IBM

note di

informatica 23

Ricerca Tecnologia Applicazioni

Direzione ricerca scientifica e tecnologica
IBM Italia

Sommario

•note di informatica

RicercaTecnologiaApplicazioni

Direzione ricerca scientifica e tecnologica
IBM Italia

23

Riconoscimento della voce: sperimentazione della dettatura automatica per la creazione di testi

Pierluigi Alto, Massimo Brandetti, Marco Ferretti, Giulio Maltese, Federico Mancini, Frank Giovanni Marbello, Anna Maria Mazza, Stefano Scarci, Giuseppe Vitillaro

Sviluppato presso il Centro Ricerca IBM di Roma, il sistema sperimentale per il riconoscimento del parlato in tempo reale e su grandi vocabolari è uscito per la prima volta dal laboratorio ed è stato utilizzato in un ambiente di lavoro.

Introduzione

Negli ultimi anni gli studi sui processi di produzione e percezione della voce hanno portato allo sviluppo di tecnologie in grado di permettere il riconoscimento automatico del parlato. Contemporaneamente l'aumento delle possibilità di integrazione dei componenti offerte dalla microelettronica ha permesso di realizzare dei riconoscitori basati su personal computer.

I sistemi disponibili commercialmente, tipicamente basati sul confronto del segnale vocale in ingresso con delle pronunce di riferimento, hanno in genere la capacità di riconoscere parole isolate estratte da piccoli dizionari (fino a qualche centinaio di parole) con una forte dipendenza dal parlatore.

Presso il Centro Ricerca IBM di Roma è stata sviluppata una macchina per dettare in grado di riconoscere in tempo reale frasi in linguaggio naturale [2]. Il sistema utilizza un vocabolario orientato ad un lessico economico e finanziario contenente più di 20000 parole. Per risolvere i problemi relativi al trattamento di grandi dizionari e all'uso del linguaggio naturale, sono stati utilizzati metodi probabilistici, che hanno permesso di raggiungere ottimi risultati sia per l'analisi acustica sia per l'analisi linguistica.

Studi sui fattori umani legati all'introduzione di uno strumento così innovativo sono stati condotti [3] al fine di mettere in luce la sua validità come strumento per la creazione automatica di testi. Terminata la prima fase di sperimentazione è sorta la necessità di studiare come il riconoscitore viene utilizzato e quale è il livello di accettazione da parte degli utenti durante la normale attività lavorativa. A tale scopo sono state selezionate due applicazioni: una riguardante la dettatura di referti radiologici, l'altra la stesura di polizze di assicurazione.

Un problema che si è posto per la realizzazione dei sistemi di riconoscimento da utilizzare nei due esperimenti è stato quello di adattare il sistema al lessico richiesto dalle diverse applicazioni. Ciò ha comportato lo sviluppo di tecniche ad hoc per permettere un rapido adattamento dei modelli acustico e linguistico del riconoscitore. I prototipi realizzati sono stati

Riconoscimento della voce: sperimentazione della dettatura automatica per la creazione di testi

installati nell'ambiente di lavoro e sono stati utilizzati da utenti senza nessuna precedente esperienza di dettatura automatica per lo svolgimento delle loro normali attività.

Struttura del riconoscitore

Per tenere conto della componente di casualità insita nel segnale vocale, il problema del riconoscimento della voce è stato affrontato, nel prototipo IBM, ricorrendo a modelli probabilistici [1][2][4]. Sia $\overline{W} = w_1, w_2, \ldots w_n$ una sequenza di parole, e sia $\overline{A}$ l'informazione acustica estratta dal segnale vocale sulla cui base il riconoscitore deve individuare le parole pronunciate. $P(\overline{W}|\overline{A})$ è la probabilità che la sequenza di parole $\overline{W}$ sia stata pronunciata, condizionata all'osservazione dell'informazione acustica $\overline{A}$. La sequenza di parole più probabile, data $\overline{A}$, è quella che rende massima $P(\overline{W}|\overline{A})$. Applicando la formula di Bayes possiamo scrivere:

$$P(\overline{W}|\overline{A}) = \frac{P(\overline{A}|\overline{W})P(\overline{W})}{P(\overline{A})} \qquad (1)$$

dove $P(\overline{W}|\overline{A})$ è la probabilità che la sequenza di parole $\overline{W}$ produca l'informazione acustica $\overline{A}$. $P(\overline{W})$ è la probabilità a priori della sequenza di parole $\overline{W}$. $P(\overline{A})$ è la probabilità della sequenza di informazioni acustiche $\overline{A}$. La $\overline{W}$ che massimizza il primo membro è quella che rende massimo il numeratore del secondo: $P(\overline{A})$ infatti non dipende da $\overline{W}$.

Come conseguenza di queste considerazioni il problema del riconoscimento della voce può essere scomposto nei seguenti passi:

1. realizzazione di un'**elaborazione acustica** che estragga dal segnale vocale un'informazione $\overline{A}$ rappresentativa delle caratteristiche acustiche e adatta ad un'analisi statistica;

2. costruzione di un **modello acustico**, che consenta di calcolare la probabilità $P(\overline{A}|\overline{W})$;

3. messa a punto di un **modello del linguaggio** che, indipendentemente dal modello acustico, fornisca la probabilità $P(\overline{W})$;

4. determinazione, mediante una efficiente **strategia di ricerca**, della sequenza di parole che ha la massima probabilità.

Mentre la fase di elaborazione acustica e la strategia di ricerca possono essere considerate indipendenti dall'applicazione, i modelli acustico e linguistico debbono essere modificati a seconda del vocabolario utilizzato.

La stazione di lavoro è costituita da un *front-end acustico*, che provvede all'acquisizione e ad una prima elaborazione del segnale vocale, e da un *PC/AT* equipaggiato con *quattro schede specializzate* che permettono il riconoscimento in tempo reale. Il front-end comprende un microfono, un amplificatore, un convertitore analogico/digitale.

Il segnale vocale viene digitalizzato ad un tasso di 20000 campioni al secondo. Da questa massa di informazioni, elaborata secondo modelli che simulano il comportamento dell'orecchio umano, viene estratto un vettore di 20 parametri acustici ogni 10 ms. Per ridurre ulteriormente la quantità di dati, si considera un insieme di vettori prototipo (nel prototipo IBM sono 200 e ad ognuno è associata un'"etichetta") e si sostituisce ciascun vettore con l'"etichetta" che meglio lo rappresenta. Questa procedura permette di trasformare un'onda acustica in un insieme discreto di "etichette" che *riassumono* le sue caratteristiche principali.

Il sistema di decodifica individua la sequenza di parole più probabile a partire dalla sequenza di etichette acustiche, combinando le informazioni provenienti dal modello acustico e dal modello del linguaggio secondo un'opportuna strategia di ricerca. Questa strategia utilizza una struttura ad albero in cui vengono eliminati i rami che corrispondono alle sequenze di parole più improbabili, secondo opportune soglie, ottenendo così una notevole riduzione delle sequenze da esaminare.

Il modello acustico

Come abbiamo visto nel paragrafo precedente, nell' approccio probabilistico il compito del modello acustico è quello di calcolare la probabilità $P(\overline{A}|\overline{W})$.

Nel prototipo IBM il modello acustico è costituito da **sorgenti di Markov** [2][5]. Una sorgente di Markov è una macchina probabilistica a stati finiti che ben si presta a rappresentare fenomeni altamente variabili che si evolvono nel tempo, quale il processo di produzione della voce. Nella fase di elaborazione acustica il segnale vocale viene codificato in una sequenza di "etichette" che rappresentano in modo semplificato le sue caratteristiche acustiche. In questo caso il modello probabilistico di una parola è realizzato costruendo una sorgente di Markov che rappresenta il processo di produzione di "etichette" acustiche per quella parola. Ad intervalli di tempo fissi la sorgente effettua una transizione, che provoca o meno il cambiamento di stato della macchina, ed emette una "etichetta" acustica. Sia le transizioni che l'emissione delle "etichette" acustiche avvengono in base a distribuzioni di probabilità che dipendono soltanto dallo stato in cui la sorgente si trova e non dalla storia precedente. Mentre è possibile osservare la sequenza di "etichette" prodotta, la sequenza di stati visitati dal modello rimane nascosta. Questi modelli vengono quindi chiamati **sorgenti di Markov nascoste**.

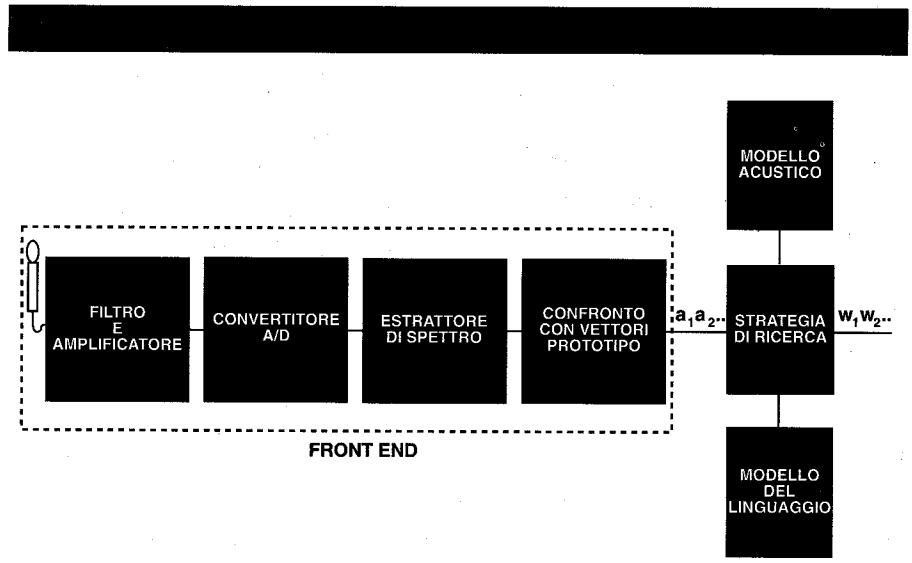
Nel modello acustico del riconoscitore ognuna delle parole appartenenti al vocabolario è rappresentata da un modello di Markov. Data la complessità del modello di ciascuna parola, per la sua costruzione ci si basa su modelli più semplici ottenuti a partire da unità elementari comuni a più parole. Esempi di unità elementari utilizzabili sono: sillabe, difoni, fonemi. Nel nostro caso è stato definito un alfabeto di 56 fonemi (o unità fonetiche) per descrivere i

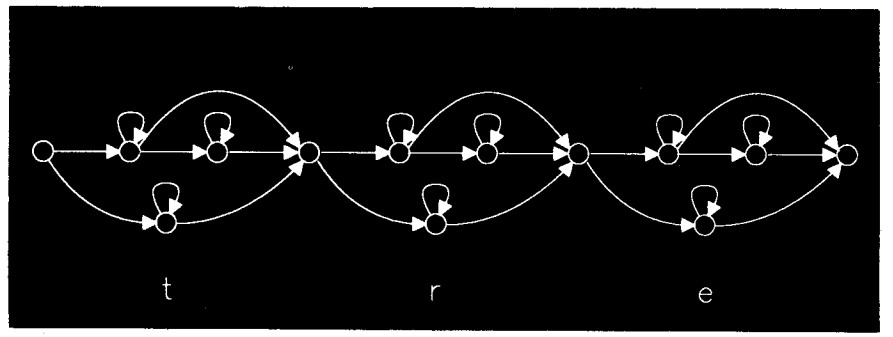
suoni elementari dell'italiano ed è stato associato ad ognuno di essi un modello di Markov che ne rappresenta le caratteristiche acustiche. Il modello di una parola viene costruito concatenando i modelli dei fonemi che la compongono. In modo analogo, il modello della frase $\overline{W} = w_1, w_2, \ldots, w_n$ è costituito dalla concatenazione dei modelli delle parole $w_1, w_1, \ldots, w_n$.

La struttura dei modelli di Markov, determinata dal numero degli stati e dalle loro interconnessioni, è la stessa per tutte le unità fonetiche e per tutti i parlatori. La distinzione tra le diverse unità fonetiche per ciascun parlatore è lasciata interamente ai parametri dei modelli ossia alle probabilità di transizione tra stati e alle probabilità di emissione delle "etichette" acustiche. Il calcolo dei parametri avviene per ogni parlatore durante la fase di *addestramento acustico* ed è reso possibile dalla conoscenza simultanea della sequenza di "etichette" $\overline{A}$ pronunciate e della composizione del testo $\overline{W}$ letto. Il ricorso a unità elementari, quali i fonemi, permette di avere un testo di addestramento indipendente dalle dimensioni del vocabolario. In questo modo non occorre che in fase di addestramento l'utente pronunci tutte le parole del dizionario (che costituirebbe un grosso limite per applicazioni basate su grandi dizionari), ma è necessario soltanto che il testo da leggere contenga più volte tutti i fonemi dell'alfabeto scelto. Una volta noti i parametri, i modelli di ogni parola vengono utilizzati in fase di riconoscimento per calcolare la probabilità $P(\overline{A} \mid \overline{W})$.

Il modello del linguaggio

La funzione del modello del linguaggio è di fornire informazioni sul modo di combinarsi delle parole del vocabolario. Se le parole del vocabolario fossero considerate equiprobabili in ogni contesto, sarebbe impossibile distinguere parole identiche dal punto di vista acustico (per esempio *anno* e *hanno*), e sarebbe difficile riconoscere correttamente parole simili (per esempio *fanno, vanno, sanno*). Il ricorso a modelli grammaticali apporta una notevole semplificazione al problema del riconoscimento poiché limita le possibili combinazioni di parole. I





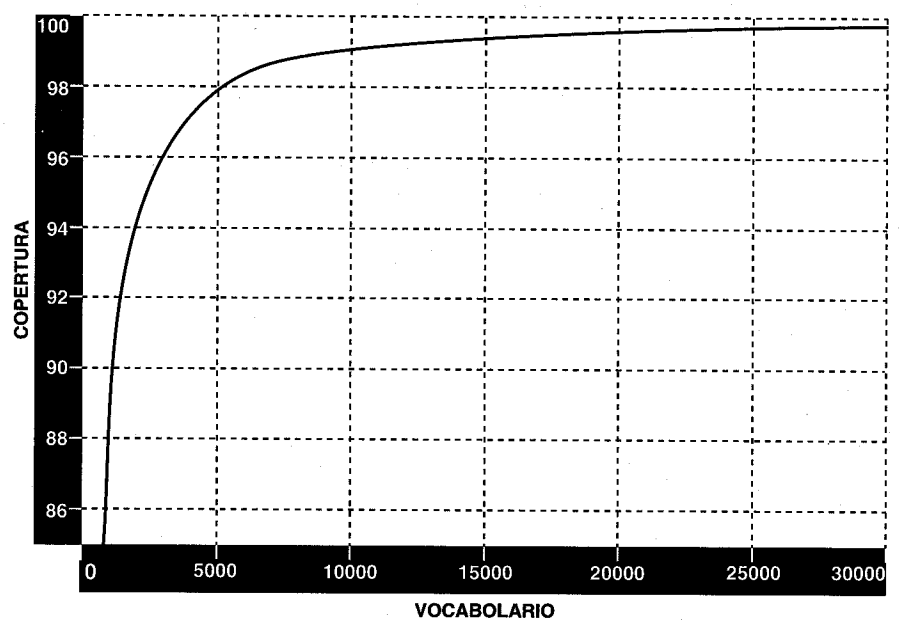


Figura 1. Schema del sistema di riconoscimento probabilistico della voce.

Figura 2. Modello markoviano della parola 'tre' ottenuto dalla concatenazione dei modelli delle unità fonetiche che la compongono.

Figura 3. Copertura del corpus A in funzione della dimensione del vocabolario.

Riconoscimento della voce: sperimentazione della dettatura automatica per la creazione di testi

linguaggi rappresentati con modelli molto rigidi sono però adeguati solo per applicazioni specifiche (per esempio porre semplici interrogazioni ad una base di dati) e non sono utilizzabili per applicazioni basate sul linguaggio naturale.

Nel prototipo IBM è stato seguito un approccio di tipo *statistico-probabilistico* che, basandosi sull'analisi di un vasto insieme di dati (*corpus*), viene utilizzato per:

1. la scelta del vocabolario;
2. la costruzione del modello del linguaggio probabilistico ossia per la stima della probabilità a priori $P(\overline{W})$[7].

Scelta del vocabolario
È evidente che un riconoscitore dotato di un dizionario di dimensione *finita* incontrerà parole a lui sconosciute, per cui potrà costituire un utile strumento di lavoro soltanto se una percentuale molto piccola delle parole che gli vengono dettate gli è ignota. Il primo grosso problema che si pone riguarda allora la scelta di quante e quali parole inserire nel vocabolario. Studi statistici mostrano che è molto vantaggioso fare uso di dizionari specializzati sul tipo di testi che si vuole dettare. Infatti la copertura di testo che si ottiene con un dizionario di dimensioni fissate è tanto maggiore quanto meglio è delimitato il dominio lessicale.

Il dizionario della versione del prototipo IBM basata sul dominio lessicale economico-finanziario è stato costruito considerando le 20000 parole più frequenti riscontrate in un *corpus* di 44 milioni di parole composto da articoli di quotidiani (*Il Sole 24 Ore*), settimanali (*Il Mondo*) e comunicati dell'agenzia *ANSA*. Tale dizionario ha una copertura del 96.5% su testi di economia e finanza non utilizzati per la sua creazione.

Calcolo della probabilità $P(\overline{W})$
La probabilità della sequenza $\overline{W} = w_1, w_2, \ldots, w_n$ può essere espressa come:

$$P(\overline{W}) = \prod_{i=1}^{n} P(w_i \mid w_1, \ldots, w_{i-1}) \quad (2)$$

Il numero di possibili sequenze $w_1, w_2, \ldots, w_n$ è talmente elevato che, anche per vocabolari di poche centinaia di parole, non è possibile stimare la probabilità che la parola w_i verrà pronunciata utilizzando per la predizione ***tutte*** le parole pronunciate prima di essa. Basandosi sulla considerazione intuitiva che le parole più recenti hanno una maggiore influenza su ciò che verrà pronunciato, sono stati utilizzati i *modelli a n-grammi* in cui la probabilità di una parola dipende solo dalle $n-1$ che la precedono. È così possibile considerare equivalenti le frasi che terminano con le stesse $n-1$ parole. Circa il valore da scegliere per n, esso deve risultare da un compromesso tra quantità di dati da trattare ed efficacia predittiva. Nel nostro caso la scelta è stata $n=3$: si parla quindi di modello a *trigrammi* [7].

Il modello di linguaggio dovrà stimare pertanto

$$P(w_i \mid w_{i-1}, w_{i-2}) \quad (3)$$

sulla base del numero di occorrenze delle possibili terne di parole nel corpus di riferimento. L'aspetto più delicato di un modello di questo genere riguarda la stima della probabilità di terne di parole mai viste. Il numero di trigrammi $w_1 w_2 w_3$ possibili è, per un vocabolario con V parole, V^3 (8×10^{11} se $V = 20000$). Poiché l'ordine di grandezza dei testi su cui vengono stimate le statistiche è in genere di $10^7 \div 10^8$, è chiaro che, anche tenendo in considerazione i vincoli grammaticali, il numero di trigrammi possibili è troppo alto per avere una quantità sufficiente di dati di addestramento per ciascuno di essi. È allora necessario in generale ricorrere a tecniche di *smoothing* che consistono nel combinare le informazioni fornite dai trigrammi con altre provenienti dalle distribuzioni di *bigrammi* (coppie di parole) e *unigrammi* (parole isolate).

Due sono le principali tecniche[1] utilizzate per valutare le $P(w_i \mid w_{i-1}, w_{i-2})$:

- *modelli* a *backing-off* in cui le distribuzioni dei trigrammi, bigrammi e unigrammi vengono utilizzate una alla volta per la stima della probabilità della parola successiva;
- *modelli interpolati* in cui tutte le distribuzioni concorrono a determinare tale probabilità:

$$P(w_3 \mid w_1, w_2) =$$

$$\lambda_3 F(w_1 w_2, w_3) + \lambda_2 F(w_2, w_3) + \quad (4)$$

$$\lambda_1 F(w_3) + \lambda_0 \frac{1}{V}$$

I coefficienti combinatori sono tali che:

$$\lambda_3 + \lambda_2 + \lambda_1 + \lambda_0 = 1 \quad (5)$$

Essi vengono stimati con l'algoritmo di *expectation-maximization*, che consiste nel massimizzare la probabilità di un testo tenuto fuori (*held-out*) dai dati di addestramento [8][9].

Una misura importante per giudicare il potere predittivo di un modello statistico di linguaggio è la *perplessità*. Essa indica il numero medio di parole considerate equiprobabili dal modello. In assenza del modello ovviamente la perplessità è pari alla dimensione del vocabolario.

Adattamento del sistema di riconoscimento

Per poter avere un sistema effettivamente utilizzabile al di fuori di un ambiente di laboratorio, bisogna che il riconoscitore sia in grado di adattarsi al lessico che caratterizza i diversi settori applicativi. Si tratta essenzialmente di:

- costruire il vocabolario relativo alla nuova applicazione;
- adattare i modelli acustico e linguistico a tale dizionario.

In vista di applicazioni future, sono

stati ideati degli algoritmi ed è stato messo a punto un insieme di programmi che permette di eseguire questo adattamento facilmente e rapidamente. Mentre però la scelta del vocabolario è strettamente legata alla natura dell'applicazione, al tipo di lessico e soprattutto alla quantità di dati disponibili, l'adattamento dei modelli può in gran parte essere automatizzato. L'obiettivo è di avere uno scheletro di macchina per dettare che possa essere *riempito* facilmente e sia pronto quindi per la decodifica delle parole appartenenti ai nuovi dizionari.

Adattamento del modello acustico

Abbiamo visto come il modello di ogni parola sia ottenuto a partire dai modelli delle unità fonetiche che la compongono. Il primo passo da compiere quando si vuole adattare il riconoscitore ad un nuovo vocabolario è, quindi, quello di effettuare la trascrizione fonetica di tutte le parole. Allo scopo di limitare il numero di trascrizioni fonetiche da effettuare per una nuova applicazione è stata costruita una *base di dati* contenente tutte le parole e le relative trascrizioni fonetiche che hanno fatto parte di vocabolari utilizzati per altre applicazioni. Effettuando una ricerca nella base di dati è possibile individuare l'insieme di parole per le quali non esiste già la trascrizione fonetica.

Di solito la trascrizione fonetica è un processo che viene compiuto in modo completamente manuale risultando così costoso e facilmente soggetto ad errori. Per questa ragione si è cercato di realizzare degli strumenti automatici che rendessero tale processo più rapido ed affidabile. I sistemi proposti per la trascrizione fonetica automatica sono basati su regole o su apprendimento automatico da un insieme di dati di addestramento. Questi sistemi non permettono però di raggiungere l'accuratezza desiderata per un riconoscitore della voce. Ciò è dovuto alla intrinseca complessità del problema ed alla difficoltà di descrivere, tramite un limitato insieme di regole, tutti i casi possibili. L'approccio seguito per risolvere il problema è stato quello di separare la conoscenza fonetica da quella lessicale [6]. Infatti, mentre la prima può essere descritta da un limitato insieme di regole, la seconda è basata in gran parte sull'esperienza umana e non può essere facilmente formalizzata. Il sistema costruito è in grado, data la stringa ortografica che rappresenta la parola, di produrre un insieme di possibili trascrizioni fonetiche che si ottengono applicando le regole per il passaggio da forma ortografica a forma fonetica. La scelta della trascrizione corretta, date tutte le alternative, avviene manualmente. Un insieme di 78 regole permette di descrivere le ambiguità esistenti nell'italiano per la traduzione della parola dalla forma ortografica a quella fonetica. Ogni regola è costituita da una parte sinistra ed una parte destra. La parte sinistra è composta da una stringa ortografica, la parte destra è costituita dalle possibili trascrizioni fonetiche della stringa. Le possibili trascrizioni alternative prodotte da questo insieme di regole vengono sottoposte ad una successiva analisi mediante un insieme di regole globali che eliminano le trascrizioni impossibili, ad esempio quelle che contengono più di una vocale accentata. In media il numero di trascrizioni prodotte per parola è 5. La trascrizione fonetica corretta è sempre presente nelle alternative che vengono presentate all'operatore. Quest'ultimo, sulla base della conoscenza lessicale e semantica della parola, è in grado di selezionare facilmente la trascrizione corretta che verrà utilizzata per rappresentare la pronuncia della parola nel riconoscitore. Questa tecnica si è dimostrata estremamente efficace ed ha permesso un rapido adattamento del modello acustico del riconoscitore della voce al nuovo vocabolario.

Adattamento del modello del linguaggio

La tecnica utilizzata per la costruzione di un modello del linguaggio probabilistico è quella vista nei paragrafi precedenti. Le probabilità $P(w_i \mid w_{i-1}, w_{i-2})$ vengono stimate con tecniche che si basano sull'interpolazione di più informazioni con pesi proporzionali al loro grado di attendibilità (formula 4). Essendo queste tecniche indipendenti dal vocabolario, è stata possibile la realizzazione di programmi che automaticamente, a partire da un dizionario e da un *corpus*, generano il modello del linguaggio.

Sperimentazione del sistema su applicazioni diverse

Sono stati effettuati numerosi esperimenti con diversi parlatori per valutare le prestazioni del sistema. Il vocabolario base utilizzato è rivolto a un dominio linguistico di tipo economico-finanziario. I risultati ottenuti (tasso di errore del 2-3% e velocità di immissione di circa 60 parole al minuto, equivalenti a circa 400 battute dattiloscritte), hanno dimostrato che il sistema costituisce una valida alternativa ai mezzi normalmente utilizzati per la creazione di testi. Per poter misurare il grado di accettabilità di un sistema così innovativo in un ambiente di lavoro, è stato necessario trovare delle applicazioni che per le loro caratteristiche ben si prestassero alla dettatura automatica. Le due applicazioni selezionate riguardano la dettatura di referti radiologici e la stesura di polizze di assicurazione. La prima è stata condotta in collaborazione con il Prof. Bruno Vidal, Primario del reparto di Radiologia dell'Ospedale "S. Maria della Pietà" di Udine; per la seconda ci si è avvalsi della collaborazione dell'Ingegner Angelo Santangelo, responsabile del Centro Elaborazione Dati della Compagnia di Assicurazioni "Lloyd Adriatico" di Trieste.

Poiché il riconoscitore è basato su un vocabolario predefinito, la scelta del dizionario ne condiziona in modo essenziale l'utilità. Il primo passo dunque è stato quello di definire un vocabolario per ciascuna applicazione e successivamente adattare il sistema a tale vocabolario utilizzando la struttura descritta nei paragrafi precedenti.

Il lessico del vocabolario base (economia e finanza) è molto *distante* da quello impiegato nella dettatura di referti medici, mentre mostra una maggiore affinità con i testi prodotti dalla compagnia assicuratrice. Nel primo caso si è quindi ritenuto necessario ricostruire interamente il dizionario basandosi esclusivamente su un *corpus* di referti radiologici, nel secondo caso si è costruito un dizionario misto, basato cioè su quello economico-finanziario e arricchito con le parole più frequenti del *corpus* applicativo. Vediamo il processo che ha portato alla definizione dei vocabolari e dei modelli di linguaggio per i due esperimenti, focalizzando l'attenzione sulle principali caratteristiche dei dati che sono stati

analizzati e sui criteri utilizzati per la selezione delle parole.

Applicazione 1: dettatura di referti radiologici
Per la definizione del vocabolario costruito per la dettatura dei referti radiologici, è stato analizzato un corpus di 5 milioni di parole (*A*) costituito da referti raccolti presso 4 ospedali e un corpus di 50000 parole (*B*) raccolto presso il reparto di radiologia che ha ospitato l'esperimento.

Per la scelta della dimensione del vocabolario ci si è basati sull'analisi dell'andamento della copertura del testo in funzione del numero di parole del dizionario (figura 3). Per questo tipo di lessico, e con la quantità di dati disponibile, un vocabolario di 5000 parole rappresenta un buon compromesso fra l'esigenza di avere elevati valori di copertura e quella di avere un sufficiente quantitativo di dati per la stima dei parametri del modello del linguaggio.

Allo scopo di ottenere una buona personalizzazione del sistema, si è scelto di inserire nel vocabolario tutte le 3200 parole distinte individuate nel corpus *B*. Il vocabolario è poi stato completato con altre 1900 parole scelte fra le più frequenti del corpus *A* e non incluse nelle precedenti 3200.

Il dizionario così costruito ha una copertura del 100% sui referti di tipo *B* e una copertura del 98% sui referti di tipo *A*.

Applicazione 2: stesura di polizze assicurative
Per estrarre le parole del vocabolario sono stati analizzati i testi prodotti da una compagnia di assicurazioni. Tali testi sono stati raccolti presso 9 dipartimenti e sono relativi ad argomenti differenti: corrispondenza, stipula di polizze di vario tipo, documenti di diversa natura, ecc.

L'analisi della copertura incrociata fra i 9 corpora, divisi per dipartimento, ha mostrato come questi facciano uso di lessici fra loro molto distanti. Si è pertanto ritenuto opportuno basare il dizionario sui testi provenienti da un solo dipartimento, costruendo così un'applicazione mirata per quel tipo di lessico.

Il corpus selezionato come il più adatto per la nostra applicazione ha 1.5 milioni di parole e una buona affinità con il lessico economico-finanziario (copertura del dizionario economico-finanziario pari al 95.2%).
Per la definizione del nuovo vocabolario sono state selezionate le 15000 parole più frequenti del dizionario base e a queste sono state aggiunte le 3100 parole più frequenti del corpus applicativo non incluse nelle precedenti 15000.

Il dizionario così costruito ha una copertura del 99% sui testi relativi alla applicazione e una copertura del 95.7% su testi di tipo economico-finanziario.

Risultati

Le due applicazioni per la dettatura di referti radiologici e di polizze di assicurazione sono state realizzate sperimentando il sistema direttamente nell'ambiente lavorativo. Vengono qui di seguito riportati i risultati di entrambi gli esperimenti. In ambedue i casi si tratta di risultati ancora non definitivi, ma che possono già dare un quadro rappresentativo delle prestazioni del sistema.

Dettatura di referti radiologici
Quattro medici hanno partecipato all'esperimento. Senza possedere particolari conoscenze informatiche, essi hanno mostrato notevole adattabilità al sistema, sia come capacità di utilizzare le varie funzioni sia per quanto riguarda l'uso della dettatura in **parlato connesso** (inserzione di una brevissima pausa tra le parole).

Sono stati dettati 152 referti per un totale di oltre 12000 parole. La copertura del vocabolario relativamente alle parole dettate è stata del 98.7%. Nella tabella 2 sono riportati, per ogni parlatore, i risultati complessivi dell'esperimento. Il tasso di errore netto si riferisce agli errori effettivamente attribuibili al riconoscitore. Il tasso di errore umano si riferisce ad errori dovuti ad un uso errato della macchina (mancato inserimento delle pause tra le parole, errori nell'uso dei comandi).
Una stima globale del comportamento del sistema si può avere sommando al tasso di errore netto la percentuale di errori dovuti all'assenza della parola nel vocabolario che, come visto, è stata in media dell'1.3%. In tal modo si ottengono dei tassi di riconoscimento variabili tra il 93.7% ed il 97% (o tra il 90.5% ed il 95.6% se si tiene conto anche degli errori umani).

In realtà, essendo un referto lungo un centinaio di parole, la correzione di 4-5 parole in media in un testo così breve si è rivelata un'operazione assai poco onerosa. Tra l'altro l'errore risultava facilmente individuabile sia perché il medico lo notava durante la trascrizione sul video, sia perché il significato della parola errata risultava in genere estraneo al contesto dettato.

Dopo circa tre mesi di utilizzo del sistema da parte dei medici dell' Istituto di Radiologia, si può affermare che l'integrazione in un contesto *produttivo* di tale macchina ha permesso di raggiungere ottimi risultati. In particolare ecco i vantaggi rilevati nell'ambito specifico della dettatura dei referti radiologici:

- sono stati notevolmente abbassati i tempi necessari per la stesura definitiva dei referti (da 2 giorni a pochi minuti);
- è possibile un maggiore controllo poiché la stesura definitiva del referto avviene a diagnosi appena effettuata avendo le lastre a disposizione;
- il sistema si è rivelato estremamente utile per gli esami di pronto soccorso notturni e festivi, quando non è di solito disponibile il servizio di dattilografia.

Dettatura di polizze di assicurazione
In questo caso sono disponibili risultati su tre parlatori. Sono state dettate oltre 8000 parole con una copertura del vocabolario intorno al 99%. A differenza dell'esperimento precedente, i parlatori avevano già esperienza nell'uso di sistemi informatici. I risultati dell'esperimento sono riportati nella tabella 3 e confermano l'affidabilità del sistema quando questo viene utilizzato con un lessico centrato sull'applicazione richiesta (si osservi il basso valore di perplessità per questa applicazione riportato in tabella 1).

Questo esperimento è stato molto utile per studiare la effettiva applicabilità del sistema nel campo assicurativo e d'ufficio in generale. In questo settore i testi dettati sono piuttosto lunghi e gli errori non sempre sono facilmente individuabili. Prescindendo dalla fase di correzione, il vantaggio maggiore ottenuto utilizzando il sistema nella stesura di documenti, è stato la possibilità di creare *velocemente* un testo e di renderlo quindi disponibile immediatamente su un supporto magnetico. Il testo è suscettibile poi di essere modificato e/o corretto successivamente per la stesura finale. I consensi maggiori di un tale sistema in un ambiente di ufficio, sono giunti dagli utenti che quotidianamente devono scrivere relazioni o documenti ma non dispongono di un servizio di dattilografia.

Tabella 1

Caratteristiche dei modelli di linguaggio per le due applicazioni.
I dati di addestramento (utilizzati per stimare i parametri del modello di linguaggio) sono espressi in milioni di parole ed i numeri di bigrammi e trigrammi in milioni.
Nell'applicazione riguardante la dettatura di polizze assicurative, è stato usato un corpus di 1.5 milioni di parole tipico del dominio lessicale dell'applicazione. Esso è stato integrato mediante un corpus di 40 milioni di parole tratte dal dominio economico-finanziario. Ciò è stato possibile per l'elevata copertura incrociata (95.2%) tra i due domini linguistici.

Parametro	Dettatura referti	Dettatura polizze
Dimensione del vocabolario	5100	18100
Dati di addestramento	4.8	1.5+40
Trigrammi diversi	0.62	3.4
Bigrammi diversi	0.19	2.3
Perplessità	38	18

Tabella 2

Risultati dell'esperimento di dettatura di referti radiologici.

Parlatore	Referti dettati	Tasso errore netto	Tasso errore umano
p1	25	1.7%	1.4%
p2	12	2.0%	1.2%
p3	79	3.5%	0.9%
p4	36	5.0%	3.2%

Tabella 3

Risultati dell'esperimento di dettatura di polizze di assicurazione.

Parlatore	Tasso errore netto	Tasso errore umano
p1	1.9%	1.0%
p2	1.4%	0.5%
p3	2.4%	0.7%

Conclusioni

Nonostante i risultati esposti siano da considerare preliminari, possono essere tratte utili indicazioni.
È da notare come i vari utenti abbiano rapidamente imparato ad usare la macchina per dettare. Ciò si può desumere dai bassi valori del tasso di errore umano che, tranne in un caso, si sono sempre mantenuti sotto il 2%.
Il tasso di errore netto è stato soddisfacente. Il numero di errori nei testi dettati dovuti alla mancanza di parole nel vocabolario è risultato molto contenuto. Un utile risultato degli esperimenti è stato quello di raccogliere indicazioni su come ampliare i vocabolari in modo tale da estendere il dominio lessicale delle applicazioni e su come migliorare l'interfaccia utente e quali funzioni sia necessario aggiungere per rendere funzionale l'uso del sistema nei vari settori applicativi.

Ringraziamento
Si desidera ringraziare il Prof. Bruno Vidal, Primario del reparto di Radiologia dell'Ospedale "S. Maria della Pietà" di Udine e l'Ingegner Angelo Santangelo, responsabile del Centro Elaborazione Dati della Compagnia di Assicurazioni "Lloyd Adriatico" di Trieste
per aver fornito la collaborazione che ha permesso di realizzare le applicazioni descritte.

Bibliografia

[1] F. Jelinek, **The Development of an Experimental Discrete Dictation Recognizer,** *Proceedings of IEEE,* vol. 73, no. 11, novembre 1985, pp. 1616-1624.

[2] P. D'Orta, M. Ferretti, A. Martelli, S. Melecrinis, S. Scarci, G. Volpi, **Vocabulary Speech Recognition: a System for the Italian Language,** *IBM Journal of Research and Development,* Vol. 32, No. 2, marzo 1988, pp. 217-226.

[3] P. Alto, M. Brandetti, M. Ferretti, G. Maltese, S. Scarci, **Esperimenti di dettatura in linguaggio naturale con un riconoscitore per 20000 parole,** *Atti del Congresso annuale AICA, Trieste, 4-6 ottobre 1989,* vol. 1, pp. 363-369.

[4] P. D'Orta, M. Ferretti, S. Scarci, **Riconoscimento della voce per la dettatura automatica di testi,** *Atti del Congresso annuale AICA, Trento, 30 settembre - 2 ottobre 1987,* vol. 2, pp. 111-128.

[5] L.R. Rabiner, B.H. Juang, **An Introduction to Hidden Markov Models,** *IEEE ASSP Magazine,* vol. 3, no. 1, gennaio 1986, pp. 4-16.

[6] S. Scarci, S. Taraglio, **Automatic Phonetic Transcription for Large-Vocabulary Speech Recognition,** *Speech 88, Seventh FASE Symposium,* Edinburgh, 22-26 agosto 1988, pp. 771-777.

[7] F. Jelinek, **Self Organized Language modeling for Speech Recognition,** *IBM Europe Institute 1986, Advances in Speech Processing,* Oberlech, Austria, 14-18 luglio 1986.

[8] F. Jelinek, R.L. Mercer, **Interpolated Estimation of Markov Source Parameters from Sparse Data,** *Pattern Recognition in Practice,* E.L.Gelsema and L.N. Kanal, Ed., North-Holland, New York, 1980, pp. 381-397.

[9] A.P. Dempster, N.M. Laird and D.B. Rubin, **Maximum Likelihood from Incomplete Data via the EM algorithm,** *J. of Roy. Stat. Soc., no. 1, pp. 1-38,* 1977.