L'Intelligenza Artificiale

di Raffaello De Masi

Riconosciamo la lingua parlata

Seconda parte

La volta scorsa ci siamo fermati a stabilire quali fossero i parametri iniziali della lingua parlata destinati a funzionare da chiave di riconoscimento: procediamo su questa strada, e con l'aiuto dell'ottimo lavoro di Pieraccini vediamo come sia possibile procedere, praticamente al riconoscimento stesso.

Come acutamente fa notare l'autore, neppure l'uomo ha la capacità di comprensione di un oggetto se prima non ne è venuto a conoscenza, o, almeno, non ne ha ricevuto una descrizione formale. È necessario, quindi, fornire alla macchina, inizialmente, un dizionario, sotto forma di parole pronunciate dall'utente; tali parole, i prototipi del riconoscimento, faranno parte del patrimonio di base della macchina.

Tali campioni, se così è lecito chiamarli, verranno immagazzinati ed utilizzati come termini di confronto: in fase di riconoscimento (supponiamo di aver superato i problemi descritti nella puntata precedente); una parola pronunciata dall'utente verrà confrontata con il patrimonio di base, e riconosciuta e rigettata.

Vediamo come è possibile eseguire la fase di riconoscimento: la parola viene suddivisa in finestre, ognuna costituita da intervalli eguali di tempo (10 millisecondi). L'analisi di ogni finestra consente di calcolarne lo spettro di energia emesso (ecco la necessità che sia lo stesso utente a fornire alla macchina il suo vocabolario di base); la rappresentazione finale della parola sarà data da un punto individuato da una serie di coordinate nello spazio ad N dimensioni, dove N è il numero di finestre in cui è stata divisa la parola stessa. Dall'esame di tali coordinate è possibile dedurre la somiglianza o meno del suono analizzato col vocabolario di base. È possibile, cioè, confrontare lo stereogramma della parola analizzata con tutti quelli presenti in memoria, e ricavare, per confronto, il significato più probabile di una parola, e, quindi, ottenere il suo riconoscimento.

Il lettore attento avrà subito diagnosticato il lato debole della configurazione: una parola, anche se pronunciata dalla stessa persona, ben difficilmente avrà sempre la stessa durata (specie se come misura viene utilizzata la finestra dei dieci millisecondi). Vale a dire che la stessa parola, pronunciata con cadenze diverse o in una frase di diverso senso, alla fine darà configurazioni parametriche diverse. È ovvio, infatti, che l'allungamento della durata di pronuncia della parola porterà alla completa non corrispondenza tra pronuncia e giacenza in vocabolario dello stesso vocabolo, anzi è più che certo che la differenza sarà ben più ampia di quella esistente tra due parole diverse. La soluzione sta nel rendere elastico lo spazio di rappresentazione della funzione n-dimensionale della parola stessa: vale a dire che, poiché è estremamente improbabile che la differenza di lunghezza della pronuncia possa riferirsi solo ad una posizione di parola, è sufficiente prevedere, nel programma di analisi la possibilità di allineare e confrontare rappresentazioni simili (proporzionali) per risolvere, almeno in linea teorica il problema. L'operazione può essere eseguita, con successo, più che confrontando i valori finali, allineando le due rappresentazioni (in termini di finestre) e scegliendo quella configurazione che meglio fa corrispondere configurazioni simili. Operazione non banale e senz'altro complessa, dal punto di vista anche del tempo di intervento, se non si potesse introdurre una semplificazione, sotto forma di una tecnica detta «allineamento temporale dinamico», frutto delle esperienze di H. Sakoe e S. Chiba.

Questa tecnica considera lo spazio (in cui viene rappresentata la parola, come abbiamo descritto precedentemente) ndimensionale come occupato da un reticolo, anch'esso di pari dimensioni, comprendente un numero finito di punti nodali. Ciascun punto mette in corrispondenza la parola analizzata con una configurazione diversa. Poiché il punto iniziale e finale del reticolo (punti estremi dell'inviluppo) sono rappresentativi del momento d'inizio della pronuncia della parola e della sua fine, tutti i possibili rami del reticolo rappresentano le distanze, in ordine di tempo, rappresentanti le finestre (e in maniera più grossolana, i fonemi). Orbene, poiché ad ogni punto nodale corrispondono n vie diverse, (dove n sono funzione di diverse variabili, come cadenza, fretta di pronuncia, ecc.), verrà scelta la frazione di reticolo successivo più simile a quella corrispondente la parola da analizzare.

Data l'altissima variabilità delle funzioni in gioco, risulta estremamente improbabile che, in base ad essa, possa verificarsi l'errore, inteso come riconoscimento di similarietà tra parole diverse.

Il problema, così risolto in via teorica, all'atto pratico risulta quanto meno oneroso per l'elevata mole di calcoli da eseguire (ancora dallo stesso autore citiamo come l'analisi di una qualsiasi frase richiede un ritmo di calcolo di oltre 16 milioni di operazioni al secondo); tutto ciò sarebbe irrealizzabile se non intervenisse, ancora una volta, un aiuto che sta a metà tra lo statistico e l'empirico. Infatti risulta inutile esplorare tutto l'albero nodale (ad esempio, è del tutto assurdo che in una parola, fenomeni successivi siano ai lati opposti come durata del tempo di pronuncia); inoltre si è notato che ben difficilmente, tranne che per scopi specifici e voluti, si hanno variazioni

notevoli nella durata totale della pronuncia della parola: in tale ipotesi il reticolo può essere esplorato solo in una fascia prossima alla sua diagonale senza perdere molto nella possibilità di riconoscimento.

Esistono, inoltre, possibilità di ridurre, ancora, tali fenomenologie piuttosto complesse: si è visto, infatti, che, generalmente, in una parola, esistono zone di stazionarietà vocale, che quasi mai differiscono anche in sequenze fonetiche diverse. Da questo si è passati alle configurazioni a finestra variabile, dove la lunghezza della finestra dipende anche dalla posizione che occupa nella parola stessa.

Ma stavamo dimenticando il problema principale. Poiché il parlato con scansione delle parole è piuttosto innaturale, e non sempre efficace, data la presenza, sovente, del rumore di fondo, come è possibile realizzare il riconoscimento di un discorso fluente?

Se si potesse essere in grado di riconoscere e codificare il punto di separazione tra una parola ed un'altra, presenti in una frase, si potrebbe adottare la tecnica precedente con buoni risultati. Purtroppo ciò non è praticamente mai possibile (risulta addirittura ancora impossibile capire come il cervello umano possa farlo, figuriamoci una macchina), anche per la presenza del fenomeno della coarticolazione, in base alla quale il fonema iniziale di una frase si fonde con quello finale della precedente. È perciò impossibile eseguire la scansione di una frase solo utilizzando le tecniche acustiche; la garanzia della esatta comprensione del parlato fluente non può fare a meno di evitare di coinvolgere tipologie diverse di ricerca, come analisi della struttura semantica, grammaticale, sintattica e lessicale.

Verso gli anni '70, negli U.S.A., un gruppo di industrie operanti nel settore dell'informatica e di strutture pubbliche, generalmente Università, parteciparono alla realizzazione di un progetto di ricerca, in parte autofinanziato, destinato a definire le tecnolgoie necessarie per un corretto riconoscimento della lingua parlata. Sebbene i risultati non siano stati del tutto coronati da successo, il progetto consentì, comunque, di porre solide basi nel campo del riconoscimento automatico della lingua parlata, oltre a raggiungere importanti risultati più generali, nel campo dei sistemi esperti e dell'intelligenza artificiale.

Gli studi eseguiti evidenziarono proprio quanto abbiamo appena detto, vale a dire, cioè, che il riconoscimento della lingua parlata non può fare a meno, oltre che di un regolare processo di acquisizione fonetica, di un'analisi grammaticale anche piuttosto spinta. Vediamo di cosa si tratta.

Un linguaggio, informatico e non, è composto di una serie di operatori che sono legati tra di loro e destinati a svolgere un certo compito in maniera razionale ed intellegibile da altri. I mezzi (la grammatica) dei linguaggi è rappresentata da unità elementari: ad esempio, le lingue umane sono formate dalle lettere dell'alfabeto; il Basic, da una serie di keyword come LET, READ, PLOT ecc. Una qualsiasi successione di tali unità rappresenta una frase. Ma non tutte le frasi hanno significato: PRLND, probabilmente, non vuol dire nulla in qualunque dialetto o lingua terrestre, tranne che per i costruttori di cambi automatici, come non ha senso la successione LZET READ 0 + + in Basic (mentre in C la seconda ha senso compiuto). Un modo banale per codificare un linguaggio sarebbe quello di elencare tutte le possibili combinazioni significative delle unità appartenenti al linguaggio stesso. Ma la cosa avrebbe ben poco senso, e probabilmente sarebbe impossibile in un linguaggio del tutto estensibile, come, ad esempio quello d'aritmetica dove gli operatori possono ripetersi in maniera del tutto infinita. Occorre quindi ricorrere ad altre tecniche, che consentano di specificare la tipologia di costruzione delle fasi.

Una di queste tecniche è costituita dagli automi a stati finiti. È questo un argomento piuttosto vasto e degno di attenzione, e probabilmente caro all'amico Giustozzi, cui non toglieremo il piacere di poterlo trattare in maniera estesa ed approfondita. In termini piuttosto banali diremo solo che, secondo tale inquadramento metodologico, una grammatica può appartenere a quattro livelli fondamentali (detti livelli di libertà). Il tipo 0 corrisponde a quello a più alto grado di libertà, vale a dire, in maniera piuttosto approssimata, che sono ammesse in questo caso, le più ampie ed elastiche strutture formali. Al tipo 3 corrispondono invece le serie e le regole grammaticali formali più precise e rigide. Gli automi a stati finiti obbediscono a grammatiche del tipo 3, ed è solo su questi che si è riusciti ad operare efficacemente per attaccare la rocca della comprensione della lingua parlata. Gli automi a stati finiti consentono di organizzare, in base ad una grammatica regolare, un numero elevato di frasi della lingua parlata pur senza giungere alla codifica, elencazione e conservazione di una lunga e non sempre manipolabile libreria di frasi fatte e comprensibili.

Non ci dilungheremo su tale argomento neppure riassumibile in maniera banale in queste pagine. Vogliamo, prima di chiudere con l'argomento, ricordare una faccia del problema solo intravista finora. Anche semplificando il problema al riconoscimento delle parole singole, rimane il fatto che il sistema di riconoscimento è sempre legato per forza di cose ad una fase d'istruzione iniziale, più o meno lunga e complessa, che è legata alle dimensioni dei vocabolari ed al numero degli utenti (ed alla loro voce, da riconoscere). È ovvio che, in prospettiva, in una utilizzazione di tipo pubblico, la cosa è improponibile. Ci viene ancora una volta in aiuto la statistica; ricordate la rappresentazione multidimensionale della parola, in termini di reticolo? Bene! riutilizzando ancora una volta la tecnica del confronto dinamico tra questi spazi multidimensionali (universi?) si è notato come, anche su una popolazione di utenti abbastanza varia, esistono raggruppamenti di punti (tanto per intenderci fonemi o finestre del tutto simili, anche se pronunciate da utenti diversi), che semplificano enormemente la tecnica del parsing. Esperienze a campione, eseguite su un centinaio di utenti, hanno evidenziato come generalmente si verifichino una ventina di raggruppamenti per ciascuna parola. Questo ha consentito una notevole semplificazione del problema, che comunque permane, vista la mole di tipologie di dati su cui è costretta ad operare.

C'è però da dire che i risultati ottenibili con tale tecnica sono comunque soggetti ad un più grosso margine d'errore. Inoltre sono tutti proponibili per dizionari dell'ordine di un centinaio di parole. Con vocabolari più ampi, i tempi d'analisi divengono estremamente lunghi e non più efficaci.

Altre tecniche sono in via di sviluppo. In questa ottica risulta interessante e gratificante sapere che un gruppo italiano di ricerca, il CSELT (Centro Studi E Laboratori Telecomunicazioni, di cui fa parte l'ing. Pieraccini, dai cui studi sono state attinte le notizie espresse in questo articolo) sta portando avanti una tecnica di riconoscimento della parola a mezzo di unità elementari fonetiche, i difoni: anche in questo caso il problema è rappresentato dal parsing della parola stessa, alla ricerca dei difoni caratteristici. Gli studi, iniziati nel 1981 hanno raggiunto stadi che lasciano prevedere che sarà possibile una procedura automatica per la ricerca dei difoni caratteristici, oggi eseguita manualmente, in modo da realizzare sistemi in grado di riconoscere frasi appartenenti a vocabolari di dimensioni di qualche migliaio di parole.

L'analisi della lingua parlata, semplificata in queste pagine, si ferma qui. La prossima volta porremo l'attenzione su un nuovo aspetto del problema di intercomunicazione tra uomo e macchina.

A risentirci.

MC

COMUNICAZIONI??

SOLO L'IMBARAZZO DELLA SCELTA!

DISK-DRIVE
STAMPANTI
COMPUTER
MONITOR
LA CASA DEL COMPUTER
IMPORTAZIONE DIRETTA

VIA MISERICORDIA, 84
56025 PONTEDERA (PI)

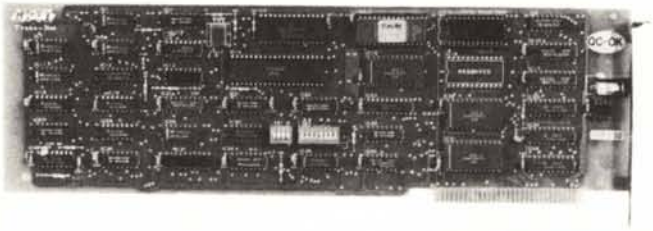
MODEM 101C (CCITT)

Interfacciabile con qualsiasi tipo di computer mediante RS-232. Velocità 300 B.P.S. full duplex. Auto Answer. Led indicatori di CX/RX/TX e Power on. Accessori optional alimentatore e cavo.

ACCOPPIATORE ACUSTICO (CCITT)

Di uso universale con RS-232. Conchiglie in gomma aggiustabili, facile da usare, 300 B.P.S., full duplex. Accessori: alimentatore.

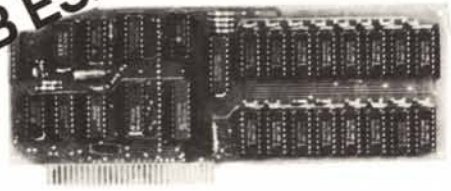
RETI LOCALI



TRANS-NET

Velocità di trasmissione: 1 Megabits/sec. Toplogia: Bus Distribuito. Distanza: 120 mt. massimo. Users gestibili: 255 massimo. Inseribile su: PC/XT/AT e compatibili. Sistema operativo: PC-DOS 2,0 - 2,11 - 3,0. Accessori Optional: cavo + terminator, repeater.

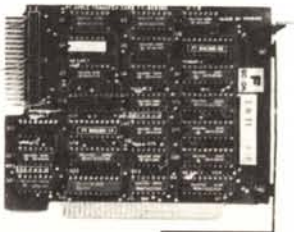
APPLE NOVITÀ 1 MB ESPANSIONE



La scheda è composta di due parti acquistabili separatamente: 1) scheda main da 512 K dotata di chips di espansione, manuale e soft, 2) scheda di espansione 512 K ulteriori da applicare sulla scheda main.

I-NET

Velocita di trasmissione: 1,2 Megabits/sec. Richiede l'installazione di un hard disk (Server) e si può configurare fino a 64 Users con 16 unità stampanti. Accessori optional: repeater.



APPLE-IBM CONNECTION CARD

Da applicare sul PC/XT trasforma un drive del PC in un drive per apple e, grazie al soft di gestione, si può formattare il dischetto in dos 3,3, trasferire files da apple a IBM e viceversa.

0587
212.312

CONTATTATECI OGGI STESSO PER MAGGIORI DETTAGLI E QUOTAZIONI

SIG.ri RIVENDITORI

PREZZI IVA ESCLUSA

DISK-DRIVE · STAMPANTI · COMPUTER · MONITOR

LA CASA DEL COMPUTER

IMPORTAZIONE DIRETTA

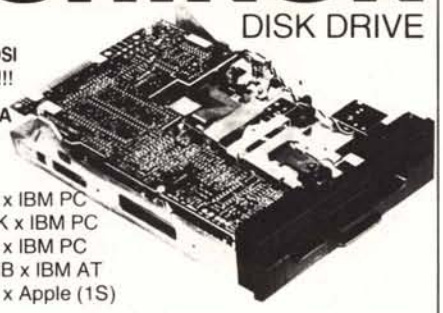
0587 212.312

VIA MISERICORDIA, 84 - 56025 PONTEDERA (PI)

CHINON DISK DRIVE

10 VOLTE PIÙ SILENZIOSI DEGLI ALTRI!!!

GARANZIA 1 ANNO



TIPI:
F-502 360K x IBM PC
F-502L 360K x IBM PC
F-561 1 MB x IBM PC
F-506 1,6 MB x IBM AT
F-051 180K x Apple (1S)

DISPONIBILI ORA I NUOVI MODELLI CON CHIAVETTA

PREZZI: DA LIT. 270.000

CHI VI DA UN ASSORTIMENTO COSÌ COMPLETO CON PREZZI SUPER COMPETITIVI???

Basta una telefonata ed in 48 ore riceverete quanto ordinato con garanzia 6 mesi od 1 anno e, se non sarete soddisfatti, vi sostituiremo l'articolo con lo stesso modello o con altro materiale a patto he il reso ci pervenga non manomesso, in porto franco, con gli imballi originali entro 18 gg. dalla data di spedizione

AT COMPATIBILE

Versione Base: Main Board ØK espandibile d 1 M.B., alimentatore 200 W. Cabinet in metallo, tastiera **L. 2.600.000**

PC/XT TURBO

L. 1.475.000
Clock 8-4,77 MHz
Main Board Esp. 640K

NOVITÀ

N. 1 Drive DS/DD 360K, controller.
Main Board ØK espandibile A 640K,
Alimentatore 130 W, Tastiera K5 S

PC/XT STANDARD (4,77 MHz)

L. 1.299.000
Configurazione come sopra ma con Main Board 256K espandibile a 640 K.

* * * Per le interfacce video vedere listino

Monitor Caègi Philips Monocr. x IBM	**L. 227.000**
Monitor Ciregi sonoro	**L. 148.000**
Monitor Philips HR Colori x IBM	**L. 690.000**

INTERFACCE PER APPLE

Controller Drive App.	60.000
16K Ram Card	83.000
80 Colonne Soft/SW.	108.000
8088 Card	592.000
Eprom Writer (16-64)	98.000
Prom Writer	434.000
Z/80 Card	59.000
RS-232 con cavo	100.000
Epson Printer e cavo	88.000
Grappler Pr. e cavo	98.000
AD-DA 12B./16 Canali	504.000
AD Card	177.000
AD-Da 8 Bit/19 Canali	336.000
IEEE-488 con cavo.	238.000
6809 Card	322.000
Communication Card	110.000
Super Serial Card	129.000
Pal Color Card	83.000
RGB Card (8 color).	124.000
RGB II (16 color)	194.000
Stereo Music Card	138.000
Scheda parlante	78.000
Wild Card	78.000
Scheda orologio	87.000
6522 Card	93.000
Forth Card	131.000
I.C. Test Card	198.000
80 Colonne + 64K IIE	54.000
80 Colonne x IIE	26.000
Adattatore Drive IIC	20.000
Adatt. Joystik IIC	14.000
Sch. orologio Prodos	120.000
Apple-IBM Conn. Card	590.000
512K Ram (ok) Esp. 1M	532.000
Esp. ulteriori 512K	240.000
Kit 8 Ram 4164 (64K)	34.800
Kit 8 Ram 256 (256K)	102.000

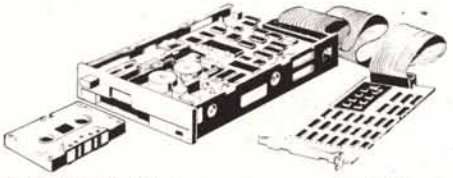
INTERFACCE PC/XT IBM

H.D. Controller 6210	330.000
Controller + cavo	120.000
Printer Card IBM	72.000
Color Graph. 2/L IBM	190.000
Mono/Col/Gr/Prin CR	340.000
Mono/Cr/Print Herc. 2	240.000
Multif. 256K Oran IBM	220.000
Multif. 384K Oram IBM	270.000
AD-DA Card IBM	435.000
Kit Ram 64K (9 Chip)	39.150
RS-232 Card IBM	108.000
Game I/O Card IBM	72.000
I/O Plus Card IBM	200.000
Eprom Writer 16/128	345.000
8255 Card IBM	270.000
IEEE-488 Card IBM.	570.000
Espansione 384K Ok	148.000
Espansione 512K (Ok)	138.000
Rete loc. I-Net + cavo	980.000
Rete loc. RPTI TR/Net	1.320.000
8087 Coprocessore PC	390.000
Mon/Col/Gr/Pr Amdek	490.000
Mono/Col/Gr Alta Ris	400.000
E.G.A. Color/Gr H.R.	980.000

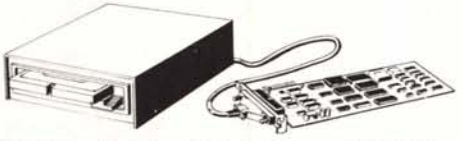
INTERFACCE AT IBM

AT Controller X 2FDD	278.000
AT Parall/Serial C.	224.000
AT Multi Serial (4S).	392.000
At Espans. 2,5 MB Ok	376.000
AT Espans. 3,5 MB Ok	520.000
AT Multifunc. 2,5 MB	490.000
AT Multifunc. 3,5 MB	590.000
Kit Ram 256K	114.750
Controller HDD + 2FDD	1.024.000

STREAMER 20 M.B.



TEAC MT-25T - Sofisticato sistema corredato di interfaccia e soft di gestione. Da collocarsi internamente al PC/XT/AT. La copia di 23 MB viene eseguita in 9 minuti circa su cassette tipo «COMPACT» da 500/600 FT.



SUPER 5 - Versatile unità di back-up per PC/XT/AT corredato di interfaccia e soft di gestione. Di semplice e veloce uso in quanto provvede ad eseguire la copia di 20 MB in soli 5 minuti. Usa cassette da 600 FT tipo «COMPACT». È dotato di cabinet metallico e cavo di collegamento all'interfaccia. Consigliato per installazioni esterne al sistema.

HARD DISK

Delle migliori marche come i nuovissimi Epson con ricovero automatico delle testine nella «Shipping zone» al momento dello spegnimento del sistema.

Epson HD-830 10 MB senza/contr.	**L. 1.090.000**
Seagate ST-225 20 MB senza/contr.	**L. 1.190.000**
Seagate ST-4051 40 MB senza/contr.	**L. 2.430.000**

DATAFLEX PROFESSIONAL 5¼

100% CERTIFICATI ERROR FREE

CONFEZIONE BULK 250/PZ. **L. 1.760** DOPPIA/DOPPIA

CON BOX IN PLASTICA OMAGGIO!!! SCONTI PER QUANTITÀ

SINGOLA F. - DOPPIA D.		DOPPIA F. - DOPPIA D.	
200 Pezzi	**L. 1990**	200 Pezzi	**L. 2650**
100 Pezzi	**L. 2100**	100 Pezzi	**L. 2800**
30 Pezzi	**L. 2350**	30 Pezzi	**L. 3150**
ALTA DENSITÀ PER AT. **L. 7.800**			

- I dischetti dataflex sono prodotti da uno dei più grossi fabbricanti americani che garantisce l'altissima qualità ed affidabilità.
- Uno speciale ed esclusivo strato «Multicot» protegge la superficie dall'usura del contatto con le testine garantendo minimo ben
- 10.000.000 di passaggi!!!
  La sicurezza dei Vs. dati è assicurata dall'ineccepibile supporto magnetico di primissima qualità.

DATO L'INSTABILE MERCATO DEI CAMBI PREGASI TELEFONARE PER CONFERMA PREZZI E DISPONIBILITÀ — RICHIEDETECI IL CATALOGO — SCONTI AI SIG. RIVENDITORI

Nella giungla dei compatibili oggi c'è un nuovo re:

Quasar

UN RE IN PRIMO PIANO

Il prezzo di un compatibile è importante. Ma noi vi garantiamo un prodotto il cui rapporto tra prezzo, qualità ed affidabilità è il migliore in assoluto.

UN RE CHE NON TEME CONFRONTI

È MS DOS compatibile (sa utilizzare tutti i maggiori programmi esistenti sul mercato dei personal computers). Possiamo paragonarlo al PC XT, ma con qualche caratteristica in più. Raggiunge la massima espansione di memoria - 640 Kbytes - direttamente su piastra madre e può passare dal clock standard di 4 MHz a quello, molto più redditizio, di 7 MHz, tramite un semplice comando da tastiera.

UN RE UNA DINASTIA

Ecco i diversi allestimenti che differenziano le macchine:

- con hard disk da 10 o da 20 Mbytes
- con la scheda per la rete locale
- con il modem completamente automatico o con il modemphone, che comprende anche l'apparecchio telefonico
- con la scheda color/graphic o la hercules o la monochrome
- con il mouse
- con l'A/D D/A converter
- con la scheda 8255 per 48 linee programmabili di I/O oppure senza drivers per applicazioni diverse
- con monitor monocromatico da 12 o 15 pollici o con quelli a colori da 14 pollici in media e alta risoluzione
- con stampante da 80 col/100 cps sino al top della gamma, la stampante laser.

UN RE PARTICOLARMENTE DOTATO

La dotazione standard è realizzata per soddisfare anche i più esigenti: due floppy disk drivers (256 Kbytes di memoria), scheda color/graphic, porta parallela e cavo per stampante. Ma l'ulteriore crescita del vostro reale amico dipende soltanto da voi.

UN GRANDE RE UN BEL RISPARMIO

Sua maestà ha un'altra grande dote, l'economicità. Interpellateci perché da noi i fatti non sono parole.

Quasar

QUASAR SRL - Via Diagonale 319 - 13050 Pratrivero (VC)
Tel. (015) 778804 - Tlx. 211401 MILFIL I