STATISTICA, anno LXII, n. 3, 2002

UN METODO STATISTICO PER IL RICONOSCIMENTO DEL PARLATORE BASATO SULL'ANALISI DELLE FORMANTI

T. Bove, P. E. Giua, A. Forte, C. Rossi

1. INTRODUZIONE: IL RICONOSCIMENTO DEL PARLATORE IN CAMPO FORENSE

È chiaro dall'esperienza di tutti i giorni che il segnale vocale contiene informazioni sulla persona che lo ha emesso. Ci sono determinate circostanze nelle quali è cruciale essere in grado di determinare univocamente l' identità di una persona attraverso il solo segnale vocale. Per esempio, un testimone di un crimine potrebbe aver udito parlare il criminale mascherato, oppure la registrazione di una conversazione telefonica può dover essere comparata con la voce di un sospetto criminale. Nel riconoscimento del parlatore, un campione di voce è attribuito ad una persona in base alle sue caratteristiche acustiche e fonetiche. Lo strumento di comparazione può essere, o l'orecchio, e in questo caso il riconoscimento avviene mediante ripetute prove di ascolto fatte da un gruppo di persone qualificate, o un sistema automatico, che acquisisce il segnale vocale, lo analizza, ne estrae opportuni parametri ed infine lo elabora mediante metodologie probabilistico/statistiche. Esiste un'ampia letteratura dedicata al problema dell'identificazione del parlatore (Bimbot *et al.,* 1995; Furui, 1989; Matsui e Furui, 1994; Ibba e Paoloni, 1993). Negli anni '70 era prevalente l'uso dell'analisi spettrografica dei segnali vocali, invece negli anni recenti l'analisi statistica di parametri acustici opportunamente definiti ha avuto ampio sviluppo in modo da prendere in considerazione aspetti stocastici legati alla variabilità. Il parlato di un individuo, infatti, non è sempre uguale, sia per quanto riguarda le proprietà legate al meccanismo fisico di produzione della voce, sia per quanto riguarda quelle legate al sistema linguistico.

Se noi immaginiamo la molteplicità delle misure che caratterizzano un parlatore come un punto in uno spazio multidimensionale, vediamo che non è un singolo punto che caratterizza un parlatore ma un'area di variabilità, cioè una distribuzione di probabilità.

2. PROVE E TESTIMONI SCIENTIFICI

Tra gli scopi della "Scienza Forense", intendendo con questo tutte le testimonianze di tipo scientifico (analisi biochimiche su reperti, analisi tecniche di ogni

tipo su registrazioni, impronte digitali, foto, impronte di DNA ecc.) fornite in aula di tribunale si possono annoverare:

– stabilire che una persona era in un certo luogo in un dato periodo di tempo;
– stabilire che una persona ha compiuto una certa attività;
– stabilire che una certa azione è stata compiuta con un dato strumento;
– stabilire l'esistenza di una relazione (per esempio di paternità) tra due persone.

Lo scopo generale dell'applicazione delle tecniche scientifiche può essere ricondotto all'identificazione di un soggetto, di solito imputato di un ben determinato crimine. Vengono utilizzati, pertanto, dei sistemi scientifici di identificazione.

Un sistema scientifico di identificazione dovrebbe possedere alcune specifiche caratteristiche:

– deve identificare caratteristiche possedute in modo "unico" da ogni individuo;
– le caratteristiche di interesse devono essere invarianti nel tempo;
– tali caratteristiche devono essere rilevate in modo univoco, cosicché due diversi esperti in diversi momenti raggiungano le medesime conclusioni;
– deve essere in grado di posizionare in modo corretto il soggetto sulla scena del crimine;
– deve essere ragionevolmente semplice e poco costoso.

Naturalmente i sistemi reali non possiedono i requisiti elencati. La conseguenza è l'introduzione di elementi di "incertezza".

Ne segue che, in generale, un testimone "scientifico" non sarà in grado di dire se due campioni provengano con certezza da uno stesso individuo. Spesso si potrà solo giungere alla valutazione della probabilità (o della verosimiglianza) che i due campioni provengano dalla stessa persona.

3. L'INTERPRETAZIONE DELLE PROVE SCIENTIFICHE: IL RAPPORTO DELLE VEROSIMIGLIANZE

Una prova scientifica generalmente implica che lo scienziato forense effettui osservazioni su alcuni aspetti relativi al caso in esame e, sulla base di precedenti esperienze, effettui dei procedimenti inferenziali e ne riferisca alla Corte (Aitken, 1995; Robertson e Vignaux, 1995). Il minimo requisito richiesto ad una prova scientifica è che sia rilevante. Secondo la definizione in uso nelle Corti di Giustizia degli Stati Uniti:

"prova rilevante significa prova che abbia la tendenza a rendere ogni fatto che sia conseguenza dell'azione criminale più probabile o meno probabile di quanto non sia senza la produzione di quella prova".

In parole più precise, l'evento (o la misura) su cui effettuare le osservazioni, usato come prova, deve essere correlato con il fatto su cui si sta indagando e con le sue conseguenze (altri fatti derivanti dall'azione criminale).

Da questo deriva che una prova ideale deve produrre probabilità condizionate pari a 0 o 1. In generale raramente si hanno a disposizione prove ideali, occorrerà, pertanto, ragionare su probabilità condizionate nell'intervallo aperto (0,1). *Una*

prova rilevante è un evento che è *più probabile* quando l'azione (o ipotesi) che si tenta di provare è vera che quando è falsa (o anche viceversa).

Una prova forte è, invece, un evento che è *molto più probabile* quando l'azione (o ipotesi) che si tenta di provare è vera che quando è falsa (o anche viceversa). Ne segue che siamo in grado di calcolare il valore probativo o forza di un'evidenza solo avendo osservato la frequenza del verificarsi della prova quando l'ipotesi è verificata e quando non lo è. *Non è possibile valutare la forza di una prova osservando solo quei casi in cui l'ipotesi è verificata*. In ogni caso, alla fine del procedimento inferenziale saremo in grado di valutare la probabilità della prova scientifica se l'ipotesi è vera e non quale sia la probabilità che l'ipotesi sia vera. La misura relativa del valore della prova V(E) è, allora, naturalmente data dall'odds ratio:

$$V(E)=P(E|H_1)/P(E|H_2) \quad (1)$$

dove $P(E|H_1)$ è la probabilità condizionata della prova E se è vera l'azione o ipotesi denotata con H_1, mentre $P(E|H_2)$ è la probabilità condizionata di E se non è vera H_1, ovvero se è vera qualche ipotesi alternativa definita da H_2.

Per calcolare correttamente il numeratore del rapporto di verosimiglianza si deve valutare la probabilità dell'elemento di prova E nell'ipotesi (nulla) che l'individuo sotto processo abbia effettivamente commesso l'azione criminosa (tesi accusatoria). Tale valutazione deve essere sempre confrontata con il denominatore, che misura la probabilità di quello stesso elemento di prova nell'ipotesi (alternativa) che sia vero quanto sostenuto dalla difesa (tesi difensiva). Tale procedimento non implica l'uso delle probabilità a priori di nessuna delle due ipotesi. Se un giudice ritiene che un particolare elemento di prova sia "incriminante" questo può solo significare che tale elemento risulta più probabile nell'ipotesi sostenuta dall'accusa che in quella sostenuta dalla difesa.

Per concludere:

– uno scienziato forense non può dire quanto sia probabile che la tesi sostenuta dall'accusa (ipotesi nulla) sia vera, ma solo quanto risulti probabile una certa prova nell'ipotesi che tale tesi sia vera sempre e solo in rapporto a quanto sia probabile la stessa prova, essendo vera la tesi della difesa;

– l'indicatore che quantifica il confronto è solo il rapporto di verosimiglianza;

– in linea di principio una prova risulta rilevante se tale rapporto è maggiore o minore di 1. In caso contrario è irrilevante perché ugualmente probabile nelle due ipotesi;

– sebbene rilevante una prova può essere esclusa o per incompatibilità con altre o perché il suo valore probativo (misurato dal rapporto di verosimiglianza) non è sufficiente per bilanciare il costo della sua ammissione in termini di tempo, denaro, confusione o pregiudizio.

Occorre sottolineare l'impossibilità di effettuare un'inferenza corretta in mancanza della valutazione appropriata del denominatore del rapporto di verosimiglianza. Questo implica la corretta definizione dell'ipotesi alternativa (ipotesi difensiva), la corretta definizione dei modelli statistici delle verosimiglianze (numeratore e denominatore) relative alla prova considerata e la stima degli eventuali parametri incogniti che compaiono in tali funzioni. È bene ribadire che, in man-

canza di un opportuno modello per il denominatore, ogni inferenza è priva di senso e non è neppure possibile la decisione in merito alla rilevanza della prova considerata. Valori anche grandi del numeratore, infatti, possono dar luogo a rapporti di verosimiglianza non significativamente diversi da 1 e, persino inferiori a tale valore, ovvero in favore della difesa. È immediata l'estensione dal caso in cui la prova sia un evento al caso in cui sia una misura (continua o discreta), dove alla probabilità dell'evento si sostituisca la distribuzione di probabilità della misura (eventualmente la densità o densità generalizzata).

4. L'IPOTESI ALTERNATIVA

Ricordiamo ancora che il teste scientifico non deve e, in realtà, non può valutare la probabilità di una certa ipotesi (accusatoria o difensiva), ma solamente fornire correttamente il valore del rapporto di verosimiglianza della (o delle) prove.

La rilevanza di una prova dipende interamente dalla sua capacità di discriminare tra l'ipotesi accusatoria e un'ipotesi alternativa scelta dalla difesa. Tale ipotesi deve essere in qualche modo una negazione della prima (incompatibilità), ma la sua definizione, in termini concreti e matematicamente e statisticamente trattabili, è un problema delicato. Occorre ricordare che non è possibile utilizzare nessuna prova scientifica se non si è in grado di valutare correttamente la sua probabilità (o distribuzione) subordinatamente sia all'ipotesi accusatoria che all'ipotesi alternativa perché il rapporto di verosimiglianza le prevede entrambe. Precisiamo che è molto difficile, se non impossibile, valutare le probabilità della prova considerata avendo a disposizione un'ipotesi alternativa vaga o mal definita. Entrambe le ipotesi a confronto devono essere chiare e ben definite ed entrambe devono entrare esplicitamente nelle valutazioni probabilistiche.

Tale esigenza, dal punto di vista matematico-statistico, implica ed è implicata dalla scelta di ipotesi H_1 e H_2 incompatibili e, pur di definire opportunamente lo spazio campione, esaustive. Qualsiasi altra scelta non permette la corretta valutazione del rapporto di verosimiglianza.

Per concludere:

– una singola prova non può da sola provare un'ipotesi considerata in assoluto, può soltanto discriminare tra ipotesi diverse;

– il valore di una prova in sostegno all'accusa può essere diverso in relazione a diverse ipotesi difensive;

– il punto fondamentale è: "qual è l'ipotesi alternativa appropriata?"

– Nel caso in cui una traccia sia stata lasciata dal colpevole sulla scena del crimine, il rapporto di verosimiglianza misura la probabilità della traccia nel caso che sia stato l'imputato a lasciarla, divisa per la probabilità della traccia nel caso in cui a lasciarla sia stato qualcun altro. Come vada scelto tale ipotetico individuo alternativo dipende dallo stato di conoscenza generale sul crimine e dalla linea della difesa;

– può essere necessario confrontare l'unica tesi accusatoria con più tesi difensive che danno luogo a diversi valori del rapporto di verosimiglianza;

– le due ipotesi via via a confronto devono essere incompatibili;
– in un corretto schema procedurale tutte le ipotesi da considerare dovrebbero essere note ai testi scientifici in anticipo in modo da consentire la stima di tutte le probabilità condizionate e il calcolo di tutti i corrispondenti rapporti di verosimiglianza.

5. LE REGISTRAZIONI TELEFONICHE E LE TECNICHE DI RICONOSCIMENTO DEL PARLATORE

Nei processi giudiziari, una possibile prova scientifica per identificare le persone è spesso la registrazione telefonica.

Il problema dell'identificazione di un parlatore usando la sua voce trova applicazioni in molti campi, ma in ambito forense è reso più difficile dal fatto che le telefonate registrate e messe a disposizione degli esperti spesso, ma non sempre come la cronaca recente ci ha insegnato, sono anonime e con fini illeciti. Pertanto hanno dimensioni limitate e qualità scadente.

Inoltre, la qualità dei dati di confronto è altrettanto mediocre: infatti, raramente le persone sospettate di un crimine sono disposte a rilasciare saggi di voce nel modo richiesto e per il tempo necessario, da usare per il confronto.

In più, essi possono introdurre volontariamente alterazioni nella voce per ridurre la somiglianza con il campione iniziale e complicare l'eventuale riconoscimento.

A tutto ciò si sommano disturbi della linea telefonica, rumori ed altre condizioni ambientali che spesso non sono riproducibili e rendono ancor più problematica l'identificazione.

Nonostante queste difficoltà, data l'importanza della prova, sono state comunque sviluppate varie tecniche per riconoscere un parlatore. Si tratta di tecniche di riconoscimento che possono essere "soggettive" o automatiche.

Con l'attributo "soggettivo" si indicano tutti quei metodi fondati sulle abilità sensoriali umane: tra questi esistono molti metodi che fanno uso di prove d'ascolto e sono basati sulla capacità dell'orecchio, eventualmente addestrato in modo specifico, di confrontare segnali vocali di diversa origine e stabilirne la compatibilità. Altri possibili metodi soggettivi consistono nel confrontare sonogrammi (suono accompagnato dal diagramma dell'energia) e sono dunque basati sul giudizio visivo di un esperto.

La principale obiezione sollevata contro questi tipi di tecniche è proprio la loro soggettività: i risultati ed i giudizi che da esse derivano, non possono essere quantificati con delle metriche riproducibili ed indipendenti dal soggetto che li ha espressi. Di conseguenza, non è facile calcolare le probabilità degli errori di prima e seconda specie.

Tali errori possono essere così formulati:

1. rifiutare l'ipotesi che la voce registrata durante l'intercettazione e la voce registrata per il confronto non siano uguali, quando in realtà lo sono (1ª specie);

2. accettare l'ipotesi che le due voci registrate siano uguali quando in realtà non lo sono (2ª specie).

Queste formulazioni si giustificano ricordando che l'errore di seconda specie, in una analisi inferenziale, è considerato il più grave tra i due: è un errore, infatti, più grave in generale condannare un innocente che assolvere un colpevole.

Accanto alle tecniche soggettive, esistono poi metodi automatici di riconoscimento, basati sull'analisi di parametri acustici estratti direttamente dal segnale. Tale estrazione è fatta strumentalmente in modo automatico mediante software appositamente predisposto con l'eventuale intervento attivo di un operatore opportunamente addestrato.

In quest'ultimo caso, si dovrebbe parlare più precisamente di metodi semi-automatici, in quanto il processo di estrazione dei parametri e, di conseguenza, il risultato dell'estrazione stessa, sono soggetti al giudizio dell'esperto che li ha effettuati.

La differenza tra i metodi semi-automatici e quelli soggettivi sta nel fatto che usando i primi è possibile comunque valutare abbastanza facilmente la probabilità di errore e la loro affidabilità; inoltre i risultati sono riproducibili.

Il processo di decisione consiste nello scegliere tra le due ipotesi H_1 e H_2 precedentemente definite e nel valutare le probabilità di errore.

Nel fare ciò, bisogna distinguere tra due possibili alternative note a priori vedi commento pagina precedente:

1. la voce anonima appartiene sicuramente ad uno dei sospettati;
2. la voce anonima può non appartenere ad alcun individuo del gruppo.

Nel primo caso si parla di test di tipo "chiuso": è il più favorevole dato che ammette solo esiti positivi (ossia l'associazione del campione vocale con una delle voci di confronto).

La seconda alternativa è quella che si presenta più frequentemente nella realtà e consente di applicare solo test di tipo "aperto": in essi va tenuta in considerazione l'ipotesi che la voce non appartenga ad alcuno dei sospettati.

Se si sfruttano tecniche automatiche o semi-automatiche bisogna stabilire una misura di distanza o di dissimilarità tra parlatori; viene quindi stimata la distribuzione della distanza del parlatore anonimo da se stesso e la distribuzione della distanza dell'anonimo da tutti gli altri parlatori inseriti in una Base-Dati di confronto che si possa considerare rappresentativa della popolazione di riferimento cui si ritiene appartenga la voce dell'anonimo. Sulla base di tali distribuzioni si definisce un valore soglia, che identifica il criterio di decisione. Infine, confrontando la distanza tra voce anonima e voce dell'indiziato con tale soglia, si stabilisce se le due voci sono uguali o diverse, avendo definito a priori il criterio di decisione (scelta di un valore massimo ammissibile per l'errore di seconda specie, scelta di un livello fissato per il rapporto di verosimiglianza…).

Qualsiasi procedura presuppone, comunque, due fasi fondamentali.

1. La fase di addestramento del sistema, che consiste nell'insieme di procedure che permettono di caratterizzare la variabilità statistica della voce dell'indiziato (o dell'anonimo) rispetto a quella dei parlatori della popolazione di riferimento e, di conseguenza, stimare le due verosimiglianze che sono alla base del processo di decisione e determinare la soglia di decisione in base al criterio adottato.

2. La fase di riconoscimento che consiste nella procedura che permette di classificare probabilisticamente la voce dell'anonimo (o dell'indiziato) rispetto alle due classi ottenute, a partire dalla soglia di decisione, nella fase di addestramento.

6. IL RICONOSCIMENTO BASATO SULLE FORMANTI

Un metodo per l'identificazione del parlatore è basato sull'analisi delle formanti delle vocali. Le formanti, abbreviazione di frequenze formanti, sono ottenute dallo spettro di frequenza del segnale. Le frequenze formanti corrispondono ai massimi relativi dell'inviluppo dello spettro. Le formanti permettono di distinguere le vocali fra di loro. Nel caso di analisi di registrazioni su linea telefonica, quali quelle considerate nel seguito, è possibile rilevare solo la frequenza fondamentale della voce e le prime tre formanti per ogni vocale. Ogni parlatore è caratterizzato, in relazione ad una singola vocale, pertanto, da una matrice di quattro colonne ed un numero di righe pari al numero di osservazioni per quella particolare vocale. In questo modo, egli è rappresentato globalmente da una distribuzione di frequenze in uno spazio a 16 dimensioni (4 per ogni vocale), essendo in generale esclusa la vocale "u" scarsamente rappresentata nel parlato in lingua italiana. Scegliendo come regola di decisione, per esempio, un approccio basato sul criterio di Verosimiglianza, la classificazione avverrà sulla base della distribuzione di opportuni indici di dissomiglianza o di distanza costruiti a partire dai dati disponibili.

Considereremo, per esemplificare, solo due ipotesi alternative; la prima è l'usuale ipotesi accusatoria che il parlatore anonimo e l'imputato siano caratterizzati dalla stessa distribuzione statistica, l'ipotesi difensiva sarà quella che le due distribuzioni statistiche del parlatore anonimo e dell'imputato siano diverse, come sempre accade nei procedimenti. Questo confronto si basa su un opportuno indice di distanza. Preliminarmente, pertanto, è necessario scegliere tale indice di cui valutare la rilevanza come "prova scientifica. Osserviamo che si parla di identità o di differenza delle distribuzioni e non degli individui.

6.1. *La fase di addestramento*

Prima di procedere alla costruzione dei modelli statistici necessari per la vera e propria fase di addestramento, occorre prendere in considerazione eventuali dati mancanti legati a svariati problemi di tipo tecnico, come, per esempio, i casi di saturazione del segnale (Rosati, 2001) e procedere alla ricostruzione mediante imputazione.

6.1.1. *La pre-elaborazione per la ricostruzione dei dati mancanti*

Possono verificarsi essenzialmente due situazioni diverse:

1. può non essere rilevata la frequenza fondamentale f_0;
2. può non essere rilevata la formante f_3;

Le due situazioni possono verificarsi contemporaneamente e in un numero consistente di ripetizioni della vocale o in un numero moderato.

La figura 1 mostra una porzione di una matrice di dati in cui sono evidenti entrambe le situazioni.

Un'ampia sperimentazione (Rosati, 2001) suggerisce di procedere come segue.

1. Se sono mancanti alcuni dati relativi alla frequenza fondamentale (non numerosi) si imputa la stessa valutando la media sui dati rilevati della stessa colonna della matrice.
2. Se i dati mancanti sono numerosi si esclude la colonna e si riporta la matrice dei dati ad una di dimensione inferiore.
3. Se sono mancanti alcuni dati relativi alla formante f_3 (non numerosi) si imputa la stessa valutando attraverso regressione multipla la media condizionata ai valori delle altre formanti.
4. Se i dati mancanti sono numerosi si esclude la colonna e si riporta la matrice dei dati ad una di dimensione inferiore.

In relazione al caso presentato in figura 1 è opportuno imputare i dati mancanti sulla prima colonna ed escludere la terza dalle analisi.

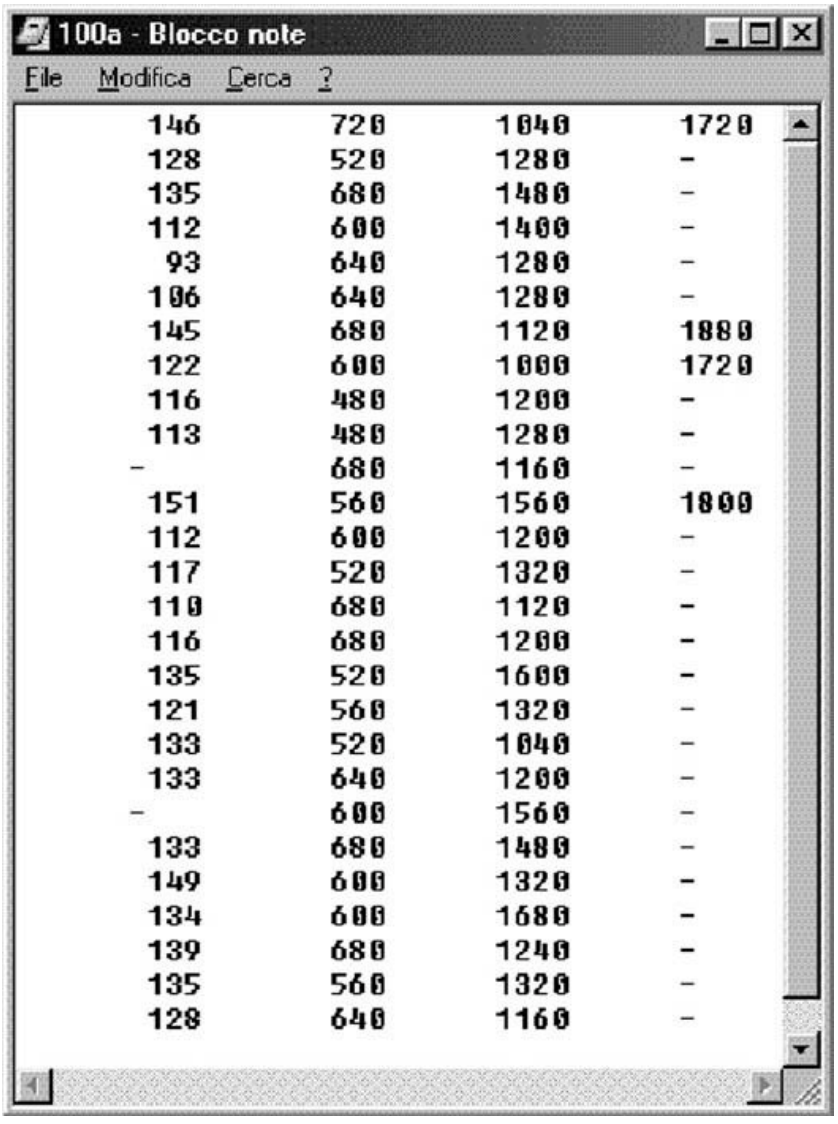

Figura 1 – Matrice dei dati incompleta sulla colonna relativa alla frequenza fondamentale (pochi dati mancanti) e alla formante f_3 (numerosi dati mancati).

6.1.2. *La modellizzazione e la stima delle verosimiglianze*

Per affrontare il problema occorre arrivare al calcolo del rapporto di verosimiglianza delle due ipotesi considerate e, per prima cosa, occorre modellare le due funzioni di verosimiglianza da utilizzare per la costruzione del rapporto, una volta scelto l'indice di distanza da utilizzare. In precedenti lavori (Calvani, 1996; Forte, 1998; Ghizzoni, 1999; Rossi, 1996) si è ampiamente dimostrata l'adeguatezza della distanza di Mahalanobis, calcolata per ogni singola vocale, come indice che ben descrive e discrimina la variabilità intra e inter parlatore (mediamente più rilevante di altri indici). Per stimare le due funzioni di verosimiglianza, però, sorge un problema dovuto al fatto che, in generale, non si dispone di ripetizioni per la voce dell'anonimo e per quella dell'imputato, occorre, pertanto, utilizzare un metodo di ricampionamento per studiare la variabilità intra parlatore. Nel caso in esame si è scelto di utilizzare il metodo bootstrap, ricampionando dall'urna delle vocali del parlatore di interesse per tener conto delle correlazioni tra le formanti di una stessa vocale (Rossi, 1998). Un'ampia sperimentazione (Rosati, 2001) ha mostrato le buone prestazioni del modello di urna scelto. Si può, quindi, sempre partire dalle due matrici dei dati contenenti le formanti delle ripetizioni bootstrap del parlatore da riconoscere e le formanti del Data-Base di riferimento per la stima delle due funzioni di verosimiglianza.

Un primo passo consiste nell'analizzare separatamente, per ogni parlatore, le quattro vocali "A", "E", "I" ed "O". Per ogni vocale, quindi, sarà possibile avere un campione di distanze statistiche tra parlatori diversi ed uno per lo stesso parlatore. Siano infatti $\boldsymbol{X}_A^{(k)}$, $\boldsymbol{X}_E^{(k)}$, $\boldsymbol{X}_I^{(k)}$ e $\boldsymbol{X}_O^{(k)}$, i vettori caratteristici relativi, rispettivamente alle vocali "A", "E", "I" ed "O" del parlatore *k*-esimo. Ogni vettore caratteristico contiene i valori delle 4 formanti determinate per ogni vocale, relative ad n_V osservazioni diverse effettuate per ogni vocale (n_A osservazioni per la vocale "A", n_E osservazioni per la vocale "E", e così via).

Per esempio, il vettore caratteristico della vocale "A" del parlatore *k*-esimo sarà rappresentato da una matrice $n_A \times 4$ del tipo:

$$\boldsymbol{X}_A^{(k)} = \begin{bmatrix} f_{A0}(1)^{(k)} & f_{A1}(1)^{(k)} & f_{A2}(1)^{(k)} & f_{A3}(1)^{(k)} \\ f_{A0}(2)^{(k)} & f_{A1}(2)^{(k)} & f_{A2}(2)^{(k)} & f_{A3}(2)^{(k)} \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ f_{A0}(n_A)^{(k)} & f_{A1}(n_A)^{(k)} & f_{A2}(n_A)^{(k)} & f_{A3}(n_A)^{(k)} \end{bmatrix}$$

dove $f_{Ah}(l)^{(k)}$ rappresenta il valore della *h*-esima frequenza formante che è stato rilevato nella *l*-esima osservazione, effettuata sul *k*-esimo parlatore.

Per "calibrare" il metodo nella fase di addestramento, è necessario riferirlo ad un determinato parlatore, che nel processo giudiziario è l'imputato (o l'anonimo). Per valutare il rapporto di verosimiglianza, sarà allora necessario calcolare la distribuzione della distanza sia nell'ipotesi accusatoria (confronto tra l'imputato e sé stesso) sia in quella difensiva (confronto tra l'imputato e la popolazione di riferimento).

Allora, i campioni della distanza statistica nell'ipotesi accusatoria verranno calcolati considerando la distanza tra due osservazioni, estratte casualmente, dello stesso parlatore che rappresenta l'imputato, riferite ad una determinata vocale. Analogamente, i campioni della distanza statistica nell'ipotesi difensiva verranno calcolati considerando la distanza tra le osservazioni dell'imputato e di tutti gli altri, sempre riferite ad una determinata vocale. Ripetendo tale procedimento più volte si otterranno due campioni di distanze per ogni vocale, della numerosità desiderata e di dimensioni pari a 4 (il numero delle vocali considerate). In totale quindi si avranno 4 vettori di distanze intra-parlatore e 4 vettori di distanze inter-parlatore che tengono conto delle correlazioni tra le formanti di una stessa vocale. Quindi usando la norma di Mahalanobis che, tra tutte quelle sperimentate, si è rivelata la più efficiente, si possono ottenere due soli vettori di distanze: uno "intra-" e l'altro "inter-" parlatore, che tengono conto delle correlazioni esistenti tra le distanze relative alle diverse vocali (Rossi, 1998).

A partire dai due vettori delle distanze si può procedere alla stima delle distribuzioni di interesse necessarie per il calcolo del rapporto di verosimiglianza. In precedenti lavori (Paronetto, 1995; Calvani, 1996) si è dimostrato come non sia possibile ipotizzare modelli parametrici per descrivere le verosimiglianze. Occorre, pertanto, utilizzare una modellizzazione non parametrica. Si è scelto di utilizzare il metodo di stima non parametrico kernel, con componenti gaussiane e con ottimizzazione del parametro di smussamento ottenuta con il metodo "direct plug in" (Forte, 1998; Ghizzoni, 1999).

Dalle verosimiglianze così stimate si ricava immediatamente il corretto rapporto di verosimiglianza e si identifica univocamente la soglia di decisione coerente con il criterio scelto.

6.2. *La fase di riconoscimento*

Il riconoscimento avviene semplicemente confrontando il valore della distanza tra l'anonimo e l'imputato con la soglia di decisione: se il valore è inferiore alla soglia, si conclude che le due voci sono uguali, se è maggiore che le due voci sono distinte. Si valuta, di conseguenza, la probabilità dei due tipi di errore.

Lo schema riportato in figura 2, riassume la procedura completa.

7. UN'APPLICAZIONE A DATI REALI

Il metodo descritto sopra è stato applicato a dati reali rilevati dal Servizio di Polizia Scientifica di Roma. Essi sono relativi a più di 100 registrazioni di parlatori italiani adulti di sesso maschile; ogni registrazione comprende un numero diverso di osservazioni per ogni vocale. Da un punto di vista pratico, quindi, ogni parlatore è descritto da quattro matrici di quattro colonne (o meno in caso di esclusione dovuta a dati mancanti) ed un numero variabile di righe. Per la stima della distribuzione della distanza intra-parlatore è necessaria la generazione di un opportuno campione bootstrap. Per l'analisi dei dati è stato scritto un programma nel linguaggio S-PLUS 2000, che consente, scelto un parlatore da un database considerato di riferimento, di ottenere:

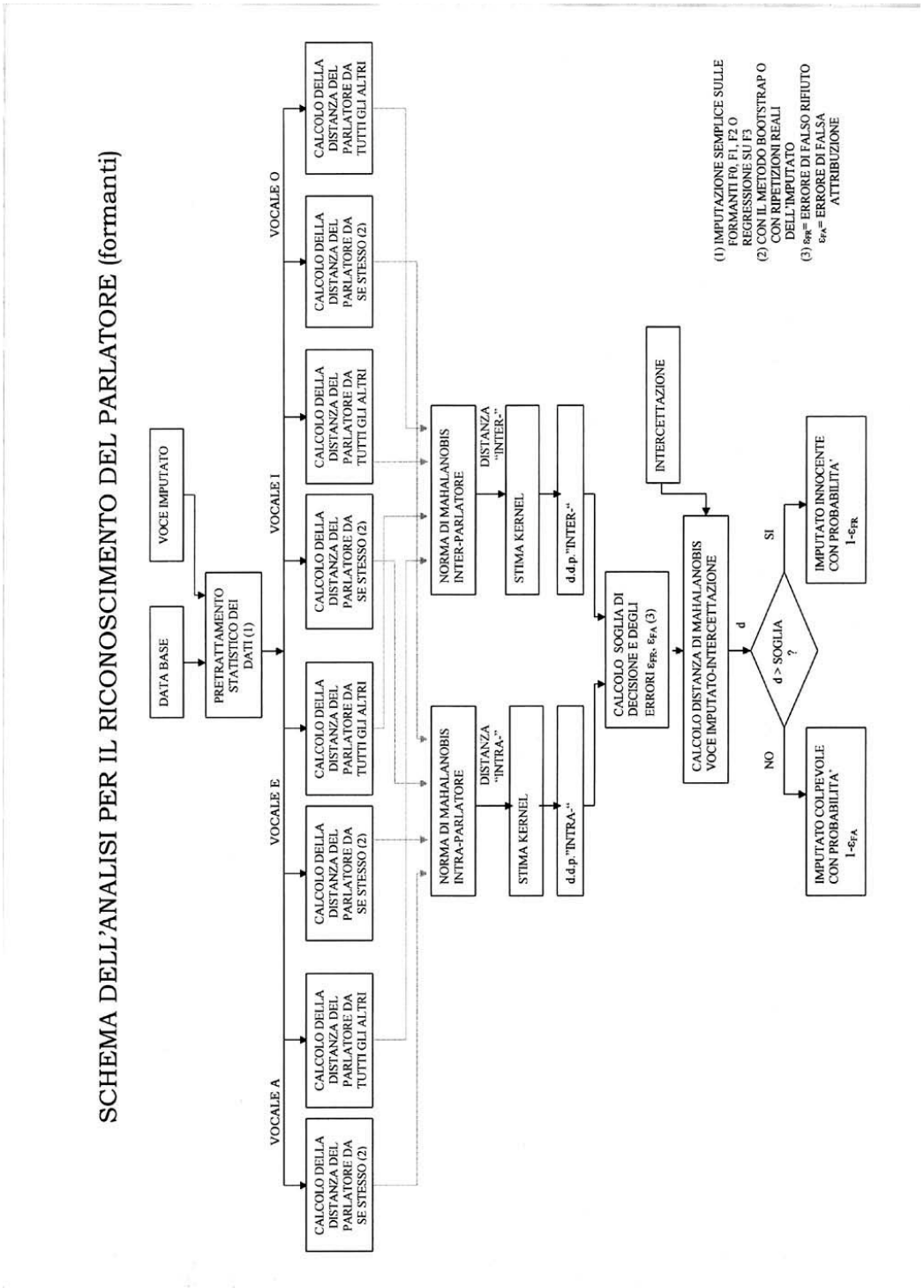

Figura 2 – Schema della procedura per il riconoscimento del parlatore sulla base delle formanti.

– le distanze inter-parlatore in relazione a tutti gli altri soggetti inseriti nel data-base di riferimento;

– la generazione di un campione bootstrap per il calcolo di un uguale numero di distanze intra-parlatore;

– le distanze intra-parlatore calcolate sulla base del campione bootstrap generato;

– la stima delle verosimiglianze, il calcolo della soglia di decisione e delle probabilità dei due tipi di errore;

Il programma, che presenta delle interfacce standard Windows, guida l'operatore nella selezione dei parametri utilizzati nell'analisi. Esso consente di effettuare inoltre il confronto tra il parlatore in esame ed un parlatore "anonimo", la cui voce è stata intercettata tramite registrazione e dalla quale sono state estratte le formanti relative alle vocali. Sulla base della soglia di decisione calcolata e delle probabilità dei due tipi di errore (a priori), è possibile quindi prendere la decisione circa l'appartenenza delle due voci alla stessa persona o meno. Dopo il confronto, è possibile conoscere anche le probabilità a posteriori dei due tipi di errore.

In Figura 3 è mostrata una finestra tipica presentata dal programma relativamente alla fase di addestramento, in cui è presente il grafico delle due verosimiglianze stimate con il metodo Kernel per tutte le vocali (tranne la "u") relativamente al parlatore "1" del database ed un resoconto dettagliato, contenente tutti i parametri di interesse. La soglia di decisione è ottenuta dal rapporto di verosimiglianza pari a 1.

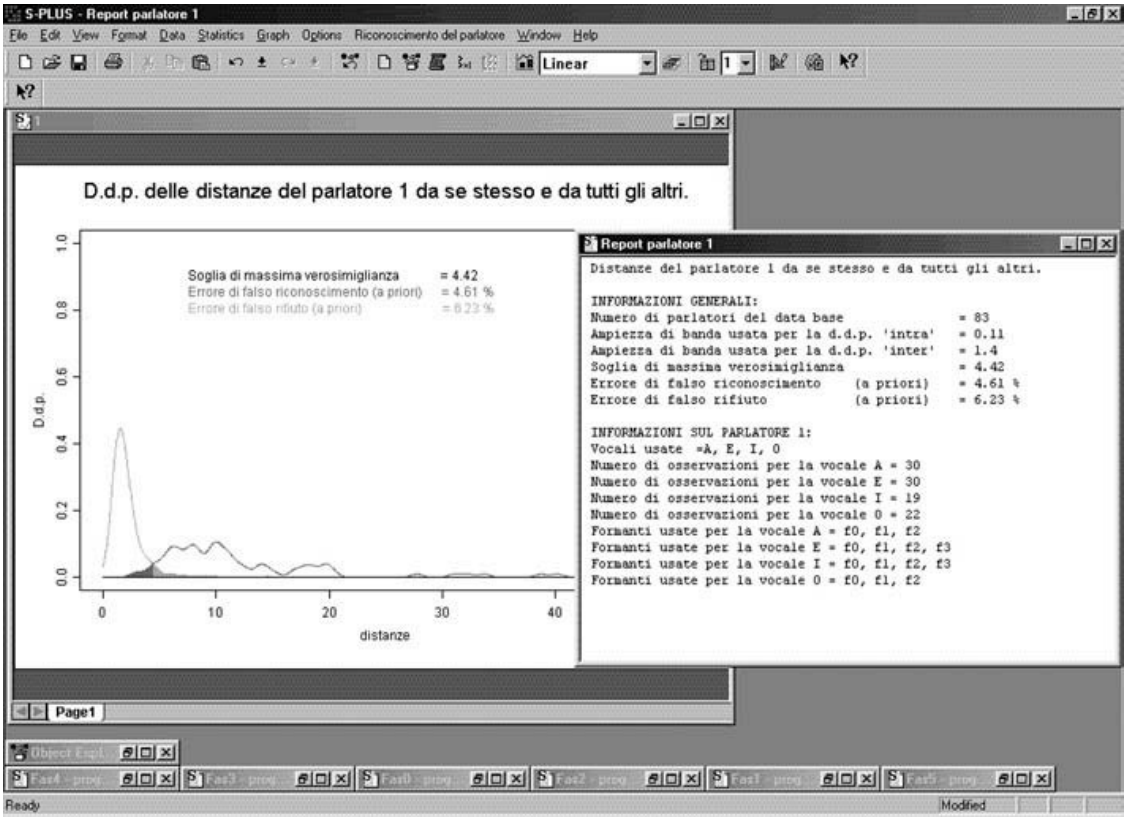

Figura 3 – Finestra relativa alla fase di addestramento: verosimiglianze relative alla distanza (norma di Mahalanobis) intra- e inter-parlatore per le vocali "a", "e", "i", "o", stimata con il metodo Kernel.

Nel caso esaminato, ad esempio, si vede che per calcolare il rapporto di verosimiglianza sono state usate le quattro vocali "a", "e", "i", "o", ed un numero di formanti variabile da vocale a vocale; ciò è dovuto al fatto che per molti parlatori del database non era disponibile alcuna osservazione relativa alla formante f_3 delle vocali "a" e "o". Il programma riconosce automaticamente tali mancanze e permette all'utente di scegliere se continuare ad utilizzare quei parlatori per costruire il database (non analizzando quindi le formanti mancanti) oppure eliminarli.

Gli errori di falso riconoscimento e falso rifiuto si riducono notevolmente utilizzando le quattro vocali contemporaneamente, combinate mediante la norma di Mahalanobis, anziché le singole vocali separatamente. Tuttavia, l'utente può scegliere quali e quante vocali utilizzare, così come quali e quante formanti, purché siano disponibili i file delle relative misure.

In Figura 4 è mostrato il risultato del confronto tra il parlatore "1" e il "118" (fase di riconoscimento), supposto anonimo (cioè l'intercettazione). In essa sono mostrate ancora le verosimiglianze intra- e inter-parlatore di Figura 2, con l'ulteriore informazione sugli errori di falso riconoscimento e falso rifiuto a posteriori (p osservata). Nell'esempio mostrato, poiché la soglia di verosimiglianza è pari a 4.42, mentre l'intercettazione presenta una distanza dal parlatore "1" pari a 8.52, è più verosimile che le due voci appartengano a due diverse persone, con probabilità di errore pari all'1.55%. Se, tuttavia, si decide che le due voci appartengano allo stesso parlatore, l'errore ha probabilità pari a 37.35%.

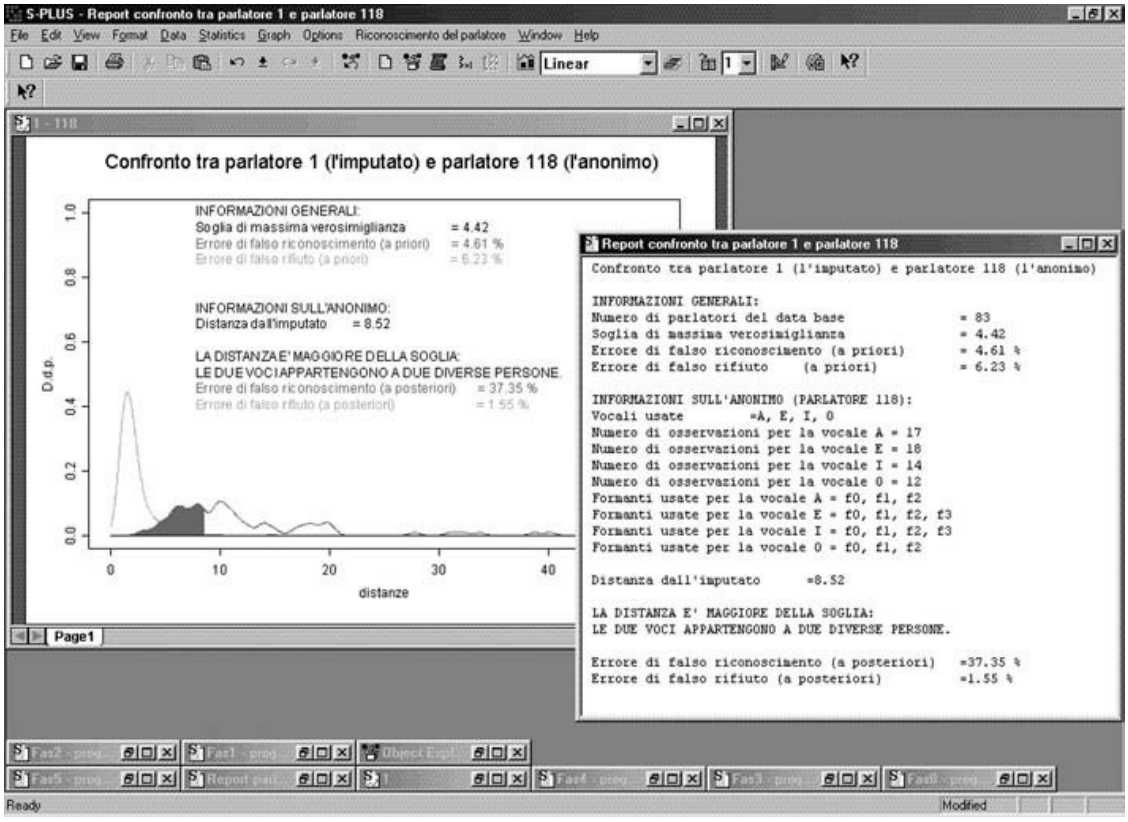

Figura 4 – Finestra relativa alla fase di riconoscimento: risultato del confronto, basato sulle vocali "a", "e", "i", "o", tra parlatore "1" e parlatore "118".

8. CONCLUSIONI E SVILUPPI FUTURI

L'utilizzo di ulteriori parametri acustici e l'applicazione di altri metodi e modelli probabilistico-statistici contribuiranno a migliorare i metodi di identificazione del parlatore nell'ambito investigativo e forense, ma tali miglioramenti dipenderanno fortemente da un modello più completo dello "spazio del parlatore" e, soprattutto, da un approccio di tipo integrato che combini sia i diversi modelli, sia i diversi parametri utilizzati in un contesto meta-analitico. La ricerca trarrà inoltre grande vantaggio dall'uso di banche dati contenenti campioni provenienti da un gran numero di parlatori che possano essere considerati rappresentativi in una determinata area di interesse.

Alcuni risultati preliminari ottenuti applicando lo stesso approccio metodologico mostrato per le formanti ma utilizzando un altro tipo di misure effettuate sul segnale digitalizzato, i cosiddetti coefficienti Mel-Cepstrum, mostrano che, basandosi su tali misure, è possibile sostanzialmente dimezzare il livello di errore e che l'utilizzo congiunto di formanti e Mel-Cepstrum in una procedura integrata può ulteriormente ridurre le due probabilità (Causarano, 2001).

Servizio Polizia Scientifica – Roma TOMMASO BOVE

Istituto di Acustica O. M. Corbino – CNR Roma PAOLO EMILIO GIUA

Dipartimento di Matematica
Università di Roma "Tor Vergata" ALESSANDRA FORTE
CARLA ROSSI

RINGRAZIAMENTI

Il presente lavoro è stato svolto nell'ambito del progetto europeo OISIN, S.M.A.R.T (Statistical Methods Applied to the Recognition of the Talker). In questo contesto si ringraziano il signor Stefano Delfino e tutti i tecnici della polizia scientifica e la Dott.ssa Gabriella Fanello Marcucci, unitamente alla direzione di Radio Radicale, per aver reso disponibili alcune registrazioni della trasmissione "Filo Diretto".

Si ringrazia l'anonimo referee che, con le sue critiche, ha permesso di giungere ad una stesura dell'articolo più ampia e dettagliata, maggiormente comprensibile anche a non esperti di riconoscimento del parlatore in ambito giudiziario.

RIFERIMENTI BIBLIOGRAFICI

AITKEN C.G.G., (1995), *Statistics and the evaluation of evidence in forensic science*, Wiley, New York.

BIMBOT F., MAGRIN-CHAGNOLLEAU I., MATHAN L., (1995), *Second-order statistical measures for text-independent speaker identification*, "Speech communication", 17 pp. 177-192.

CALVANI F. (1996), *Il problema dell'errore di assegnazione nel riconoscimento del parlatore,* Tesi di Laurea in Matematica, Università degli Studi di Roma 'Tor Vergata'.

CAUSARANO A. (2001), *Il problema del riconoscimento del parlatore: stima delle distanze intra e inter parlatore per diverse caratteristiche del segnale*. Tesi di Laurea in Matematica, Università degli studi di Roma "Tor Vergata".

FORTE A., (1998), *Analisi critica di metodi per la classificazione e l'identificazione del parlatore nelle scienze forensi*, Tesi di Laurea in Matematica, Università degli Studi di Roma 'Tor Vergata', 1998.

FURUI S., (1989), *Digital speach processing, synthesis and recognition*, Marcel Dekker inc., New York.

GHIZZONI A., (1999), *Il problema dell'identificazione del parlatore nelle scienze forensi: modelli, metodi di classificazione e analisi di dati*, Tesi di Laurea in Matematica, Università degli Studi di Roma 'Tor Vergata'.

IBBA G., PAOLONI A., (1993), *Analisi delle voci: il parlatore ignoto*, "Poste e Telecomunicazioni", 1**,** pp. 14-25.

MATSUI T., FURUI S., (1994), *Adaptation of Tied Mixture Based Phoneme Models for Text-Prompted Speaker Verification*, in: "Proceedings ICASSP" 1**,** pp. 125-128.

PARONETTO B., (1995), *Problemi di stima di densità per l'identificazione del parlatore: utilizzo del metodo kernel,* Tesi di Laurea in Matematica non pubblicata, Università degli Studi di Roma 'La Sapienza'.

ROBERTSON B., VIGNAUX G.A., (1995), *Interpreting Evidence: evaluating forensic science in the courtroom*, Wiley, New York.

ROSATI F. (2001), *Sperimentazione del metodo Bootstrap nel problema del riconoscimento del parlatore*, Tesi di Laurea in Matematica, Università di 'Tor Vergata'.

ROSSI C., (1996), *Il problema di decisione nell'identificazione del parlatore*, in: "Caratterizzazione del Parlatore", F. Fedi & A. Paoloni eds., Fondazione Ugo Bordoni Roma, pp. 173-176.

ROSSI C., (1998), *Classification and Decision Making in Forensic Sciences: the Speaker Identification Problem*, in: "Advances in Data Sciences and Classification", Rizzi A., Vichi M., Bock H.H. eds., Springer Verlag, Heidelberg, pp. 647-654.

RIASSUNTO

Un metodo statistico per il riconoscimento del parlatore basato sull'analisi delle formanti

Nel presente contributo viene analizzato un metodo per il riconoscimento del parlatore basato sulle formanti, ovvero sull'analisi della frequenza fondamentale e delle prime tre formanti delle vocali "a", "e", "i" e "o". Sulla base di tali misure, il metodo proposto stima in modo non parametrico le densità di probabilità delle distanze di Mahalanobis di un parlatore da se stesso (distanza intra-parlatore)e dalla popolazione di riferimento (distanza inter-parlatore) per ogni singola vocale. Viene quindi calcolata la norma di Mahalanobis per le quattro vocali congiuntamente e se ne stima la densità di probabilità. Si definisce, quindi, la regola di classificazione di ogni nuovo soggetto sulla base del criterio di assegnazione di massima verosimiglianza. Questo permette di determinare un'unica soglia di decisione e, di conseguenza le probabilità dei due possibili errori: falso rifiuto e falso riconoscimento. Il metodo è stato sperimentato su un insieme di dati reali forniti dal Servizio Polizia Scientifica di Roma nell'ambito di un progetto europeo.

SUMMARY

A method for speaker recognition based on formants

In this paper, a method for the forensic identification of speakers is presented, based on the analysis of the pitch and the first three formants of the four vowel: "a", "e", "i" and "o". Using these data, the method estimates the probability density function (pdf) of the

Mahalanobis distance both of the defendant from himself (intra-distance estimation) and from the voices of the control set (inter-distance estimation), using the Kernel method for each vowel. The Mahalanobis norm is then used to estimate the pdf related to the four vowel. The sample under study is then classified according to the Maximum Likelihood Criterion approach. This allows one to estimate a unique decision threshold and the probabilities of the two possible classification errors (false acceptance and false rejection). The method has been applied to real data provided by Police Scientific Service of Rome, in the framework of a European Project.