MINISTERO DELLO SVILUPPO ECONOMICO
DIREZIONE GENERALE PER LA TUTELA DELLA PROPRIETA' INDUSTRIALE
UFFICIO ITALIANO BREVETTI E MARCHI

UIBM

DOMANDA NUMERO	**101995900428613**
Data Deposito	**17/03/1995**
Data Pubblicazione	**17/09/1996**

Sezione	Classe	Sottoclasse	Gruppo	Sottogruppo
G	10	L		

Titolo

SISTEMA DI RICONOSCIMENTO DI PARLATO CONTINUO

DESCRIZIONE dell'invenzione industriale dal titolo:
"Sistema di riconoscimento di parlato continuo"
di: ISTITUTO TRENTINO DI CULTURA, nazionalità italiana, Via S. Croce 77, 38100 Trento
Inventori designati: Giuliano ANTONIOL, Fabio BRUGNARA, Mauro CETTOLO, Marcello FEDERICO
Depositata il: 17 marzo 1995 TO 95A000200

* * *

DESCRIZIONE

La presente invenzione fa riferimento in generale ai sistemi di riconoscimento vocali, in particolare ai sistemi di riconoscimento di parlato continuo, e più specificatamente fa riferimento alla stima del modello del linguaggio e alla sua rappresentazione in un sistema di riconoscimento vocale.

I sistemi di riconoscimento vocale sono ormai abbastanza diffusi nella tecnica e vengono impiegati in una varietà di applicazioni in cui un essere umano può comunicare a voce con un elaboratore elettronico. Tali sistemi consentono infatti di convertire un segnale acustico, la voce umana, in informazioni codificate in formato digitale rappresentanti le parole pronunciate da un parlatore umano. In tal modo è possibile realizzare un'interfaccia uomo-macchina particolarmente semplice, confortevole ed efficiente da

usare, per l'uomo, in sostituzione di dispositivi quali tastiere alfanumeriche e simili.

In sostituzione di tali dispositivi un sistema di questo tipo impiega un sensore acustico, ad esempio un microfono, collegato ad un circuito convertitore analogico-digitale al fine di trasformare il segnale acustico emesso dal parlatore in un formato compatibile per un elaboratore elettronico. Infine un sistema di questo tipo impiega un'unità di elaborazione per decodificare il segnale acustico convertito in un segnale indicativo delle parole pronunciate.

L'unità di elaborazione può essere dedicata a tale compito oppure può venire realizzata utilizzando parte delle risorse elaborative dell'elaboratore elettronico destinato a ricevere il segnale decodificato, in modo ben noto nella tecnica. Tale unità di elaborazione in genere necessita di una potenza considerevole data la complessità del compito che è chiamata a svolgere per cui vi è l'esigenza di migliorare l'efficienza e le prestazioni di tale tipo di sistemi. Un'applicazione caratteristica vede l'uso di tali sistemi come interfacce nelle apparecchiature, e nei sistemi, di dettatura automatica.

La maggior parte degli attuali prototipi di sistemi per il riconoscimento di parlato continuo ese-

guono un processo di decodifica attraverso un algoritmo di ricerca che opera su una rappresentazione dello spazio di ricerca.

Data una rappresentazione di un segnale di ingresso Y, il compito del processo di decodifica è quello di calcolare la stringa (sequenza) di parole W che massimizza il seguente criterio di decisione Bayesiano:

$$\hat{W} = \arg\max_{W} \Pr(W)\Pr(Y|W)$$

dove $\Pr(W)$ è la probabilità a priori, o linguistica, di una stringa W e $\Pr(Y|W)$ è la sua probabilità acustica, ovvero la probabilità che Y corrisponda a tale stringa.

La probabilità acustica di una stringa viene calcolata utilizzando dei particolari modelli stocastici, detti modelli di Markov nascosti (hidden Markov model, HMM), si veda in proposito "A tutorial on hidden Markov models and selected applications in speech recognition" di L.R. Rabiner in "Readings in speech recognition" di A. Weibel e K. Lee, pagg. 267-296, Morgan Kaufmann, 1990, che rappresentano le unità fonetiche della lingua riconosciuta. Ogni parola del dizionario viene modellata tramite una o più sequenze di tali unità.

La probabilità linguistica Pr(W) viene calcolata dal modello del linguaggio tramite un modello stocastico detto a bigrammi. Si veda in proposito "Self-organized language modeling for speech recognition" in "Readings in speech recognition" di A. Weibel e K. Lee, pagg. 450-505, Morgan Kaufmann, 1990,

La ricerca della cosiddetta stringa ottima viene effettuata da un algoritmo di ricerca a fascio, noto anche come beam-search, si veda in proposito H. Ney, D. Mergel, A. Noll, and A. Paesler "Data driven search organization for continuous speech recognition" IEEE transactions on signal processing Vol. 40, No. 2, pp. 272-281, Feb. 1992 su una rete a stati finiti che rappresenta in termini probabilistici l'insieme delle stringhe di parole accettate dal riconoscitore. Durante la ricerca, la valutazione di ipotesi parziali avviene utilizzando sia le probabilità linguistiche che quelle acustiche, ottenute confrontando la rappresentazione del segnale con i modelli delle parole.

Lo scopo della presente invenzione è quello di realizzare un sistema per il riconoscimento di parlato continuo più affidabile ed efficiente, cioè con prestazioni migliorate, rispetto a sistemi di questo tipo secondo la tecnica nota.

Secondo la presente invenzione, tale scopo viene raggiunto grazie ad un sistema di riconoscimento di parlato continuo avente le caratteristiche indicate nelle rivendicazioni che seguono la presente descrizione.

Ulteriori vantaggi e caratteristiche della presente invenzione risulteranno evidenti dalla seguente dettagliata descrizione, effettuata con l'ausilio degli annessi disegni, forniti a titolo di esempio non limitativo, in cui:

- la figura 1 è una rappresentazione schematica di una struttura dati del sistema secondo la presente invenzione, e
- la figura 2 illustra schematicamente un passo dell'algoritmo di costruzione della struttura dati utilizzata dal sistema.

Il sistema secondo l'invenzione utilizza un originale metodo di stima delle probabilità del modello del linguaggio, a partire da un campione di testi, ed una rappresentazione originale dello spazio di ricerca per l'algoritmo di decodifica, che verranno descritti nel seguito.

Naturalmente il sistema secondo la presente invenzione è configurato in modo da eseguire le operazioni e le fasi di elaborazione, cosiddette a basso

livello, del segnale acustico da riconoscere e che non verranno qui descritte in dettaglio in quanto non differiscono da analoghe operazioni svolte da sistemi di questo tipo e sono quindi ampiamente note nella tecnica. La presente descrizione è invece fondamentalmente intesa ad illustrare le componenti originali del sistema secondo l'invenzione.

In particolare le differenze rispetto alla tecnica nota sono relative al modello del linguaggio, alla sua stima e rappresentazione, nel sistema secondo l'invenzione, ragion per cui la presente descrizione verterà sostanzialmente su tali aspetti. Per le componenti e le fasi del sistema qui non esplicitamente descritti si può assumere che siano di tipo noto nella tecnica.

Parimenti sono da considerarsi note le tecniche e le metodologie di realizzazione del sistema secondo l'invenzione per il quale possono essere adottate architetture elaborative di tipo tradizionale ed alla portata di un tecnico esperto del settore.

Stima del modello nel linguaggio

Il modello del linguaggio statistico costituisce la conoscenza linguistica del dominio considerato. La sua funzione è di suggerire all'algoritmo di decodifica le parole più probabili che possono seguire un

certo contesto, in genere costituito da una o più parole. Nel caso in cui il contesto sia formato da una sola parola, il modello del linguaggio utilizzato è quello dei bigrammi, ossia coppie di parole consecutive. Un modello a bigrammi stima la probabilità di un testo, o sequenza, di parole $W = w_1, \ldots, w_N$ mediante la seguente approssimazione:

$$Pr(W) = Pr(w_1) \prod_{t=2}^{N} Pr(w_t|w_{t-1}).$$

Questa approssimazione, che formalmente assume un processo di Markov omogeneo, richiede di stimare la probabilità $Pr(z|y)$ di un generico bigramma yz.

La stima delle probabilità solitamente si basa sui conteggi di bigrammi relativi ad un testo di apprendimento, che rispecchia il più possibile il linguaggio del dominio considerato. Un fenomeno tipico dei testi è però la cosiddetta sparsezza dei dati, che in pratica vuol dire che vi sono molti bigrammi rari e pochi bigrammi molto frequenti. Inoltre, vi sono molti bigrammi possibili che non compaiono mai nel testo di apprendimento ai quali il modello del linguaggio deve comunque assegnare una probabilità non nulla.

Nell'approccio più utilizzato per stimare un

modello del linguaggio a bigrammi si fa uso di:

- una funzione di sconto che toglie dalle frequenze relative dei bigrammi $f(z|y)$ porzioni di probabilità che, messe assieme, costituiscono la probabilità totale $\lambda(y)$ da assegnare ai bigrammi mai visti nel contesto y;
- una funzione di ridistribuzione che suddivide la probabilità totale $\lambda(y)$ tra i bigrammi mai visti nel contesto y;
- uno schema di calcolo che combina le due funzioni precedenti per calcolare la probabilità di un generico bigramma.

Anche se esistono molti modi per calcolare la funzione di sconto, la probabilità dei bigrammi a frequenza zero viene di solito ridistribuita in proporzione alla probabilità a priori delle singole parole, o unigrammi, $Pr(z)$. Quest'ultima probabilità può essere calcolata con metodi tradizionali quali ad esempio la frequenza relativa $f(z)$. Uno schema di calcolo noto in letteratura e qui utilizzato è quello interpolato:

$$Pr(z|y) = \begin{cases} f'(z|y) + \lambda(y)Pr(z) & \text{se } c(y) > 0 \\ Pr(z) & \text{se } c(y) = 0 \end{cases} \qquad (1)$$

in cui con $c(\cdot)$ si indica il numero delle occorrenze

(comparizioni) in un testo.

Secondo lo schema interpolato la probabilità del bigramma è espressa come interpolazione della frequenza relativa scontata $f'(\cdot)$ e della funzione di ridistribuzione.

La funzione di sconto qui utilizzata è quella lineare per la quale:

$f'(z|y) = (1-\lambda(y))f(z|y)$

Un modello interpolato lineare comporta la stima dei parametri $\lambda(y)$ per ogni parola y del vocabolario V.

La stima del modello interpolato lineare si basa sulla combinazione di una tecnica di stima nota come cross-validation, e di un metodo di interpolazione tra stimatori, noto come stacked estimation.

Senza perdere in generalità, il modello interpolato lineare può essere riscritto come:

$$Pr(z|y) = (1-\lambda(y))f(z|y) + \lambda(y)Pr(z) \qquad (2)$$

dove $0 < \lambda(y) \leq 1 \quad \forall y$ e $\lambda(y) = 1$ se $c(y) = 0$. Ciascun parametro $\lambda(y)$ può essere stimato in modo che massimizzi la seguente funzione, denominata "leaving-one-out likelihood", su un testo di apprendimento W:

$$LL = \sum_{y \in V} \sum_{yz \in W} log((1-\lambda(y))f^*(z|y) + \lambda(y)Pr(z)) \qquad (3)$$

dove $f^*(z|y)$ è la frequenza relativa calcolata su W dopo aver tolto una occorrenza di yz. Questo criterio di stima, derivato combinando il criterio di massima verosimiglianza con una tecnica di cross-validation detta Leaving-One-Out, permette di simulare nella funzione di stima i bigrammi mai visti.

Applicando un teorema di Baum ed Egon, si veda in proposito "An inequality with applications to statistical predictions for functions of Markov processes and to a model for ecology" di L.E. Baum e J.A. Egon in Bull. Amer. Math. Soc., 73:360-363, 1967, si può ricavare una formula iterativa per calcolare i valori dei parametri che massimizzano LL localmente, rispetto a dei valori iniziali. La formula iterativa è la seguente:

$$\lambda^{(n+1)}(y) = \frac{1}{|S_y|} \sum_{y z \in S_y} \frac{\lambda^{(n)}(y) Pr(z)}{(1 - \lambda^{(n)}(y)) f^*(z \mid y) + \lambda^{(n)}(y) Pr(z)} \forall y \in V \qquad (4)$$

dove S_y indica l'insieme delle occorrenze dei bigrammi che iniziano con y nel testo di apprendimento. Le iterazioni su ogni parametro vengono controllate secondo un altro criterio di cross-validation.

Di fatto, prima di iniziare l'apprendimento dei parametri, le occorrenze di bigrammi nel testo di apprendimento vengono divise casualmente in due parti,

qui indicate con W_1 e W_2, secondo il rapporto 3:4. La massimizzazione di LL avviene su W_1 e le iterazioni del generico parametro $\lambda(y)$ vengono interrotte se portano ad una diminuzione della verosimiglianza dei bigrammi che iniziano con y nel campione W_2. Al termine dell'apprendimento le frequenze relative vengono riconteggiate su tutto il testo di apprendimento W.

Questa tecnica implica ovviamente un costo aggiuntivo in termini di materiale utilizzato per l'addestramento dei parametri. Una parte consistente del testo di apprendimento viene infatti utilizzata solo per controllare l'algoritmo di massimizzazione.

Metodo di stima stacked

Per risolvere questo problema viene introdotto un metodo di stima originale basato sull'interpolazione di più stimatori.

L'interpolazione di stimatori è una tecnica utilizzata nella teoria della regressione, si veda in proposito "Stacked regressions" di L. Breiman, Technical Report 367, Dept. of Statistics, University of California, Berkeley, Cal. Agosto 1992. Il metodo proposto si ispira a questa tecnica. L'approccio replica ad un diverso livello ciò che avviene per il modello interpolato stesso. Vengono cioè stimati diversi modelli lineari interpolati Pr^1, ..., Pr^m e

quindi combinati come segue:

$$\sum_{i=1}^{m} \alpha_i Pr^i : \alpha_i \geq 0 \; i = 1, \ldots, m \;\; \sum_{i=1}^{m} \alpha_i = 1$$

Ogni modello del linguaggio viene stimato su una diversa partizione casuale del testo di apprendimento nei due insiemi W_1 e W_2 secondo le stesse proporzioni. Al termine gli m modelli del linguaggio così stimati vengono combinati calcolandone la media. Il modello che ne risulta è il seguente:

$$Pr(z \mid y) = \frac{1}{m} \sum_{i=1}^{m} (1 - \lambda^i(y)) f(z \mid y) + \frac{1}{m} \sum_{i=1}^{m} \lambda^i(y) Pr(z)$$

dove λ_i è il vettore di parametri calcolato con l'i-esima partizione del testo di apprendimento.

Il modello che ne risulta ha la stessa forma matematica del modello interpolato semplice, che può essere esteso ad n-grammi con n > 2 e a metodi di combinazione diversi dalla semplice media aritmetica.

Verranno ora illustrati in maggior dettaglio i passi impiegati per stimare il modello del linguaggio a bigrammi qui considerato. Il punto di partenza è sempre un testo di apprendimento che per convenienza può essere considerato come un campione casuale di bigrammi indipendenti e identicamente distribuiti. La

stima utilizza come passo intermedio un algoritmo di stima basato su una cross-validation che necessita di due campioni di apprendimento: uno per la stima dei parametri mediante la formula iterativa (4) ed uno per valutare la condizione di terminazione delle iterazioni.

La stima vera e propria è ottenuta utilizzando questo algoritmo su m partizioni casuali del testo di apprendimento e quindi calcolando la media dei parametri stimati in ciascuna partizione. Verrà ora descritto il primo algoritmo di stima.

Algoritmo di stima con cross-validation (W_1, W_2)

1. Siano W_1 e W_2 due campioni casuali di bigrammi e sia W_2/y il sottoinsieme dei bigrammi in W_2 che iniziano con y
2. Calcola le frequenze relative $f(z|y)$ su W_1
3. Inizializza tutti i parametri $\lambda(y) = 0.5$
4. Per ciascun parametro $\lambda(y)$ itera la formula (4) fintantoché la verosimiglianza di W_2/y calcolata con la formula (3) aumenta.

L'algoritmo di stima con cross-validation viene utilizzato come passo intermedio nell'algoritmo di stima stacked. Il testo di apprendimento viene partizionato casualmente m volte in due sottocampioni di

apprendimento sui quali viene applicato il precedente algoritmo. Si ottengono così m diverse stime dei parametri di interpolazione delle quali viene calcolato il valore medio. Infine, le frequenze relative vengono calcolate su tutto il testo di apprendimento. Quest'ultimo passo completa la stima del modello del linguaggio a bigrammi interpolato lineare.

Algoritmo di stima stacked (W)

1. Sia W il campione casuale di bigrammi nel testo di apprendimento
2. per i = 1, ..., m
3. Calcola una partizione casuale di W in due insiemi W_1 e W_2 secondo la proporzione 2:3
4. Calcola mediante l'algoritmo di stima con cross-validation (W_1,W_2) il vettore di parametri $\lambda^i = \{\lambda(y):y \in V\}$
5. Calcola il vettore medio $\lambda = (1/m)\ \Sigma^m_{i=1}\lambda^i$
6. Calcola le frequenze relative f(z|y) su W.

Rappresentazione del modello del linguaggio

La rete a stati finiti su cui l'algoritmo di decodifica effettua la ricerca della soluzione ottima è costruita imponendo un duplice insieme di vincoli: un insieme acustico, limitando le sequenze di fonemi ammesse a corrispondere alle trascrizioni fonetiche delle parole, ed un insieme linguistico, associando

alle coppie di parole le probabilità stimate tramite la formula indicata con (2). Per una trattazione relativa alle reti a stati finiti si veda "Introduction to Automata Theory, Language and Computation" di J. Hopcroft e J. Ullman, Addison-Wesley, 1979.

I vincoli acustici: l'albero del lessico.

Il primo insieme di vincoli viene imposto in modo da sfruttare la somiglianza acustica delle parole. In effetti, in un vocabolario di dimensioni medio-grandi vi sono molte parole che condividono la parte iniziale della loro trascrizione fonetica. Per questa ragione, l'insieme di parole viene organizzato ad albero.

L'albero ha una radice e tante foglie quante sono le parole del lessico. Gli archi entranti nelle foglie sono etichettati sia con l'ultimo fonema che con la stringa della parola a cui la foglia si riferisce; tutti gli archi rimanenti sono etichettati solamente con fonemi. Per ogni parola del dizionario, esiste un cammino che, a partire dalla radice, passa attraverso archi etichettati secondo la trascrizione fonetica della parola stessa e termina nella foglia che la identifica.

Parole che condividono la parte iniziale della trascrizione fonetica condividono anche il loro cam-

mino fino al punto in cui la trascrizione coincide. Parole omofone, cioè con la stessa trascrizione fonetica, condividono il cammino fino al penultimo arco, restando l'ultimo distinto per permettere la corrispondenza biunivoca tra foglie e parole.

I vincoli linguistici: gli alberi dei successori

Al fine di inserire i vincoli linguistici definiti dal modello del linguaggio nella rete, per ogni parola del dizionario l'insieme dei successori effettivamente osservati nel testo di apprendimento viene organizzato ad albero, esattamente come avviene per l'intero lessico. In questo modo, se la parola y è un successore osservato della parola z, allora l'albero dei successori di z avrà una foglia relativa alla parola y.

Le probabilità fornite dal modello del linguaggio vengono quindi inserite nella rete associandole a degli archi non etichettati, detti per questo vuoti, che connettono l'albero dell'intero lessico e gli alberi dei successori, secondo le modalità descritte di seguito. In figura 1 è rappresentata, per una migliore comprensione ed a titolo esemplificativo, una porzione di rete a stati finiti ad albero per la rappresentazione del modello del linguaggio. In figura con AL è indicato l'albero del lessico mentre con

as(x), as(y), as(z) sono indicati gli alberi dei successori delle parole x, y e z rispettivamente.

Se y è un successore osservato di x allora la probabilità Pr(y|x) viene assegnata ad un arco vuoto che connette la foglia relativa a y dell'albero as(x) dei successori di x con la radice dell'albero as(y) dei successori di y. Ogni foglia dell'albero del lessico completo AL è connessa alla radice dell'albero dei successori della parola che essa identifica, sia essa y, tramite un arco vuoto a cui è associata la probabilità dell'unigramma Pr(y). Dalla radice dell'albero as(y) dei successori di y parte un arco vuoto verso la radice dell'albero dell'intero lessico AL con associata la quantità di probabilità $\lambda(y)$.

La fattorizzazione delle probabilità

Se la ricerca della soluzione ottima è fatta sulla rete della figura 1, l'informazione acustica, associata ai modelli di Markov nascosti con cui i fonemi vengono modellati, e quella linguistica, specificata dagli archi vuoti, vengono utilizzate in zone della rete nettamente distinte.

Allo scopo di utilizzare l'informazione linguistica in anticipo rispetto al punto in cui essa è disponibile nella rete di figura 1, si effettua una fattorizzazione delle probabilità. Quando più parole

condividono un fonema, sia all'interno dell'albero del lessico AL che negli alberi dei successori as(w) è possibile utilizzare la probabilità più alta tra quelle degli archi uscenti dalle foglie che identificano tali parole. In figura 2 è illustrato, per una migliore comprensione, un possibile passaggio di un'operazione di fattorizzazione delle probabilità della rete rappresentata in figura 1, come verrà meglio specificato nel seguito.

La fattorizzazione delle probabilità della rete avviene mediante l'applicazione dell'algoritmo di fattorizzazione delle probabilità, che verrà descritto nel seguito, sull'albero dell'intero lessico AL e sugli alberi dei successori as(w). Tale algoritmo richiede che le probabilità di tutti gli archi siano unitarie eccetto quelle degli archi entranti nelle foglie, vincolate solo ad essere non nulle. In effetti, l'applicazione dell'algoritmo di fattorizzazione delle probabilità è preceduta dallo spostamento delle probabilità del modello del linguaggio all'interno degli alberi, come illustrato in figura 2. In tal modo le probabilità degli archi vuoti uscenti dalle foglie vengono trasferite sugli archi entranti nelle foglie medesime; dopodiché, la fattorizzazione delle probabilità può aver luogo.

L'ottimizzazione della rete

L'uso del massimo tra tutte le probabilità delle parole che condividono un certo fonema comporta l'applicazione del valore corretto del modello del linguaggio non appena la parola non condivide più alcun fonema con altre parole. Inoltre, la fattorizzazione delle probabilità implica che a tutti gli archi vuoti uscenti dalle foglie degli alberi rimanga associato il valore di probabilità 1. Tali archi possono essere quindi eliminati, collassando gli stati da essi collegati.

La rete così ottenuta viene infine ridotta utilizzando uno degli algoritmi noti in letteratura per la minimizzazione del numero di stati di un automa a stati finiti deterministico, ad esempio nel già citato testo "The Design and Analysis of Computer Algorithms" di A. Aho, J. Hopcroft e J. Ullman, Addison-Wesley, 1974. Siccome la rete non è deterministica, data la presenza di archi vuoti, ed è probabilistica, avendo i suoi archi, vuoti e non, associata una probabilità, l'uso di uno di tali algoritmi rende necessaria l'adozione di alcuni accorgimenti.

Innanzitutto, agli archi vuoti va associato un simbolo fittizio in modo che essi siano considerati a tutti gli effetti degli archi etichettati. In se-

condo luogo, dato che questi algoritmi basano il loro funzionamento sull'etichetta associata agli archi, ogni arco viene etichettato con un'unica stringa ottenuta concatenando il simbolo del fonema, la probabilità e, nei casi in cui è presente, la parola.

Algoritmo di fattorizzazione delle probabilità in un albero

```
T: albero da fattorizzare;
a, b, n, s: stati di T;
r: radice di T;
F(n): insieme degli stati successori dello stato n;
p(a, b): probabilità dell'arco da a a b

fattorizza (T)
∀ n ∈ F(r)
   aggiorna (r, n)

aggiorna (a,b)
 se b è foglia di T
    ritorna;
 ∀ s ∈ F(b)
    aggiorna (b,s)
 p:=max {p(b,s):s∈F(b)}
 ∀ s ∈ F(b)
    p(b,s):=p(b,s)/p
 p(a, b):=p
```

Dal modello del linguaggio alla rete

I passi sopra descritti per la costruzione della rete che rappresenta il modello del linguaggio vengono di seguito riportati in forma algoritmica:

1. Costruisci l'albero dell'intero lessico
2. Per ogni parola del lessico, costruisci l'albero dei successori visti nel testo di apprendimento
3. Inserisci le probabilità fornite dal modello del linguaggio tramite delle transizioni vuote
4. Trasferisci le probabilità all'interno degli alberi
5. Fattorizza le probabilità negli alberi
6. Elimina le transizioni vuote superflue
7. Etichetta le transizioni vuote rimanenti con un simbolo fittizio ε
8. Etichetta ogni arco con la stringa ottenuta concatenando il fonema o il simbolo ε, la probabilità e, se presente, la parola
9. Ottimizza la rete
10. Riassegna ad ogni arco il fonema o il simbolo ε, la probabilità ed eventualmente la parola a partire dalla stringa ottenuta al passo 8.

Le soluzioni originali, secondo la presente invenzione, per la stima del modello del linguaggio e per la costruzione della rete con cui il modello del linguaggio viene rappresentato sono state utilizzate dalla richiedente per la realizzazione di un sistema di riconoscimento di parlato continuo, basato su mo-

delli di Markov nascósti. Il dominio applicativo è quello della refertazione radiologica in lingua italiana. L'originale topologia della rete permette di ottenere una contenuta dimensione dinamica del processo di riconoscimento.

Il sistema secondo la presente invenzione, tuttavia, è applicabile in tutti quei settori in cui è riscontrabile una affinità con le problematiche specifiche del riconoscimento di parlato. Ad esempio, lo stesso approccio può essere impiegato per il riconoscimento di caratteri. Le tecniche originali proposte nella presente invenzione sono quindi immediatamente trasferibili in tale ambito.

Più in generale, le soluzioni proposte sono trasferibili in tutti quei settori in cui si effettua una classificazione di sequenze di simboli tali che:

- la classificazione avviene mediante un algoritmo di ricerca a fascio basato su programmazione dinamica;
- le sequenze di simboli sono modellabili da un modello del linguaggio a bigrammi.

Naturalmente, fermo restando il principio dell'invenzione, i particolari di realizzazione e le forme di attuazione potranno essere ampiamente variati rispetto a quanto descritto ed illustrato, sen-

za per questo uscire dall'ambito della presente invenzione.

RIVENDICAZIONI

1. Sistema di riconoscimento di parlato continuo configurato in modo tale da compiere le seguenti operazioni:

- acquisire un segnale acustico comprendente parole pronunciate da un parlatore,
- elaborare detto segnale acustico in modo da generare un segnale indicativo di parametri acustici presenti in detto segnale acustico,
- decodificare detto segnale indicativo di parametri acustici in modo da generare un segnale di uscita indicativo delle parole pronunciate da detto parlatore,

detta operazione di decodifica di detto segnale indicativo di parametri acustici comprendendo un'operazione di confronto con un modello del linguaggio rappresentativo di un linguaggio e con un lessico relativo alle parole pronunciate da detto parlatore,

in cui detto modello del linguaggio è rappresentato mediante una rete a stati finiti ad albero di detto lessico, detta rete a stati finiti essendo una rete probabilistica,

caratterizzato dal fatto che detta rete viene costruita, in una fase preliminare, impiegando un modello del linguaggio interpolato lineare per as-

segnare le probabilità a detta rete.

2. Sistema secondo la rivendicazione 1, caratterizzato dal fatto che detta rete a stati finiti comprende una trascrizione fonetica di dette parole di detto lessico.

3. Sistema secondo la rivendicazione 1 o la 2, caratterizzato dal fatto che detto modello del linguaggio è basato su bigrammi.

4. Sistema secondo la rivendicazione 1 o la 2, caratterizzato dal fatto che detto modello del linguaggio può essere esteso ad n-grammi con $n > 2$.

5. Sistema secondo la rivendicazione 3, caratterizzato dal fatto che impiega, per assegnare ad ogni bigramma la rispettiva probabilità, la seguente funzione:

$$Pr(z|y) = \begin{cases} f'(z|y) + \lambda(y)Pr(z) & \text{se } c(y) > 0 \\ Pr(z) & \text{se } c(y) = 0 \end{cases}$$

essendo $Pr(z|y)$ la probabilità di un generico bigramma yz, essendo $\lambda(y)$ la probabilità totale assegnata ai bigrammi a frequenza nulla nel contesto y, $Pr(z)$ la probabilità a priori di z, $f'(z|y)$ essendo data da:

$$f'(z|y) = (1-\lambda(y))f(z|y)$$

essendo $f(z|y)$ la frequenza relativa del bigramma yz

ed essendo c(y) il numero di occorrenze di y in un segnale acustico campione.

6. Sistema secondo la rivendicazione 5, caratterizzato dal fatto che detto modello interpolato lineare utilizza la seguente funzione:

$Pr(z|y) = (1-\lambda(y)f(z|y) + \lambda(y)Pr(z))$

essendo $0 < \lambda(y) \leq 1 \quad \forall y$ e $\lambda(y) = 1$ se $c(y) = 0$.

7. Sistema secondo la rivendicazione 6, caratterizzato dal fatto che detto modello interpolato lineare comporta la stima di detto parametro $\lambda(y)$ per ogni parola y di detto lessico ed utilizza un procedimento di stima del tipo cross-validation ed un procedimento di interpolazione tra stimatori del tipo stacked estimation per stimare detti parametri $\lambda(y)$.

8. Sistema secondo la rivendicazione 7, caratterizzato dal fatto che ogni parametro $\lambda(y)$ viene stimato in modo che massimizzi una funzione del tipo leaving-one-out likelihood, denominata LL, definita dalla seguente formula:

$$LL = \sum_{y \in V} \sum_{yz \in W} log((1 - \lambda(y))f^*(z \mid y) + \lambda(y)Pr(z))$$

su detto testo di apprendimento, indicato con W, essendo $f^*(z|y)$ la frequenza relativa calcolata sul segnale campione W dopo aver tolto una occorrenza di yz

ed essendo V detto lessico.

9. Sistema secondo la rivendicazione 8, caratterizzato dal fatto che impiega, per calcolare i valori dei parametri $\lambda(y)$ che massimizzano LL localmente rispetto a valori iniziali, la seguente formula iterativa:

$$\lambda^{(n+1)}(y) = \frac{1}{|S_y|} \sum_{yz \in S_y} \frac{\lambda^{(n)}(y) Pr(z)}{(1 - \lambda^{(n)}(y)) f^*(z \mid y) + \lambda^{(n)}(y) Pr(z)} \forall y \in V$$

in cui S_y indica l'insieme delle occorrenze dei bigrammi che iniziano con y in detto testo di apprendimento W.

10. Sistema secondo la rivendicazione 9, caratterizzato dal fatto che prima di iniziare la stima di detti parametri $\lambda(y)$, le occorrenze di bigrammi in detto segnale campione vengono divise casualmente in due parti, W_1 e W_2, sostanzialmente secondo il rapporto 3:4, e la massimizzazione di LL avviene su W_1 e le iterazioni di un generico parametro $\lambda(y)$ vengono interrotte se portano ad una diminuzione della verosimiglianza dei bigrammi che iniziano con y nella parte W_2.

11. Sistema secondo la rivendicazione 10, caratterizzato dal fatto che utilizza un metodo di stima basato sull'interpolazione di più stimatori, in cui

vengono stimati m, con m > 1, modelli lineari interpolati, tra loro differenti, Pr^1, ..., Pr^m e quindi combinati come segue:

$$\sum_{i=1}^{m} \alpha_i Pr^i : \alpha_i \geq 0 \; i = 1, \ldots, m \;\; \sum_{i=1}^{m} \alpha_i = 1.$$

ogni modello del linguaggio essendo stimato su una diversa partizione casuale del testo di apprendimento nei due insiemi W_1 e W_2 secondo le stesse proporzioni.

12. Sistema secondo la rivendicazione 11, caratterizzato dal fatto che i modelli del linguaggio stimati vengono combinati calcolandone la media in modo tale per cui il modello risultante è il seguente:

$$Pr(z \mid y) = \frac{1}{m} \sum_{i=1}^{m} (1 - \lambda^i(y)) f(z \mid y) + \frac{1}{m} \sum_{i=1}^{m} \lambda^i(y) Pr(z)$$

in cui λ_i è un vettore di parametri calcolato con un i-esima partizione del testo di apprendimento.

13. Sistema secondo la rivendicazione 12, caratterizzato dal fatto che, per stimare detti parametri $\lambda(y)$, esegue le seguenti operazioni:

- essendo W un campione casuale di bigrammi
- per i = 1, ..., m
- calcolare una partizione casuale di W in due insiemi W_1 e W_2 secondo una proporzione 2:3

- calcolare mediante un procedimento di stima con cross-validation (W_1,W_2) il vettore di parametri $\lambda^i = \{\lambda(y) : y \in V\}$
- calcolare il vettore medio $\lambda = (1/m)\ \Sigma^m_{i=1}\lambda^i$
- calcolare le frequenze relative $f(z|y)$ su W

detto procedimento di stima con cross-validation comprendendo le seguenti fasi:

- essendo W_1 e W_2 due campioni casuali di bigrammi ed essendo W_2/y un sottoinsieme dei bigrammi in W_2 inizianti con y
- calcolare le frequenze relative $f(z|y)$ su W_1
- inizializzare tutti i parametri $\lambda(y) = 0.5$
- per ciascun parametro $\lambda(y)$ iterare detta formula iterativa fintantoché aumenta la verosimiglianza di W_2/y calcolata con detta formula:

$$LL = \sum_{y \in V} \sum_{yz \in W} log\left((1 - \lambda(y))f^*(z \mid y) + \lambda(y)Pr(z)\right)$$

14. Sistema secondo la rivendicazione 13, caratterizzato dal fatto che detta rete a stati finiti è costruita imponendo due insiemi di vincoli:

un insieme acustico, limitando le sequenze di fonemi ammesse a corrispondere alle trascrizioni fonetiche delle parole,

ed un insieme linguistico, associando alle cop-

pie di parole dette probabilità stimate.

15. Sistema secondo la rivendicazione 14, caratterizzato dal fatto che detto primo insieme di vincoli viene imposto in modo da sfruttare la somiglianza acustica delle parole e l'insieme di parole viene organizzato ad albero.

16. Sistema secondo la rivendicazione 14, caratterizzato dal fatto che detto secondo insieme di vincoli viene imposto in modo tale per cui per ogni parola del dizionario l'insieme dei successori effettivamente osservati nel testo di apprendimento viene organizzato ad albero.

17. Sistema secondo la rivendicazione 15 e la 16, caratterizzato dal fatto che esegue una fattorizzazione delle probabilità di detta rete a stati finiti mediante l'applicazione di un procedimento di fattorizzazione delle probabilità, su un albero dell'intero lessico (AL) e su alberi di successori (as(w)).

18. Sistema secondo la rivendicazione 17, caratterizzato dal fatto che per costruire detta rete rappresentante il modello del linguaggio esegue le seguenti operazioni:

- costruire detto albero dell'intero lessico (AL)
- per ogni parola del lessico, costruire l'albero dei successori comparenti in detto testo di ap-

prendimento

- inserire le probabilità fornite dal modello del linguaggio tramite transizioni vuote
- trasferire le probabilità all'interno degli alberi
- fattorizzare le probabilità negli alberi
- eliminare le transizioni vuote superflue
- etichettare le transizioni vuote rimanenti con un simbolo fittizio
- etichettare ogni arco con la stringa ottenuta concatenando il fonema o il simbolo fittizio, la probabilità e, se presente, la parola
- ottimizzare la rete
- riassegnare ad ogni arco il fonema o il simbolo, la probabilità ed eventualmente la parola a partire dalla stringa ottenuta nell'operazione di etichettatura degli archi.

19. Sistema secondo la rivendicazione 18, caratterizzato dal fatto che detta operazione di fattorizzare le probabilità negli alberi comprende un procedimento costituito dalle seguenti operazioni (si veda la Figura 2):

```
fattorizza (T)
∀ n ∈ F(r)
    aggiorna (r, n)
```

```
aggiorna (a,b)
 se b è foglia di T
     ritorna;
 ∀ s ∈ F(b)
     aggiorna (b,s)
p:=max {p(b,s):s∈F(b)}
 ∀ s ∈ F(b)
     p(b,s):=p(b,s)/p
 p(a, b):=p
```

in cui: T è l'albero da fattorizzare;

a, b, n, s sono stati di T;

r è la radice di T;

F(n) è l'insieme degli stati successori dello stato n;

p(a, b) è la probabilità dell'arco da a a b.

Il tutto sostanzialmente come descritto ed illustrato e per gli scopi specificati.

PER INCARICO

Ing. Mauro MARCHITELLI
N. Iscriz. ALBO 567
(in proprio e per gli altri)

FIG. 1

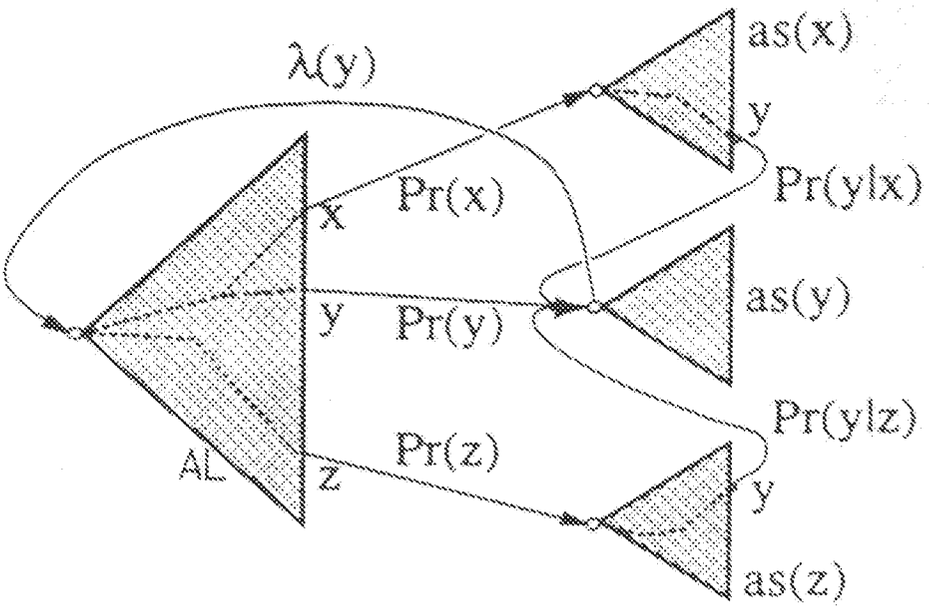
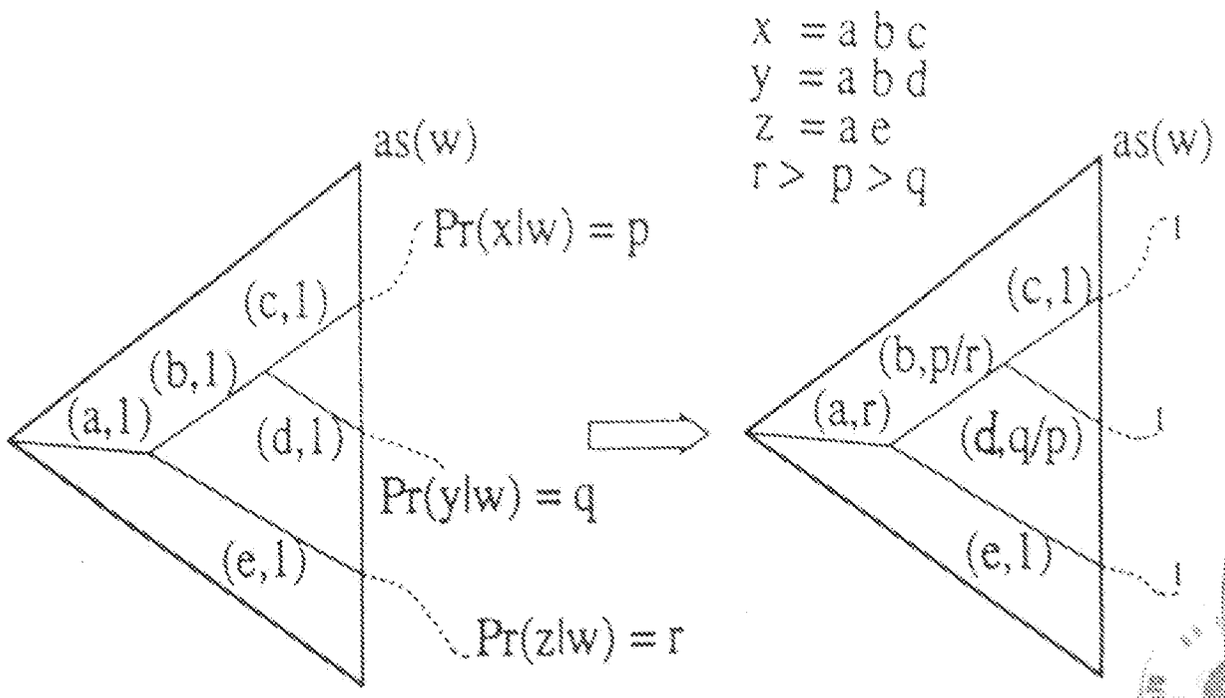

λ(y)
as(x)
y
Pr(y|x)
x
Pr(x)
y
Pr(y)
as(y)
Pr(y|z)
AL
z
Pr(z)
y
as(z)

FIG. 2


x = a b c
y = a b d
z = a e
r > p > q
as(w)
Pr(x|w) = p
(c,1)
(b,1)
(a,1)
(d,1)
Pr(y|w) = q
(e,1)
Pr(z|w) = r
as(w)
1
(c,1)
(b,p/r)
(a,r)
(d,q/p)
(e,1)
1