*STATISTICA, anno LXII, n. 3, 2002*

# UNA PROPOSTA DI META-ANALISI BASATA SULLA COMBINAZIONE DI CLASSIFICATORI PER IL PROBLEMA DEL RICONOSCIMENTO DEL PARLATORE (*)

P. Brutti, F. Fabi, G. Jona Lasinio

## 1. INTRODUZIONE

Recentemente nell'ambito delle ricerche inerenti le tecniche di riconoscimento del parlatore si sta affermando un nuovo paradigma: combinare più classificatori allo scopo di sviluppare sistemi di identificazione compositi che migliorino le prestazioni dei sistemi componenti (Avnimelech e Intrator, 1999; Cacciatore e Nowlan, 1994; Chen *et al.*, 1995-1998; Xu *et al.*, 1992-1993; Duin *et al.*, 1998, 2000; Jacobs, 1995; Moerland e Mayoraz, 1999; Rida *et al.*, 1999).

Due ordini di motivi giustificano il ricorso ad un tale approccio. Anzitutto nel problema del riconoscimento del parlatore, come d'altra parte in tutte le applicazioni di tipo *pattern recognition*, numerose sono le tecniche di classificazione proposte e, talvolta, profondamente differenti i principi su cui si basano. Sintetizzando oltre il lecito potremmo anzitutto identificare due grandi gruppi di metodi: da una parte quelli basati su una qualche tipologia di caratteristiche estratte dai dati "grezzi" (Figura 1) e dall'altra tutti quelli che, non ricorrendovi, possono essere denominati sintattico-strutturali. Ciascuno di tali gruppi, poi, include algoritmi dalle giustificazioni teoriche più diverse. Per fare un esempio, osserviamo che solo nel primo di questi si collocano, assieme ai $k$–NN[1], i numerosi classificatori neurali nonché la maggior parte dei classificatori basati su distanze. Si consideri inoltre che, in generale, le prestazioni di un classificatore dipendono fortemente dall'applicazione che ci si trova ad affrontare e che solo raramente si è effettivamente in grado di individuare una procedura ottimale. Concreta appare allora la necessità di sviluppare tecniche di *pooling* efficaci.

Il secondo dei motivi cui si accennava è da ricercarsi nel fatto che, come in molti problemi di riconoscimento, anche in quello da noi considerato lunga è la lista delle caratteristiche dal contenuto informativo "semi-complementare" utilizzabili al fine di rappresentare ed identificare strutture salienti nei dati:

---

(*) 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*).

[1] *k–Nearest Neighborhood* (vedi Hastie et al., 2001).

coefficienti della predizione lineare (LPC), *cepstrum* (CEPS), *MEL-cepstrum* (MEL-CEPS), spettri ai terzi d'ottava e coefficienti *wavelet*, sono solo alcuni esempi. Intuibili risultano allora essere le difficoltà in cui ci si imbatte qualora si tenti di sfruttarle congiuntamente con fini classificatori. Pur anche ammettendo, infatti, che si riesca ad individuare un insieme di caratteristiche per cui non risulti necessario ricorrere a normalizzazioni, operare aggregandole in un unico vettore di grandi dimensioni equivarrebbe ad andare incontro a tutta una serie di problemi, dall'instabilità numerica all'ingestibilità computazionale, connessi con ciò che Bellman (1961) definì flagello della dimensionalità (*curse of dimensionality*). Ciò detto, possiamo concludere che anche per il problema delle "diverse caratteristiche" la combinazione di classificatori operanti su un ristretto numero di variabili in input si propone come una fra le soluzioni potenzialmente adottabili.

Giustificatane la filosofia, passiamo ora ad introdurre COMBY, il modello da noi proposto al fine di implementare la metodologia appena descritta. Fatta l'ipotesi che i classificatori scelti appartengano al primo dei suddetti due gruppi, procediamo guardando al problema di classificazione da un punto di vista statistico. In quest'ottica ciascun classificatore altro non fa se non stimare la probabilità che ciascuna delle classi caratterizzanti il problema sia quella vera, quella cioè, da cui effettivamente provengono i dati considerati (condizionatamente al valore assunto dal vettore delle caratteristiche). Combinare un certo numero di classificatori, allora, equivale a mettere assieme in modo opportuno altrettante distribuzioni di probabilità aventi come supporto comune l'insieme delle classi $\{C_1,\ldots,C_k\}$. Un modo fra gli altri (Chen, 1997, 1998) di effettuare una tale operazione, scelto per la sua efficacia e facilità implementativa, verrà descritto nel prossimo paragrafo.

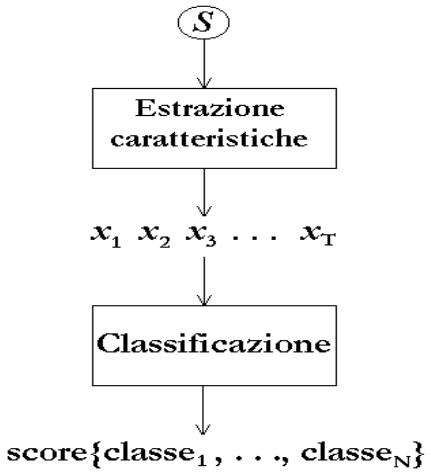


*Figura 1* – Sistema di riconoscimento basato sull'estrazione di caratteristiche dal segnale *S* (diagramma funzionale).

A conclusione di questa introduzione, quindi, possiamo dire che il presente lavoro si propone essenzialmente di proseguire la sperimentazione sul parlato italiano nella direzione della costruzione di una struttura di indagine meta-

analitica, riprendendo le linee già tracciate in Fabi *et. al* (2001) ed in Bove *et al.* (2001), in cui le problematiche inerenti l'uso forense delle procedure di riconoscimento del parlatore vengono illustrate insieme ai risultati di un'analisi comparativa tra l'approccio correntemente in uso in Italia ed alcune nuove metodologie. A tal proposito facciamo però notare che i risultati qui riportati sono ancora in via di completamento. In particolare il numero di tracce vocali considerate nella sperimentazione, al momento limitato a 50, sarà presto incrementato in modo da sfruttare a pieno l'informazione contenuta nella base di dati resaci disponibile.

## 2. METODOLOGIA

Rimandando ai lavori in bibliografia per dettagli ed approfondimenti tecnici, in questo paragrafo ci proponiamo di descrivere succintamente logica e limiti della famiglia di modelli a cui le architetture utilizzate appartengono.

### 2.1. *Modelli a moduli di esperti*

Sia i modelli a moduli di esperti (*mixtures of experts* – ME), introdotti da Jacobs *et al.* (1991), che i modelli gerarchici a moduli di esperti (*hierarchical mixtures of experts* – HME), proposti da Jordan e Jacobs (1994), definiscono due importanti paradigmi di apprendimento dai dati che risultano di interesse comune per i ricercatori nelle aree dell'apprendimento automatico (*machine learning*) e della statistica.

Nella sua generalità, il problema da affrontare consta nell'individuazione di un'applicazione che, conformemente ad un dato criterio, presenti una struttura differente in regioni differenti dello spazio delle covariate (o di ingresso). L'approccio implementato dai modelli qui esaminati si rifà al principio *divide et impera* di romana memoria e consiste essenzialmente nella costruzione di un complesso e potente modello mediante combinazione di architetture localizzate più semplici ed interpretabili. Negli ME, ad esempio, gli esperti sono generalmente reti neurali *feed forward* [2] multistrato (ad esempio percettroni multistrato) o monostrato (ad esempio modelli lineari generalizzati[3]).

Queste reti sono addestrate[4] simultaneamente in modo tale che gli esperti

[2] Bishop, 1995.

[3] Particolarmente utilizzate e studiate (Jiang e Tanner, 1999-2000; Xu, 1998; Zeevi *et al.*, 1998) sono le architetture aventi per funzione arbitro un modello logistico (McCullagh e Nelder, 1989) e per esperti, dei modelli lineari o dei modelli logistici a seconda che il problema da affrontare sia rispettivamente una regressione o una classificazione.

[4] Addestrare una rete equivale, in termini statistici, a stimare i parametri del relativo modello. Il termine "addestramento" conferisce linguisticamente una dimensione temporale a tale procedura, ed in effetti, data la complessità dei modelli coinvolti, l'utilizzo di algoritmi iterativi è la prassi (Bishop, 1995; McCullagh e Nelder, 1989).

competano per apprendere i *pattern* in input divenendo "responsabili" di una particolare regione dello spazio delle covariate. Questa competizione viene mediata dalla cosiddetta rete o funzione arbitro (*gating network*) la quale produce una serie di coefficienti di ponderazione locali, ossia funzionalmente dipendenti dal vettore delle covariate, atti a pesare i contributi dei diversi esperti sulla base delle prestazioni conseguite da ciascuno di essi nelle diverse regioni dello spazio delle covariate al termine della procedura di apprendimento.

Più in particolare, possiamo illustrare il funzionamento di un modello ME formalizzando ciascun esperto come un processo statistico che genera un output $\mathbf{y}$ secondo una certa misura di probabilità $\mathbf{P}(\mathbf{y} \mid \mathbf{x}, \Theta_e)$ condizionata all'ingresso $\mathbf{x}$ ed al valore assunto dal vettore dei parametri $\Theta_e$, e la rete arbitro come un classificatore che, condizionatamente al vettore dei parametri $\Theta_a$, trasforma il vettore delle covariate $\mathbf{x}$ nelle probabilità $\{g_j(\mathbf{x})\}_{j\in\{1,\ldots,\mathsf{N}(e)\}} = \{\mathbf{P}(e_j \mid \mathbf{x}, \Theta_a)\}_{j\in\{1,\ldots,\mathsf{N}(e)\}}$ che ciascuno degli $\mathsf{N}(e)$ esperti $e_j$ sia in grado di generare l'output desiderato (corretta classificazione). Con questa notazione, possiamo dire che la rete arbitro, assegnando a ciascun punto $\mathbf{x}$ dello spazio delle covariate $X$ un vettore di probabilità $\mathbf{g}(\mathbf{x}) = [g_1(\mathbf{x}), \ldots, g_{\mathsf{N}(e)}(\mathbf{x})]^{\mathrm{T}}$ altro non fa che generare una partizione sfumata (*fuzzy* o *soft partition*) di $X$ costituita da $\mathsf{N}(e)$ regioni $\{\mathsf{R}_1, \ldots, \mathsf{R}_{\mathsf{N}(e)}\}$ il cui generico elemento $\mathsf{R}_j$, avendo funzione di appartenenza pari proprio a $g_j(\mathbf{x})$, risulta essere controllato dal $j$-esimo esperto. In definitiva, dunque, poiché la probabilità che il $j$-esimo esperto risponda correttamente all'interrogazione è data dalla probabilità che venga scelto per la probabilità che risponda correttamente,

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

la probabilità di osservare un comportamento corretto dell'intero modello sarà chiaramente data dalla somma delle precedenti probabilità, ovvero da una *mistura* (McLachlan e Peel, 2000; Titterington *et al.*, 1985) delle misure di probabilità $\mathbf{P}(\mathbf{y} \mid \mathbf{x}, \Theta_{e_j})$ con coefficienti di ponderazione $g_j(\mathbf{x})$ dipendenti esplicitamente dal vettore delle covariate $\mathbf{x}$.

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

Il valore atteso di tale distribuzione di probabilità, pari a:

$$\boldsymbol{\mu} = \sum_{j=1}^{\mathsf{N}(e)} g_j(\mathbf{x})\boldsymbol{\mu}_j, \tag{2}$$

risulta essere, una volta stimati i parametri $\Theta_a$, e $\{\Theta_{e_j}\}_{j\in\{1,\ldots,\mathsf{N}(e)\}}$, la scelta naturale per l'uscita della rete o, in altre parole, per la stima delle probabilità di appartenenza alle diverse classi del vettore in ingresso **x**.

Osserviamo a questo punto che la struttura fin qui descritta è del tutto generale e variamente specificabile nei suoi costituenti atomici: gli esperti e la rete arbitro. Nel caso delle architetture gerarchiche, ad esempio, la suddetta procedura di partizionamento dello spazio delle covariate viene iterata ricorsivamente assieme a quella di assegnazione dinamica di ciascuna regione all'esperto localmente ottimo. Ovvio risultato di tale parametrizzazione gerarchica della misura $\mathbf{P}(e \mid \mathbf{x}, \Theta_a)$ è l'individuazione di una famiglia di partizioni annidate dello spazio delle covariate tale per cui ciascuno dei suoi elementi, ossia ciascuna partizione dalla più fine alla più grossolana, risulta essere descritta da un particolare livello della struttura ad albero in cui il modello HME si presenta naturalmente organizzato (Figure 2 e 4). Le architetture gerarchiche a moduli di esperti, dunque, altro non sono se non alberi decisionali (Hastie *et al.*, 2001) in cui sia le decisioni, prese a livello dei nodi non-terminali da delle reti arbitro, che gli output, generati dagli esperti posti sulle foglie dell'albero stesso, risultano formalizzati probabilisticamente[5]. Se allora pensiamo ciascun nodo non-terminale come una funzione base avente per supporto il sottoinsieme di $X$ contenente tutti i valori che effettivamente "transitano" per quel nodo, tale insieme coinciderà esattamente con l'unione dei supporti delle funzioni base associate a ciascuno dei discendenti del nodo considerato. Come in un'analisi di tipo *wavelet*, dunque, la struttura[6] annidata della partizione sfumata generata da un HME, ci permette di studiare i dati a diversi livelli di risoluzione ciascuno corrispondente ad una particolare altezza dell'albero decisionale.

Ovviamente la semplice gerarchizzazione della misura di probabilità associata alla rete arbitro non solo non è l'unica delle specificazioni possibili ma non è neanche la migliore tra quelle implementabili al fine di affrontare i problemi di dimensionalità cui si è accennato nel paragrafo introduttivo. Vediamo allora come modificare la modellistica descritta conformemente alle suddette esigenze.

---

[5] In generale gli HME offrono prestazioni migliori (Waterhouse, 1997; Jordan e Jacobs, 1994; Domeniconi e Jordan, 2001) rispetto ad altre implementazioni degli alberi decisionali, come CART (Breiman *et. al.*, 1984) e MARS (Friedman, 1991). Per intuirne le ragioni soffermiamoci sul caso regressivo ed utilizziamo l'errore quadratico medio come criterio per valutare la bontà di uno stimatore $\hat{\mathbf{y}}(\mathbf{x})$. Ora, sebbene la divisione dello spazio delle covariate, e la conseguente diminuzione della distorsione dello stimatore $\hat{\mathbf{y}}(\mathbf{x})$, sia comune a tutti i suddetti modelli, l'utilizzo, tipico degli HME, di margini "sfumati" (*soft*) invece che rigidi (*hard*), genera stimatori più "smussati" ossia dalla varianza ridotta, data l'influenza esercitata sulle stime dei parametri del modello *j*-esimo dai dati allocati in regioni attigue ad $\mathsf{R}_j$.

[6] Le regioni generate dall'albero decisionale sono in definitiva politopi che generalizzano le decomposizioni di tipo *wavelet* (Vidakovic, 1999) nel caso di spazi multi-dimensionali (Donoho, 1997).

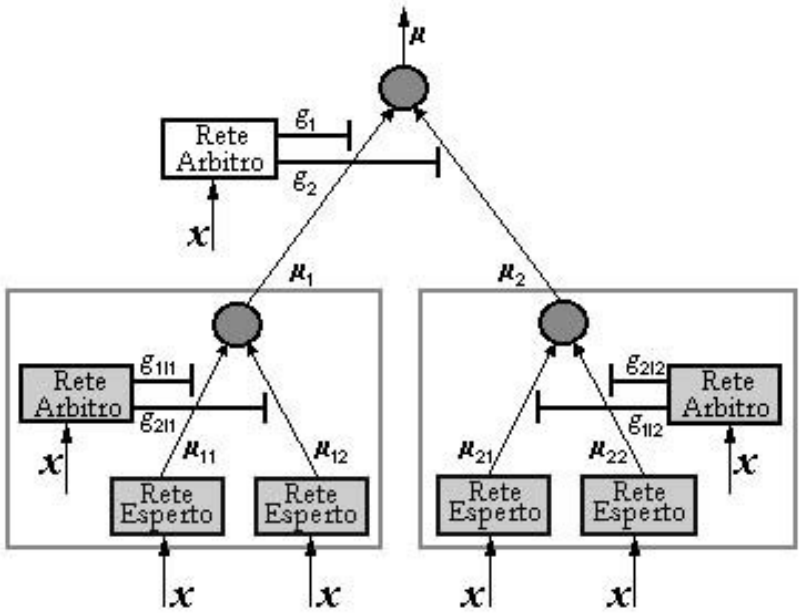


*Figura 2* – Rappresentazione grafica di un modello HME a due livelli con fattore di ramificazione (*branching factor*) pari a 2 visto come rete neurale modulare (adattato da Jordan e Jacobs 1994). Notare come sia le reti esperti che le reti arbitro ricevano in entrata il vettore di covariate **x**.

## 2.2. *Dai modelli ME alle combinazioni generalizzate: COMBY*

Come visto, punto di forza degli ME è la loro capacità di associare adattivamente ciascun elemento di un insieme di esperti al compito che meglio riesce a svolgere. Per questa ragione gli ME sembrerebbero essere particolarmente promettenti al fine di risolvere efficacemente il primo dei nostri problemi: il *pooling* di classificatori. Le vere difficoltà, però, si incontrano nel momento in cui si tenti di abbinare l'impiego di più caratteristiche acustiche[7] ad un sistema di riconoscimento basato sulla più classica versione di tali modelli. Questo perché altro modo non v'è in tale ambito se non quello di condensarle in un unico vettore dalle dimensioni computazionalmente ingestibili (*flagello della dimensionalità*).

Con l'intento di ovviare a tale limite, Chen e Chi (1998) propongono le seguenti varianti da applicare all'architettura poc'anzi descritta:

1. si introduce un insieme o banco[8] di reti arbitro (il "combinatore" di Figura 3) capace di ponderare gli output dei vari esperti sulla base di un trattamento separato dei $K$ vettori di caratteristiche, $\{\mathbf{x}_1,\ldots,\mathbf{x}_k\}$, estratte dal segnale grezzo.
2. a ciascun vettore di caratteristiche $\mathbf{x}_k \;\forall\, k \in \{1,\ldots,K\}$ si fa corrispondere un *pool* di esperti costituito da $N_k$ classificatori tra loro differenti (e.g. VQ, MLP, GMM).

[7] Si veda Fabi *et al.* (2001) per una descrizione dettagliata delle stesse.
[8] Vedi anche Xu *et al.* (1995).

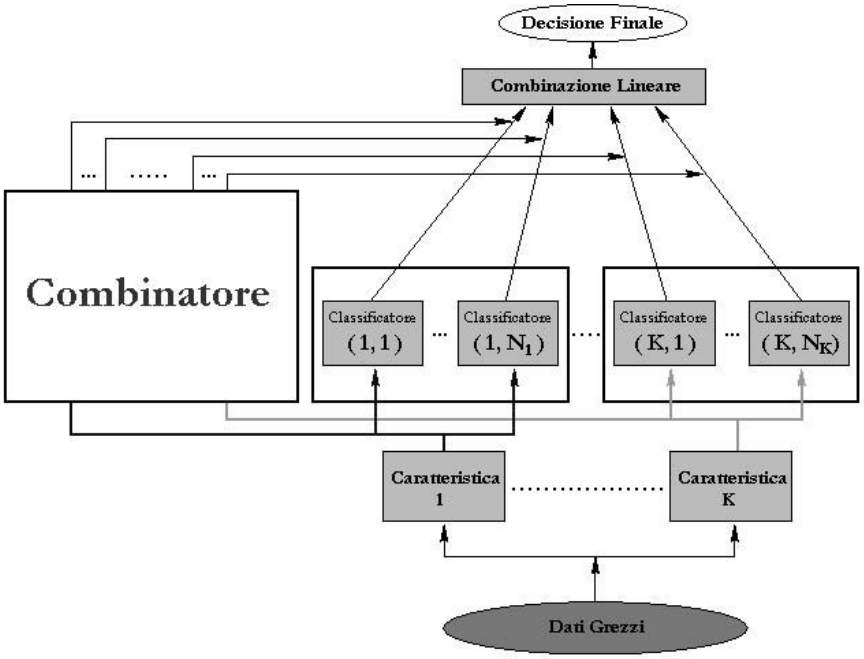


*Figura 3* – Schematizzazione di COMBY, la procedura meta-analitica da noi implementata (adattato da Chen e Chi, 1998*b*). Come si può vedere essa prevede l'estrazione di $K$ differenti vettori di caratteristiche (nel nostro caso 12 CEPS, 12 MEL-CEPS e 12 LPC → $K = 3$) dallo stesso insieme di dati e l'utilizzazione di $N$ classificatori, $N_k$ dei quali addestrati sulla base del $k$-esimo insieme di caratteristiche. Il combinatore provvede infine a ponderare opportunamente ciascuna combinazione classificatore – caratteristica.

Il modello risultante, pur permettendo ancora il ricorso ad esperti di varia foggia[9], riesce ad aggirare i più volte ricordati problemi dimensionali semplicemente derogando al gruppo di reti arbitro il compito di armonizzare, tramite mistura, i risultati delle diverse coppie classificatore – caratteristica.

La nostra proposta sebbene affine a quella di Chen, si differenzia da questa per la tipologia dei modelli impiegati. Più in particolare si è deciso di utilizzare come funzione arbitro un modello logistico (multivariato) data la superiorità dimostrata dalle architetture che implementano tale soluzione in precedenti sperimentazioni (Moerland, 2000). Oltre a ciò, il *pool* di esperti utilizzati da COMBY è composto dai seguenti classificatori:

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

L'idea alla base di questo fortunato quanto classico modello (Reynolds, 1992; Nedic e Bourlard, 2000), è quella di stimare la distribuzione di probabilità delle caratteristiche estratte dal segnale utilizzando un'opportuna mistura adattiva[10] di distribuzioni gaussiane e quindi testare l'ipotesi nulla (vedi Fabi *et al.*, 2001) che

[9] L'unico vincolo è rappresentato dal fatto che, previa normalizzazione, gli esperti debbano stimare la probabilità a posteriori delle diverse classi.

[10] L'adattività consiste nel fatto che il numero di distribuzioni coinvolte nella mistura viene selezionato automaticamente sulla base dei dati osservati. Il metodo implementato è quello descritto in Priebe (1994).

due tracce distinte siano attribuibili al medesimo parlatore semplicemente utilizzando un test di rapporto del verosimiglianze. La stima dei parametri viene effettuata secondo una particolare istanza dell'algoritmo EM.

- *Mixture of Probabilistic Principal Component Analysers* (MPPCA) (Tipping e Bishop, 1997)
- *Mixture of Probabilistic Factor Analysers* (MPFA) (Ghahramani e Hinton, 1996)

Come nel caso degli AGMM, anche i due classificatori qui considerati si basano su particolari tipologie di misture ed adottano sempre la medesima e poc'anzi ricordata procedura decisionale per testare l'ipotesi d'interesse. Essenzialmente entrambi i modelli probabilistici risultano definiti come combinazioni lineari convesse delle misure di probabilità indotte dai modelli generativi (lineari) seguenti:

$$\mathbf{x} = \mathbf{W}\mathbf{z} + \boldsymbol{\mu} + \boldsymbol{\varepsilon} \quad \text{in cui:} \quad \begin{cases} \mathbf{z} \sim N(\mathbf{0}, \mathbf{I}_\ell) \;\rightarrow\; \text{variabili latenti} \;\; \ell < \dim(\mathbf{x}) = d \\ \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \mathbf{R}) \\ \mathbf{R} = \sigma^2 \mathbf{I}_d \;\; (\text{PPCA}) \;\; \text{oppure} \;\; \mathbf{R} \;\; \text{diagonale} \;\; (\text{PFA}) \end{cases}$$

L'intento ($\ell < d$)[11] è chiaramente quello di individuare una proiezione del vettore di caratteristiche **x** che ne riduca la dimensione catturandone al contempo la struttura di correlazione, mentre il metodo di stima adottato è basato, anche in questo caso, sull'algoritmo EM (McLachlan e Peel, 2000; McLachlan e Krishnan, 1997).

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

L'unico classificatore supervisionato fra quelli considerati è un percettrone con uno strato nascosto costituito da 20 unità[12] e funzione di attivazione *softmax* (logistica multivariata). L'addestramento è stato condotto mediante *backpropagation* utilizzando come algoritmo di ottimizzazione non lineare l'SCG (*scaled conjugated gradient*; Bishop, 1995). Si è deciso di utilizzare tale classificatore data la sua ben fondata popolarità in problemi di riconoscimento del parlatore (Bennani *et al.* 1990; Bennani, 1995; Chen *et. al.*, 1995; Oglesby e Mason, 1990; Rudasi e Zahorian, 1991).

Osserviamo infine che non essendoci vincoli particolari circa la "natura" degli esperti introducibili nell'architettura appena definita, sembra piuttosto interessante, soprattutto per le applicazioni in ambito forense, la possibilità di annoverare tra questi anche linguisti o fonetisti previa accurata elicitazione probabilistica delle loro analisi.

[11] Sia la dimensione dello spazio latente che il numero di elementi della mistura sono stati individuati utilizzando procedure di validazione incrociata.

[12] Anche in questo caso si è ricorso a tecniche di validazione incrociata.

### 2.3. *Modelli grafici ed ME: implementazione ed algoritmi di apprendimento*

Il principio di Massima Verosimiglianza (*maximum likelihood* – ML), implementato seconda un'opportuna istanza dell'algoritmo EM, è alla base della proposta avanzata in Jordan e Jacobs (1994) per la stima dei parametri caratteristici del modello ME. Tale scelta è motivata essenzialmente da due risultati sperimentali:

1. l'impossibilità dell'algoritmo basato su gradiente ascendente originariamente utilizzato in Jacobs *et al.* (1991), di sfruttare la modularità della modellistica ME,
2. la più rapida convergenza dell'algoritmo EM applicato ad un HME rispetto ad una *backpropagation* di dimensioni comparabili.

Osserviamo però che ricorrendo al suddetto criterio di stima ciò che si tenta di fare è minimizzare la differenza tra i target e le previsioni del modello non già sui dati in test, bensì su quelli in *training*. Può allora accadere che un modello addestrato mediante ML, seppur sufficientemente flessibile, abbia una scarsa capacità di generalizzazione (o predittiva) perché troppo ben adattato al training set (*overfitting*).

Una soluzione classica a tale problema introdotta da Waterhouse (1997) nell'ambito della modellistica ME, è quella detta dell'*early-stopping*, consistente essenzialmente nell'arrestare la procedura di addestramento prima della convergenza dei parametri seguendo una procedura di validazione incrociata. Nello stesso lavoro Waterhouse propone anche un'altra tecnica più squisitamente bayesiana (*ensemble learning*) basata essenzialmente su di una particolare approssimazione della a posteriori suggerita dai lavori di Hinton e van Camp (1993) e Neal e Hinton (1993).

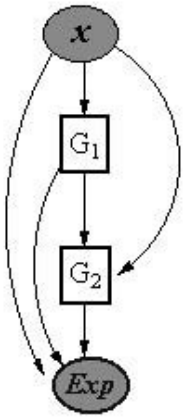


*Figura 4* – HME a due livelli visto come modello grafico. Gli archi del grafo orientato e aciclico (*directed acyclical graph*) in figura, rappresentano relazioni di indipendenza condizionata: ogni nodo è indipendente dagli altri condizionatamente al valore assunto dai nodi nel suo *markov blanket* (vedi Jensen, 2001), qui coincidente con l'insieme dei nodi ad esso connessi. I nodi denotati con **Exp** ed $\mathbf{x}$ risultano anneriti in quanto osservati. In fase di addestramento (o stima), infatti, si utilizzano insiemi di dati completamente etichettati ossia coppie $[\mathbf{x}_i, \mathbf{y}_i]$ in cui $\mathbf{y}_i$ rappresenta la classe di appartenenza dell'*i*-esimo esempio ossia il valore che ciascun esperto deve predire correttamente noto il vettore $\mathbf{x}$.

Più in generale possiamo pensare i modelli ME in termini di modelli grafici (Figura 4) ed utilizzare tutto l'armamentario tecnico sviluppato in questo contesto e presentato in testi come Jensen (2001) e Spiegelhalter *et al.* (1999), per individuare quelle probabilità a posteriori[13] che in (1) abbiamo denotato con $\mathbf{P}(\mathbf{y}|\mathbf{x})$. A tal proposito facciamo notare esplicitamente che a differenza di quanto avviene per i modelli ME ed HME classici, la stima dei parametri di COMBY è articolata in due fasi: nella prima ciascun esperto viene addestrato separatamente sull'intero data-set mentre, nella seconda, è la rete arbitro ad apprendere come ripartite lo spazio delle covariate tra i vari esperti osservando le loro previsioni a parametri "bloccati".

## 3. APPLICAZIONE A DATI REALI

### 3.1. *La base di dati ed il metodo di valutazione dei classificatori*

I metodi poc'anzi descritti sono stati applicati a dati reali rilevati dal Servizio di Polizia Scientifica di Roma (per una descrizione dettagliata della base di dati si veda Fabi *et al.* 2001), secondo il seguente protocollo di analisi (Reynolds, 1992):

- Addestramento – nella prima fase, denominata *addestramento*, viene creato il data-set mediante cui verranno successivamente stimati i parametri del modello. Più in particolare, dopo aver sottoposto ciascuna registrazione ad un opportuno pretrattamento, si provvede:
  - alla sua suddivisione in frammenti (*frames*) di lunghezza fissata (circa *23ms*, valore standard in letteratura),
  - all'estrazione, per ciascuno dei suddetti frammenti, di caratteristiche significative ai fini della loro caratterizzazione (vedi § 3.2).

- Riconoscimento – nella seconda fase, denominata *riconoscimento*, si procede secondo i seguenti passi:
  - si pretratta e frammenta il segnale vocale del parlatore ignoto con la medesima tecnica utilizzata nella fase di addestramento,
  - da ciascun frammento della traccia vocale presa in esame viene nuovamente estratto lo stesso tipo di caratteristiche utilizzate nella fase di addestramento ottenendo così una sequenza di vettori:
    $$\{\mathbf{x}_1,\ldots,\mathbf{x}_t,\ldots\},$$
  - si costruiscono a partire dalla sequenza $\{\mathbf{x}_1,\ldots,\mathbf{x}_t,\ldots\}$ dei segmenti sovrapposti di $T$ vettori[14]:
    $$\mathbf{x}^{(h)} = \{\mathbf{x}_h, \mathbf{x}_{h+1},\ldots,\mathbf{x}_{h+(T-1)}\},$$

[13] In questo contesto il termine "a posteriori" si riferisce semplicemente al condizionamento degli output della rete al vettore di covariate osservato e non necessariamente implica il ricorso a tecniche bayesiane di stima dei parametri.

[14] Nelle nostre sperimentazioni si è scelto $T$=100 con registrazioni di durata media pari a 20 secondi.

- ciascun segmento $\mathbf{x}^{(h)}$ viene considerato come una registrazione separata e passato al modello addestrato durante la prima fase,
- per ciascun vettore appartenente al segmento utilizzato il modello fornisce una classificazione. Si ha un'identificazione solo se più del 50% di tali classificazioni concordano nell'attribuire il segmento in esame ad un particolare parlatore.

Le prestazioni dei singoli modelli sono valutate mediante i seguenti indici[15] (Hand, 1997):

$$\text{tasso di identificazione} = 100 \times \left[ \frac{\text{numero segmenti } \mathbf{x}^{(h)} \text{ identificati correttamente}}{\text{numero totale di segmenti } \mathbf{x}^{(h)}} \right] = (\mathsf{TI})$$

$$\text{tasso di sostituzione} = 100 \times \left[ \frac{\text{numero segmenti } \mathbf{x}^{(h)} \text{ identificati non correttamente}}{\text{numero totale di segmenti } \mathbf{x}^{(h)}} \right] = (\mathsf{TS})$$

$$\text{tasso di rifiuto} = 100 - \left[ (\mathsf{TI}) + (\mathsf{TS}) \right] = (\mathsf{TR})$$

$$\text{tasso di affidabilità} = 100 \times \left[ \frac{(\mathsf{TI})}{100 - (\mathsf{TR})} \right] = (\mathsf{TA})$$

La sperimentazione è stata condotta su tre parlatori in particolare Giulio, Lorenzo e Tony, gli unici per i quali si disponeva di più campioni di segnale vocale. Questi sono stati utilizzati sia in fase di addestramento che di riconoscimento ponendoli a confronto con un insieme di 50 parlatori estratti dalla suddetta base di dati.

### 3.2. *Le caratteristiche utilizzate*

In linea con i già citati lavori svolti nell'ambito del progetto S.M.A.R.T. e con la letteratura corrente su questo argomento (Sambur, 1975; Furui, 1981*a*, 1981*b*, 1994; Reynolds, 1992) si è proceduto ad analizzare il comportamento a fini classificatori delle seguenti caratteristiche:

- 12 coefficienti cepstrali, ottenuti dalla trasformata di Fourier inversa del logaritmo dello spettro del segnale,
- 12 coefficienti MEL-cepstrum, ottenuti considerando i coefficienti cepstrali su scala logaritmica,
- 12 coefficienti di predizione lineare (LPC), ovvero i coefficienti di un modello autoregressivo opportunamente adattato al segnale vocale.

Di seguito riportiamo i risultati ottenuti implementando le singole architetture separatamente su ciascuna caratteristica (Tavole 1A – 1C) ed i risultati

[15] Si noti che l'errore di seconda specie calcolato in Fabi *et al.* (2001) è qui stimato dal tasso di sostituzione mentre l'errore di prima specie ha come limite superiore il tasso di rifiuto utilizzato in questa sperimentazione.

ottenuti combinando le caratteristiche sopra citate sia nell'architettura COMBY che in quelle ME ed HME (entrambe addestrare mediante *early–stopping*[16]).

Appare evidente che la strategia COMBY sia la migliore tra quelle sperimentate (Tavola 2) dato che migliora ME ed HME in tutti i settori, primo fra tutti, il tempo di addestramento necessario (circa *5* ore contro le *7.5* degli ME e le *11* degli HME). Per quanto concerne le singole caratteristiche considerate, la più efficiente a fini classificatori è risultato essere il MEL-cepstrum (Tavola 1C), fatto che conferma quanto osservato nel già citato lavoro di Fabi *et al.* (2001).

TAVOLA 1A

*Risultati delle prove effettuate utilizzando i coefficienti cepstrali*

| | ME (*Early–Stopping*) | | | | HME (*Early– Stopping*) | | | | HME (*Ensemble Learning*) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | TI | TS | TR | TA | TI | TS | TR | TA | TI | TS | TR | TA |
| Giulio | 73.0 | 11.2 | 0.7 | 73.5 | 85.8 | 2.0 | 12.2 | 97.7 | 87.0 | 4.5 | 8.5 | 95.1 |
| Lorenzo | 72.1 | 13.3 | 14.6 | 84.4 | 80.8 | 10.2 | 9.0 | 88.8 | 79.2 | 12.7 | 8.1 | 86.2 |
| Tony | 0.2 | 24.6 | 75.2 | 0.8 | 1.1 | 35.5 | 63.4 | 3.0 | 0.9 | 37.0 | 62.1 | 2.4 |
| Medie | 57.4 | 11.9 | 30.7 | 63.0 | 56.0 | 15.9 | 28.2 | 63.2 | 55.7 | 18.1 | 26.2 | 61.2 |

TAVOLA 1B

*Risultati delle prove effettuate utilizzando i coefficienti di predizione lineare* (*LPC*)

| | ME (*Early–Stopping*) | | | | HME (*Early– Stopping*) | | | | HME (*Ensemble Learning*) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | TI | TS | TR | TA | TI | TS | TR | TA | TI | TS | TR | TA |
| Giulio | 90.2 | 0.1 | 9.7 | 99.9 | 89.8 | 0.0 | 10.2 | 100.0 | 92.3 | 0.7 | 7.0 | 99.2 |
| Lorenzo | 81.7 | 11.1 | 7.2 | 88.4 | 82.0 | 8.1 | 9.9 | 91.0 | 81.6 | 1.5 | 16.9 | 98.2 |
| Tony | 0.2 | 24.6 | 75.2 | 0.8 | 1.3 | 6.9 | 91.8 | 15.8 | 1.9 | 24.1 | 74.0 | 7.3 |
| Medie | 57.4 | 11.9 | 30.7 | 63.0 | 57.7 | 5.0 | 37.3 | 68.9 | 58.6 | 8.8 | 32.6 | 68.2 |

TAVOLA 1C

*Risultati delle prove effettuate utilizzando i coefficienti cepstrali su scala MEL*

| | ME (*Early–Stopping*) | | | | HME (*Early– Stopping*) | | | | HME (*Ensemble Learning*) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | TI | TS | TR | TA | TI | TS | TR | TA | TI | TS | TR | TA |
| Giulio | 90.6 | 0.2 | 9.2 | 99.8 | 91.0 | 2.3 | 6.7 | 97.5 | 89.0 | 0.1 | 10.9 | 99.9 |
| Lorenzo | 84.0 | 0.1 | 15.9 | 99.9 | 82.0 | 5.8 | 12.2 | 93.4 | 88.3 | 4.4 | 7.3 | 95.2 |
| Tony | 0.1 | 19.6 | 80.3 | 0.5 | 0.5 | 44.6 | 54.9 | 1.1 | 1.4 | 27.6 | 71.0 | 4.8 |
| Medie | 58.2 | 6.6 | 35.1 | 66.7 | 57.8 | 17.6 | 24.6 | 64.0 | 59.6 | 10.7 | 29.7 | 66.6 |

[16] Le specifiche dei modelli HME utilizzati (profondità e fattore di ramificazione dell'albero) sono state impostate basandosi su Waterhouse (1997).

TAVOLA 2

*Risultati delle prove effettuate combinando le caratteristiche acustiche*

| | COMBY (circa 5 ore) | | | | ME (*Early– Stopping*) (circa 7.5 ore) | | | | HME (*Early– Stopping*) (circa 11 ore) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | TI | TS | TR | TA | TI | TS | TR | TA | TI | TS | TR | TA |
| 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 |
| Medie | 65.3 | 5.0 | 29.6 | 78.2 | 61.4 | 3.9 | 34.6 | 71.9 | 62.3 | 4.1 | 33.6 | 71.2 |

La sperimentazione da noi condotta mostra inoltre, piuttosto chiaramente, come l'uso di più caratteristiche combinate sia comunque l'alternativa da preferirsi. Confrontando le prestazioni di ME ed HME sul vettore delle caratteristiche combinate con le prestazioni degli stessi sulle singole si nota infatti un miglioramento sia in termini di affidabilità che di tasso di sostituzione. Anche se in via del tutto preliminare, possiamo invece dire che il miglioramento che si ottiene considerando architetture gerarchiche non sembra essere decisivo e tale da giustificare gli incrementi rilevati nei tempi di addestramento.

Bisogna osservare che per il parlatore Tony tutte le architetture e tutte le caratteristiche considerate producono risultati deludenti. La ragione è da ricercarsi nel fatto che il segnale vocale di Tony risulta estremamente saturato. Tale fenomeno induce forti distorsioni nello sviluppo spettrale del segnale e, nel nostro caso, risulta causato dal fatto che i segnali utilizzati sono di tipo telefonico. Tale mezzo, infatti, consente di ottenere una buona registrazione solo quando i livelli del segnale in ingresso rientrano entro una particolare banda di frequenza. Se il parlatore, ad esempio, usa un tono eccessivamente alto, il suo segnale vocale risulterà particolarmente saturato e, di conseguenza, scarsamente informativa sarà un'analisi condotta basandosi, come da noi fatto, su caratteristiche acustiche di tipo spettrale. Occorre comunque sottolineare che, anche in questo caso, la procedura proposta non comporta un errore "troppo grave". Il tasso più alto, infatti, non è quello di sostituzione bensì quello di rifiuto, garantendo in tal modo di non "accusare ingiustamente un innocente".

## 4. CONCLUSIONI E SVILUPPI FUTURI

In conclusione possiamo dire che la direzione scelta, ovvero l'uso di architetture combinate, sembra dare risultati particolarmente incoraggianti ed il completamento della sperimentazione sulla base di dati a disposizione del progetto sarà un primo importante passo sulla strada della validazione del modello. Oltre a ciò, in considerazione di un più efficace trattamento di segnali particolarmente degradati, promettente risulta anche essere l'introduzione nella sperimentazioni di altre tipologie di caratteristiche estraibili dal segnale vocale, quali ad esempio i coefficienti delle espansioni in serie *wavelets* e *wavelets packets*. (Vidakovic, 1999; Kadambe e Srinivasan, 1993; Maes, 1996; Keller *et al.*, 1999).

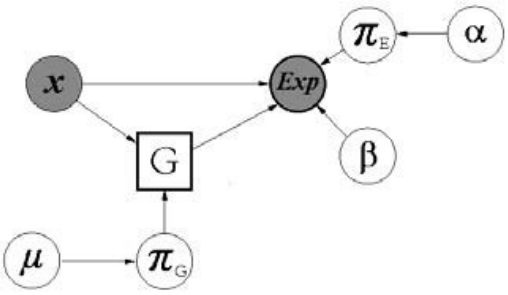


*Figura 5* – Modello grafico di ME bayesiano.

Un primo sviluppo delle metodiche proposte consisterà poi nell'implementazione di altre tecniche di apprendimento per il modello COMBY (almeno a scopo di confronto) quali ad esempio tecniche di stima bayesiana dei parametri (Jacobs *et al.* 1996-1997; Waterhouse, 1997; Waterhouse *et al.*, 1996; Ueda e Ghahramani, 2002; Bishop e Svensen, 2003). Come in Spiegelhalter e Lauritzen (1990), infatti, possiamo introdurre per ciascun nodo del grafo in Figura 4 un nodo ad esso connesso contenente il suo vettore dei parametri. Così facendo, il grafo in Figura 4 muta nel grafo in Figura 5 e tecniche simulative come il *Gibbs sampling* possono essere utilizzate per individuare la distribuzione a posteriori dei parametri dato l'insieme di esempi in addestramento e quindi, tramite integrazione, la predittiva a posteriori. Come sviluppo ulteriore, infine, ci si propone di considerare, data l'intrinseca estensione temporale dei vettori di caratteristiche estratti, le versioni dinamiche dei modelli fin qui implementati (Murphy, 2002; Meila e Jordan, 1995, Weigend *et al.*, 1995). In particolare, piuttosto promettenti sembrerebbero essere gli *hidden markov decision tree* proposti in Ghahramani *et al.* (1997) e rappresentati come modello grafico in Figura 6.

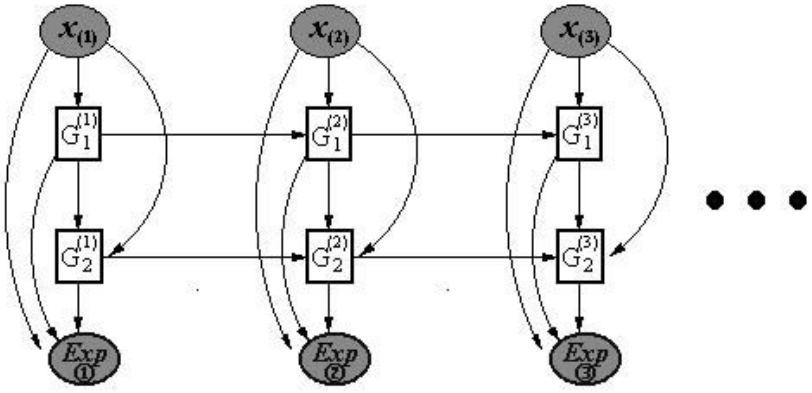


*Figura 6* – Modello grafico di *hidden markov decision tree* (adattato da Ghahramani *et al.* 1997).

*Dipartimento di Statistica, Probabilità e Statistiche Applicate*
*Università di Roma "La Sapienza"*

PIERPAOLO BRUTTI
FRANCESCO FABI
GIOVANNA JONA LASINIO

RINGRAZIAMENTI

Gli autori ringraziano il signor Stefano Delfino, tutti i tecnici della polizia scientifica e la dott.sa Gabriella Fanello Marcucci, unitamente alla direzione di Radio Radicale, per aver reso disponibili alcune registrazioni della trasmissione "Filo Diretto".

RIFERIMENTI BIBLIOGRAFICI

R. AVNIMELECH, N. INTRATOR (1999), *Boosted mixtures of experts: an ensemble learning scheme*, "Neural Computation" 11, pp. 483-497.

R. BELLMAN (1961), *Adaptive Control Processes: A Guided Tour*, Princeton University Press, Princeton, NY.

Y. BENNANI (1995), *A modular and hybrid connectionist system for speaker recognition*, "Neural Computation", 7(4), pp. 791-798.

Y. BENNANI, F. FOGELMAN, P. GALLINARI (1990), *A connectionist approach for speaker identification*, in "Proc. Int. conf. Acoust. Speech, Signal Processing", pp. 265-268.

C. M. BISHOP (1995), *Neural Networks and Pattern Recognition*, Clarendon Press, Oxford.

C. M. BISHOP, M. SVENSEN, (2003), *Bayesian Hierarchical Mixtures of Experts,* Uncertainty in Artificial Intelligence: Proceedings of the Nineteenth Conference, in stampa.

T. BOVE, A. FORTE, P. E. GIUA, C. ROSSI (2001), *Un metodo statistico per il riconoscimento del parlatore basato sull'analisi delle formanti*, "Statistica", LXII (3), pp. 313-328.

L. BREIMAN, J. H. FRIEDMAN, R. A. OLSHEN, C. J. STONE (1984), *Classification and Regression Trees*, Wadsworth.

T. W. CACCIATORE, S. J. NOWLAN (1994), *Mixtures of controllers for jump linear and non-linear plants*, in G. Tesauro, D.S. Touretzky, and T.K. Leen (eds.), "Advances in Neural Informations Processing Systems 6", Morgan Kaufmann, San Mateo, CA.

K. CHEN (1998), *A connectionist method for pattern classification with diverse features*, "Pattern Recognition Letters", 19, pp. 545-558.

K. CHEN, H. CHI (1998), *A method of combining multiple probabilistic classifiers through soft competition on different features sets*, "Neurocomputing", 20(1-3), pp. 227-252.

K. CHEN, D. XIE, H. CHI (1995), *Speaker identification based on hierarchical mixtures of experts*, Proc. World Congress on Neural Networks, Washington D.C., pp. 1493-1496.

K. CHEN, D. XIE, H. CHI (1996*b*), *A modified HME architecture for text-dependent speaker identification*, "IEEE Trans. Neural Networks", 7(5), pp. 1309-1313.

K. CHEN, D. XIE, H. CHI (1996*c*), *Speaker identification using time-delay HMEs*, "International Journal of Neural Systems", 7(1), pp. 29-43.

K. CHEN, L. XU, H. CHI (1996), *Improved learning algorithms for mixtures of experts in multiclass classification*, "Neural Networks", 12(9), pp. 1229-1252.

K. CHEN, L. WANG, H. CHI (1997), *Methods of combining multiple classifiers with different features and their applications to text-independent speaker identification*, "International Journal of Pattern Recognition and Artifificial Intelligence", 11(3), pp. 417-445

C. DOMENICONI, M. JORDAN (2001), *Discorsi sulle reti neurali e l'apprendimento*, FrancoAngeli, Milano.

D. L. DONOHO (1997), *CART and best-ortho-basis: A connection*, "Annals of Statistics", 25 (5), pp. 1870-1911.

R. P. W. DUIN, J. KITTLER, M. HATEF, J. MATAS (1998), *On Combining Classifiers*, "IEEE Trans. Pattern Analysis and Machine Intelligence", 20(3), pp. 226-239.

R. P. W. DUIN, D. M. J. TAX, M. VAN BREUKELEN, J. KITTLER (2000), *Combining multiple classifiers by averaging or by multiplying?*, "Pattern Recognition", 33, pp. 1475-1485.

F. FABI, P. BRUTTI, G. JONA LASINIO (2001), *Una metodologia per il riconoscimento del parlatore basata sulla frammentazione del segnale e sulla classificazione dei frammenti*, sottoposto per la pubblicazione

Y. FREUND, R. E. SCHAPIRE (1997), *A decision-theoretic generalisation of on-line learning and an application to boosting,* "Journal of Computer and System Sciences", 55, pp. 119-139.

J. H. FRIEDMAN (1991), *Multivariate adaptive regression spline*, "Annals of Statistics" 19(1), pp. 1-141.

J. FRITSCH, M. FINKE, A. WAIBEL (1997), *Adaptively growing hierarchical mixtures of experts*, in M. C. Mozer, M. I. Jordan and T. Petsche (eds.), "Advances in Neural Informations Processing Systems 9", MIT Press.

S. FURUI (1981*a*), *Cepstral analysis technique for automatic speaker verification*, "IEEE Tran. Acoust. Speech Signal Processing", 29(2), pp. 254-272.

S. FURUI (1981*b*), *Comparison of speaker recognition methods using statistical features and dynamic features*, "IEEE Tran. Acoust. Speech Signal Processing", 29(3), pp. 197-200.

S. FURUI (1994), *An overview of speaker recognition technology*, in "Proc. ESCA Workshop on Automatic Speaker Recognition, Identification and Verification", pp. 1-9.

Z. GHAHRAMANI, G. E. HINTON (1996), *The EM algorithm for mixtures of factor analyzers*, Technical Report CRG-TR-96-1, Department of Computer Science, University of Toronto, Toronto, Ontario.

D. J. HAND (1997), *Construction and Assessment of Classification Rules*, Wiley series in Probability and Statistics, New York.

T. HASTIE, R. TIBSHIRANI, J. FRIEDMAN (2001), *The elements of statistical learning: Data mining, Inference and Prediction*, Springer-Verlag, New York.

S. HAYKIN (1994), *Neural Networks*, Macmillan College Publishing Company, New York.

G. E. HINTON, D. VAN CAMP (1993), *Keeping neural networks simple by minimizing the description lenght of the weights*, Proceedings of COLT-93.

R. A. JACOBS, M. I. JORDAN, S. J. NOWLAN, G. E. HINTON (1991), *Adaptive mixtures of local experts*, "Neural Computation", 3(1), pp. 79-87.

R. A. JACOBS (1995), *Methods for combining experts' probability assessments*, "Neural Computation", 7(5), pp. 867-888.

R. A. JACOBS, F. PENG, M. A. TANNER (1996), *Bayesian inference in mixtures-of-Experts and Hierarchical Mixtures-of-Experts models with application to speech recognition*, JASA 91, pp. 953-960.

R. A. JACOBS, F. PENG, M. A. TANNER (1997), *A Bayesian approach to model selection in hierarchical mixtures of experts architectures*, "Neural Networks", 10(2), pp. 231-241.

F. V. JENSEN (2001), *Bayesian Networks and Decision Graphs*, Springer-Verlag, New York.

W. JIANG (2000), *The VC Dimension for Mixtures of Binary Classifiers*, "Neural Computation".

W. JIANG, M. A. TANNER (1999*a*), *On the approximation rate of hierarchical mixtures-of-experts for generalized linear models,* "Neural Comp.", 11, pp. 1183-1198.

W. JIANG, M. A. TANNER (1999*b*), *Hierarchical Mixtures-of-Experts for Exponential Family Regression Models: Approximation and Maximum Likelihood Estimation*, "Annals of Statistics" 27, pp. 987-1011.

W. JIANG, M. A. TANNER (1999*c*), *On the Identifiability of Mixtures-of-Experts*, Neural Networks 12, pp. 1253-1258.

W. JIANG, M. A. TANNER (2000), *On the Asymptotic Normality of Hierarchical Mixtures-of-Experts for Generalized Linear Models*, "IEEE Trans. Information Theory", 46(4), pp. 1005-1013.

M. I. JORDAN, R. A. JACOBS (1994), *Hierarchical mixtures of experts and EM algorithm*, "Neural Computation", 6(2), pp. 181-214.

S. KADAMBE, S. SRINIVASAN (1993), *Text-independent speaker recognition system based on adaptive wavelets*, Tech. Rep. Bell-Lab.

K. KELLER, S. BEN YOCOUB, C. MOKBEL (1999), *Combining wavelet-domain hidden markov trees with hidden markov models*, Tech. Rep. IDIAP-RR 99-14.

S. MAES (1996), *Robust speech and speaker recognition using instantaneous frequencies and amplitudes obtained with wavelet derived sinchrosqueezing measures*, in "Program on spline functions and the theory of wavelets", Montreal, Canada.

P. MCCULLAGH, J.A. NELDER (1989), *Generalised Linear Models*, Chapman&Hall, London.

G. MCLACHLAN, T. KRISHNAN (1997), *The EM algorithm and Extensions*, John Wiley & Sons, New York.

G. MCLACHLAN, D. PEEL (2000), *Finite Mixture Models*, John Wiley & Sons, Inc., New York.

M. MEILA, M. I. JORDAN (1995), *Learning fine motion by Markov mixtures of experts*, A.I. Memo No. 1567, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA.

P. MOERLAND, E. MAYORAZ (1999), *DynaBoost: combining boosted hypotheses in a dynamic way*, Tech. Rep. IDIAP-RR 99-09.

P. MOERLAND (2000), *Mixtures Models for Unsupervised and Supervised Learning*, PhD-Thesis IDIAP.

K. P. MURPHY (2002), *Dynamic Bayesian Networks: Representation, Inference and Learning*, Ph.D. thesis, Department of Computer Scince, University of California, Berkeley.

R. M. NEAL, G. E. HINTON (1998), *A new view of the EM that justifies incremental and others variants*, in M.I. Jordan (eds.), Learning in Graphical Models. Kluwer Academic Publishers.

B. NEDIC, H. BOURLARD (2000), *Recent developments in speaker verification at IDIAP*, Tech. Rep. IDIAP-RR 00-26.

J. OGLESBY, J. S. MASON (1990), *Optimization of neural models for speaker identification*, in "Proc. Int. conf. Acoust. Speech, Signal Processing", pp. 261-264.

C. E. PRIEBE (1994), *Adaptive mixture density estimation*, JASA, 89, pp. 796-806.

D. A. REYNOLDS (1992), *A Gaussian mixture modelling approach to text-independent speaker identification*, Ph.D. thesis, Department of Electrical Engineering, Georgia Institute of Technology.

A. RIDA, A. LABBI, C. PELLEGRINI (1999), *Local experts combination through density decomposition*, in "Proceedings of the Seventh International Workshop on Artificial Intelligence and Statistics", 1999. Morgan Kaufmann.

O. ROSEN, W. JIANG, M. A. TANNER (2000), *Mixtures of marginal models*, "Biometrika", 87 (2), pp. 391-404.

L. RUDASI, S. A. ZAHORIAN (1991), *Text-independent talker identification with neural network*, in "Proc. Int. conf. Acoust., Speech, Signal Processing", pp. 389-392.

M. R. SAMBUR (1975), *Selection of acoustic features for speaker identification*, "IEEE Trans. Acoust., Speech, Signal Processing", 23, pp. 176-182.

R. E. SCHAPIRE (1990), *The strength of weak learnability*, "Machine Learning" 5, pp. 197-227.

D. J. SPIEGELHALTER, S. L. LAURITZEN (1990), *Sequential Updating of Conditional Probabilities on Directed Graphical Structures*, "Networks", 20, pp. 579-605.

D. J. SPIEGELHALTER, S. L. LAURITZEN, A. P. DAWID, R. G. COWELL (1999), *Probabilistic Networks and Experts System*, Springer-Verlag, New-York.

M. E. TIPPING, C. M. BISHOP (1997), *Mixtures of probabilistic principal component analysers*, Tech. Rep. NCRG-97-003, Department of Computer Science and Applied Mathematics, Aston University, Birmingham, UK.

D. M. TITTERINGTON, A. F. M. SMITH, U. E. MAKOV (1985), *Statistical Analysis of Finite Mixture Distributions*, John Wiley, New York.

N. UEDA, Z. GHAHRAMANI (2002), *Bayesian Model Search for Mixture Models based on Optimizing Variational Bounds*, "Neural Networks", 15(10), pp. 1223-1241.

B. VIDAKOVIC (1999), *Statistical Modeling by Wavelets*, Wiley, John & Sons Inc., New York.

S. R. WATERHOUSE, A. J. ROBINSON (1994), *Classification using Hierarchical Mixtures of Experts*, in "Proc. IEEE Workshop on Neural Networks for Signal Processing", IV, pp. 177-186.

S. R. WATERHOUSE, A. J. ROBINSON , D. J. C. MACKAY (1996), *Bayesian methods for Mixtures of Experts*, in Touretzky *et al.* (eds.), "Advances in Neural Information Processing Systems 8", MIT Press, pp. 351-357.

S. R. WATERHOUSE (1997), *Classification and Regression using Mixtures of Experts*, Ph.D. Thesis. Department of Engineering, University of Cambridge.

A. S. WEIGEND, M. MANGEAS, A. N. SRIVASTAVA (1995), *Nonlinear Gated Experts for Time Series: Discovering Regimes and Avoiding Overfitting*, "International Journal of Neural Systems" 6, pp. 373-399.

L. XU (1998), *RBF nets, mixtures experts, and Bayesian Ying-Yang learning*, "Neurocomputing", 19, pp. 223-257.

L. XU, M. I. JORDAN (1993), *EM learning on a generalized finite mixture model for combining multiple classifiers*, in "Proc. World Congress on Neural Networks", San Diego, IV, pp. 227-230.

L. XU, M. I. JORDAN, G. E. HINTON (1995), *An alternative model for mixture of experts*, in J.D. Cowan, G. Tesauro, J. Alspector (eds), "Advances in Neural Information Processing System", MIT Press, pp. 633-640.

L. XU, A. KRZYZAK, C. Y. SUEN (1992), *Methods of combining multiple classifiers and their applications to handwriting recognition*, "IEEE Trans. Sys. Man. Cybern. ", 23(3), pp. 418-435.

A. ZEEVI, R. MEIR, V. MAIOROV (1998), *Error bounds for functional approximation and estimation using mixtures of experts*, "IEEE Trans. Information Theory", 44, pp. 1010-1025.

## RIASSUNTO

*Una proposta di meta-analisi basata sulla combinazione di classificatori per il problema del riconoscimento del parlatore*

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

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. Il confronto tra i diversi metodi proposti viene effettuato su di un datbase di oltre 50 segnali vocali (registrazioni telefoniche di parlatori maschi di lingua italiana).

SUMMARY

*Speaker Recognition: a proposal for a meta-analysis based on hierarchical combination of classifiers*

In practical applications of pattern recognition, there are often different features extracted from raw data used for identification. Methods of combining multiple classifiers with different features are viewed as a general problem in various application areas. In this paper some statistical methods based on mixtures of experts models able to combine classifiers and different feature sets are illustrated and applied to the text-independent speaker recognition problem. The comparison takes place on the basis of a dataset composed of more then 50 vocal signals (Italian speakers, phone calls recording).