{\mathbf{x}_1, \dots, \mathbf{x}_v, \dots\}$ we build T vectors of overlapping segments $\mathbf{x}^{(h)} = \{\mathbf{x}_h, \mathbf{x}_{h+1}, \dots, \mathbf{x}_{h+(T-1)}\}$ (following Reynolds (1992)). Each $\mathbf{x}^{(h)}$ is seen as a separate registration and fed to the model(s) estimated in the previous phase. For each vector in the segment the system produces a classification and we accept an identification only if more than 50% of the above classification agree in assigning segments to the defendant.

We report results from the application to the same dataset of vocal signals used in previous sections. For each feature we have 4 "experts" (models):

- *Adaptive Gaussian Mixture Model* (AGMM) (Priebe, 1994)

The general idea underlying this model (Reynolds, 1992; Nedic and Bourlard, 2000) is to estimate the pdf of extracted features through an adaptive mixture of Gaussian pdf and then to perform a likelihood-ratio test in order to verify the null hypothesis that two vocal signal come from the same speaker. Parameter estimation is carried out through a specific adaptation of the EM algorithm.

- | | |
|--|--|
| <ul style="list-style-type: none"> • <i>Mixture of Probabilistic Principal Component Analysers</i> (MPPCA) (Bishop and Tipping, 1999) | <ul style="list-style-type: none"> • <i>Mixture of Probabilistic Factor Analysers</i> (MPFA) (Gharhamani and Hinton 1996) |
|--|--|

As for the AGMM these two classifiers are based on particular type of finite mixtures and they adopt the same testing strategy illustrated above. Both models are given by

convex linear combinations of probability measures induced by linear models defined as follows:

$$\mathbf{x} = \mathbf{W}\mathbf{z} + \boldsymbol{\mu} + \boldsymbol{\varepsilon} \text{ in which : } \begin{cases} \mathbf{z} \sim N(\mathbf{0}, \mathbf{I}_l) \text{ where } l < \dim(\mathbf{x}) = d \text{ are latent variables} \\ \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \mathbf{R}) \\ \mathbf{R} = \sigma^2 \mathbf{I}_d \text{ for PPCA or } \mathbf{R} \text{ diagonal for PFA} \end{cases}$$

The aim is to find a projection of the features vector $\mathbf{x}$ while keeping the maximum possible of the information on the correlation structure. Estimation is again carried out by EM (see McLachlan and Peel, 2000; McLachlan and Krishnan 1997).

- *Multi Layers Perceptron* (MLP) (Bishop, 1995)

This one is the only supervised classifier we use in the Comby architecture. It is a perceptron with one hidden state and 20 units, with a softmax activation function (multivariate logistic). The training is performed by backpropagation based on a non linear optimization algorithm (Scaled Conjugated Gradient, Bishop, 1995).

In order to evaluate the Comby performances we compute the following indices:

$$IR \text{ (identification rate)} = 100 N_{Ih}/N_h$$

where N_{Ih} is the number of correctly identified segments $\mathbf{x}^{(h)}$ and N_h is the total number of $\mathbf{x}^{(h)}$.

$$SR \text{ (substitution rate)} = 100 N_{nlh}/N_h$$

where N_{nlh} is the number of incorrectly classified segments $\mathbf{x}^{(h)}$.

$$RR \text{ (refusal rate)} = 100 - [IR + SR]; AR \text{ (Reliability rate)} = 100[IR/(100-SR)].$$

Results are collected in table 1. For the training of ME and HME following the literature suggestions, an early stopping strategy in both steps, Training and Recognition, has been adopted. Giulio's and Lorenzo's vocal recordings are of good quality, they speak with a normal voice level and they do not show any particular emotion while talking. Tony's signal is saturated as this speaker is usually very emotional and he speaks in a loud voice over the phone.

Table 1: Results from Comby and two different architectures of combined classifiers

	COMBY				ME				HME			
	5 hours computation				(Early-Stopping) 7.5 hours computation				(Early-Stopping) 11 hours computation			
	IR	IS	RR	AR	IR	IS	RR	AR	IR	IS	RR	AR
Giulio	98.8	0.5	0.7	99.5	92.5	0.0	7.5	100.0	94.7	2.7	2.6	97.2
Lorenzo	91.0	5.7	3.3	94.1	90.0	5.3	4.7	94.4	90.5	1.1	8.4	98.8
Tony	6.2	8.9	84.9	41.0	1.8	6.6	91.6	21.4	1.8	8.5	89.7	17.5
Average	65.3	5.0	29.6	78.2	61.4	3.9	34.6	71.9	62.3	4.1	33.6	71.2

Comby's results are encouraging. With the proposed models it identifies a saturated speaker more often than the simple clustering approach. We haven't tried yet to include formants which shown to be the "best" feature when saturated signals are treated, the reason being that we do not yet have an automatic procedure for their extraction. However in Comby and other combined architectures, computational problems are evident. At present we are working on a new Comby's version which should be more efficient. An important development of the Comby software will be to include some Hidden Markov models proposed in the literature as speakers identifiers.

6. Conclusions

The speaker identification problem is a challenging forensic problem both for statisticians and signal processing experts. The extreme within variability of vocal signals, the difficulty to identify a small number of features able to capture and "smooth" such variability are some of the points that do not allow for a theoretical "optimal" solution. The only way to proceed is by direct experimentation. The building of an "appropriate" identification technique requires the combination of several skills, here we mainly illustrated the statistical problems while we skipped all the "engineering" and software problems involved. Digital audio signals need a careful treatment before we can apply our techniques: noise filtering, signal enhancement etc. may be necessary. These too are not yet standardized procedures and they have a strong impact on the subsequent statistical analysis. The computational effort needed to apply any identification technique, is another serious problem that cannot be solved by statisticians alone. Efficient implementation of the proposed technique is still a relevant issue to be addressed. The following issues will be considered in future research: (a) We need to test the Comby software more carefully on an extended database. (b) Further classifiers-models came to our attention from the literature on this subject. A relevant suggestion is the use of appropriate Hidden Markov Models as classifiers. (c) It was pointed out that the "best" benchmark for any speaker recognition tool would be an "ideal database" built according to the phonetic structure of the reference population. Up to now this ideal database has not been built for the Italian language. We are already collaborating with phonetic experts in order to establish the structure of this database. An ideal database is under construction in Spain and we expect to be able to work on it by the end of 2004.

To conclude we have to mention that commercial software packages are already available for speaker identification purposes, however, they are not built for forensic use, for instance, they seem very sensitive to the within speakers variability and they need accurate testing before they can be used in court.

References

- Aitken C.G.G., (1995) *Statistics and the evaluation of evidence in forensic science*, Wiley, Chichester, U.K..
- Bishop, C. M. (1995) *Neural Networks and Pattern Recognition*. Clarendon Press, Oxford.
- Bove T., Forte A., Giua P.E., Rossi C. (2003) Un metodo statistico per il riconoscimento del parlatore basato sull'analisi delle formanti, accepted *Statistica*

- Brutti P., Fabi F., Jona Lasinio G. (2003) Una proposta di meta-analisi basata sulla combinazione di classificatori per il problema del riconoscimento del parlatore, accepted *Statistica*
- Chen K., Chi H. (1998b) A method of combining multiple probabilistic classifiers through soft competition on different feature sets, *Neurocomputing* 20(1-3), pp. 227-252
- Cohen A. & Froind I., (1989) On text independent speaker identification using a quadratic classifier with optimal features, *Speech Communication* 8, 35-44.
- Doddington G.R., (1985) Speaker Recognition, Identifying People by their Voices, *Proceedings IEEE* 76 no 11.
- Fabi F., Brutti P., Jona Lasinio G. (2003) Una metodologia per il riconoscimento del parlatore basata sulla frammentazione del segnale e sulla classificazione dei frammenti, accepted *Statistica*.
- Ghahramani, Z. and Hinton, G. E. (1996) *The EM algorithm for mixtures of factor analyzers*. Technical Report CRG-TR-96-1, Department of Computer Science, University of Toronto, Toronto, Ontario.
- Hastie T., Tibshirani R., Friedman J. (2001) *The elements of statistical learning: Data mining, Inference and Prediction*, Springer-Verlag.
- Jacobs R.A., Jordan M.I., Nowlan S.J., Hinton G.E. (1991) Adaptive mixtures of local experts, *Neural Computation*, 3(1), pp. 79-87.
- Jordan M.I., Jacobs R.A. (1994) Hierarchical mixtures of experts and EM algorithm, *Neural Computation*, 6(2), pp. 181-214.
- McLachlan G., Krishnan T. (1997) *The EM algorithm and Extensions*. John Wiley & Sons.
- McLachlan G., Peel D. (2000) *Finite Mixture Models*, John Wiley & Sons, Inc.
- Moerland P. (2000) *Mixtures Models for Unsupervised and Supervised Learning*, PhD-Thesis IDIAP.
- Nedic, B., Bourlard, H. (2000) Recent developments in speaker verification at IDIAP. Tech. Rep. IDIAP-RR 00-26.
- Priebe, C.E. (1994) Adaptive mixture density estimation. *JASA*, 89, 796-806.
- Reynolds, D. A. (1992) *A Gaussian mixture modelling approach to text-independent speaker identification*. Ph.D. thesis. Department of Electrical Engineering, Georgia Institute of Technology.
- Robertson B. & Vignaux G.A., (1995) *Interpreting Evidence: evaluating forensic science in the courtroom*, Wiley, NY.
- Rossi C., (1996a) La Statistica per il Riconoscimento del Parlatore: aspetti critici e utilizzo problematico, *Atti delle VI giornate di studio del gruppo di fonetica sperimentale*, Fondazione 'Ugo Bordini', Rome, Italy, 79-86.
- Rossi C., (1996b) *Il problema di decisione nell'identificazione del parlatore*, Caratterizzazione del parlatore, F. Fedi & A. Paoloni eds., Fondazione 'Ugo Bordini', Rome, Italy, 173-176.
- Tipping, M. E. and Bishop, C. M. (1997) *Mixtures of probabilistic principal component analysers*. Tech. Rep. NCRG-97-003, Department of Computer Science and Applied Mathematics, Aston University, Birmingham, UK.