

L'Intelligenza Artificiale

di Raffaello De Masi

Riconosciamo la lingua parlata

«Buongiorno. Sono un calcolatore HAL 9.000: venni attivato il 12 gennaio 1982 nei laboratori di Verbania, Illinois». Poi il discorso si fa più sconnesso. «La mia mente se ne va, lo sento! Il mio istruttore mi insegnò una filastrocca, volete ascoltarla? Oh che bel caaaasteeeeelooooo» e qui il discorso si trasforma in un mugugno senza senso.

È ancora 2001, e l'astronauta David Bowman sta disinserendo le piastre di memoria del calcolatore di bordo della Discovery, in una delle sequenze più drammatiche del film. HAL impazzito, viene disattivato, ma lo spettatore non può fare a meno di pensare che, forse, il calcolatore sta effettivamente morendo.

In effetti, per l'immaginazione, è probabilmente molto più suggestivo pensare ad un computer che parla che non ad un antropomorfo. Merito sicuramente del fatto che ormai robot manipolanti, con arti meccanici, manovelle, bottoni ed utensili, siamo, alla fin fine, abituati a vederli, ma la voce, specie se ben articolata e suggestiva come quella di HAL (nella versione italiana del film — in quella americana il «doppiatore» possiede una voce gutturale ed un po' metallica, probabilmente più consona, ma sicuramente meno impressionante), desta pur sempre un brivido sotto pelle.

All'atto pratico, e lo dimostrano i fatti, il problema della voce è quanto di più complesso esista per un computer. Ben si intenda, non si tratta di dotare un computer di voce (il problema è stato già da tempo risolto, anche su computer di basso prezzo come, ad esempio, Commodore e Spectrum); il vero problema è di rendere interattivi umani e macchine tramite la parola: in parole povere, come è possibile farli comunicare servendosi della voce?

In effetti, a ben pensarci, il problema è doppio: da una parte l'acquisizione dei dati del discorso da parte della macchina («l'ascolto»), dall'altra l'organizzazione di una risposta logica e la sua «traduzione» in una serie di suoni aventi senso per un orecchio umano.

Il linguaggio parlato è, con molta probabilità, il sistema di scambio di informazioni più potente, efficace, e rapido esistente al mondo. Si tratta, a ben guardare, di un'area d'azione coinvolgente numerosi campi (acustica, logica, grammatica, sintassi, ecc.), di una complessità tanto elevata, che probabilmente sarebbe del tutto illusorio poter schematizzare in un diagramma di flusso l'ordine e lo sviluppo anche del più semplice messaggio verbale (si pensi alla complessità di una lezione universitaria, o magari di un racconto della trama di un libro o di un film). Fortunatamente, la potenza e la versatilità del cervello umano è tale che ci accorgiamo ben poco del lavoro che costa anche la più semplice, ancorché accanita, discussione su una partita di calcio (immaginate le locazioni di memoria occupate e il lavoro dei puntatori, nel nostro cervello), o la più banale discussione con la moglie alla notizia dell'arrivo della suocera. Pertanto le chiacchiere di Bowman ed HAL, logiche, espressive, estremamente articolate, sono pura fantascienza. Ciononostante lo sforzo continuo dedicato da gruppi di ricercatori nel tentativo di utilizzare tale mezzo di comunicazione (che, non dimentichiamolo, è costato all'uomo decine di millenni d'evoluzione) ha portato al raggiungimento di certi traguardi, abbastanza soddisfacenti, e tali da consentirci di estrarre tale argomento dal limbo della fantasia per inserirlo nella cartella del «Qualcosa è già stato fatto in proposito».

Il motivo di tali sforzi è ovvio; non esiste altro mezzo di comunicazione altrettanto potente, efficace e versatile, l'abbiamo già detto: lo stesso messaggio visivo non ha senso, tranne in sporadici casi, senza la voce (mentre non è vero il contrario). Tutto ciò ha portato, così, alla nascita di una nuova scienza, specializatissima, l'A.S.R. (dall'inglese Automatic Speech Recognition: riconoscimento automatico

della voce), che, ovviamente, si basa su numerose discipline, come intelligenza artificiale, fonetica, logica, ecc.

Quali siano i campi applicativi della A.S.R. è ovvio: basti pensare al campo educativo commerciale o scientifico per vedere orizzonti di utilizzazione sterminati. I primi risultati non si sono fatti aspettare: esistono macchine che ascoltano, interpretano e rispondono a messaggi, anche se questi ultimi vanno forniti attraverso regole sintattiche e grammaticali piuttosto rigide. I tentativi effettuati (si noti il notevole impegno profuso in tal campo da uno dei colossi dell'informatica, la Texas Instruments) sono stati, comunque orientati tutti nella stessa direzione: divisione del problema in tre campi: ascolto, elaborazione dei dati, risposta.

La parola umana è frutto delle alterazioni e delle modifiche fisico-meccaniche subite dal nostro apparato acustico, rappresentato, per essere precisi, da quella parte compresa tra i polmoni che rappresentano la fonte di energia nel loro insieme, e che non partecipano alla vera e propria fonazione, e gli organi fonetici propriamente detti, costituiti dalle corde vocali vere, e dall'apparato boccale e nasale (lingua, palato, labbra, denti, ecc.). La produzione dei suoni è affidata ad una serie di meccanismi coinvolgenti uno o più di tali componenti; è possibile, così, riconoscere suoni prodotti dalle corde vocali (i cosiddetti fonemi vocalizzati), prodotti, appunto da questi componenti (ad esempio la [m], la [n], e tutte le vocali), sovente inquinati da partecipazione di altri organi, come la gola (gutturali, come la c e la g aspra, ad esempio delle parole chiodo e gamma); ma spesso il suono appare prodotto da altri organi, diversi dalle corde vocali; è il caso delle sibilanti (come la [s], la [f] e la [c] dolce), in cui la produzione del suono è affidata ad una particolare deformazione imposta al cavo orale (provate a dire salicce

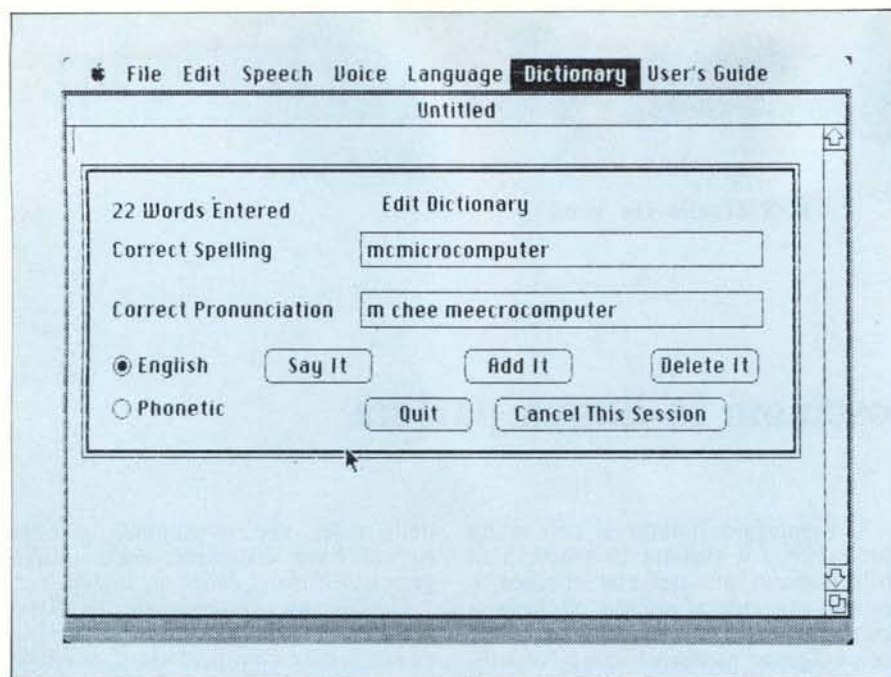


Figura 1 - Uno dei più raffinati sintetizzatori vocali su computer, lo Smooth Talker della First Byte, incorpora un dizionario, espandibile, in cui occorre inserire la parola, e la sua corrispondente pronuncia.

con un dito tra i denti). Altre categorie di suoni invocano, invece, la compressione della colonna d'aria presente nel cavo orale (ad esempio la [p], la [t] e la [d]).

Qualunque sia il meccanismo determinante il suono è possibile individuare, nella lingua parlata, una trentina di fonemi (il numero può variare da lingua a lingua): tramite essi è possibile redigere una qualsiasi frase di senso compiuto. Questo non vuol dire che il risultato sarà una frase detta da un essere umano: il fenomeno del discorrere (non dell'emettere suoni comprensibili) è qualcosa di straordinariamente complesso, solo attualmente agli albori di uno studio sistematico; tanto per fare un esempio, anche la più semplice frase, univocamente interpretabile se scritta, allorché viene pronunciata si arricchisce di un enorme corredo di sfumature, timbri, tonalità, che possono cambiarne totalmente il significato ed il valore.

Il fonema, in effetti rappresenta un modello molto rudimentale della produzione di suoni vocalizzati. Nella lingua parlata non esistono divisioni tra suono e suono, ed è anche a questo che dobbiamo la notevole ricchezza di sfumature che caratterizza anche il più semplice discorso umano, e che ci permette di verificare ed intendere sensi nascosti, stati d'animo e retorica insiti nel discorso del nostro interlocutore. Il concetto di fonema, del tutto teorico, ci consente di affrontare su base analitica il problema, ma all'atto pratico non ha alcun senso; il parlato è formato da una serie di suoni formanti le

parole e queste, ancora, formanti una frase. In un discorso, tranne casi eccezionali e per particolari esigenze sintattiche e di enfasi, una frase viene espressa come una serie di suoni del tutto legati l'uno all'altro, e solo l'eccezionale capacità di analisi del cervello umano permette di distinguere, del tutto inconsciamente, le parole in esso contenute; ma non è tutto: si noti come esistono costrutti del tutto diversi a seconda della particolare forma costruttiva del fraseggio stesso; tanto per intenderci una vocale in fine di parola (la maggior parte dei casi nella lingua italiana) si dimezza, come durata di pronuncia se la parola successiva inizia ancora con una vocale. Ancora, la stessa sillaba viene pronunciata in modo diverso se depositaria o meno di accento tonico; e, di più, si verifica un aumento della durata del suono in corrispondenza del soggetto della frase stessa.

Ovviamente, a ciò si aggiungono altre variabili, legate soprattutto a differenze di pronuncia, come le forme dialettali e i difetti di pronuncia e, ancora, la possibilità di differenza di parole e frasi dette dalla stessa persona in tempi ed occasioni diverse. Tutto ciò ci porta a concludere che la parola è qualche cosa di estremamente difficile da analizzare. Ma bisogna pur farlo: vediamo qualche tecnica.

Attualmente, in laboratorio è già stato possibile, anche su sistemi abbastanza avanzati, realizzare una macchina capace di obbedire ad ordini dati a voce (e, eventualmente, rispondere ancora a voce). L'analisi della voce,

per la notevole complessità del modello da realizzare, risulta ovviamente di complessità superiore a quanto anche il più perfezionato dei sistemi possa oggi affrontare. Ci vengono però in aiuto alcune tecniche che, sebbene sfrondate delle loro particolarità più complesse, consentono una analisi abbastanza fedele del fenomeno.

È ovvio che il riconoscimento della voce, per poter essere ristretto in un campo analizzabile da un computer, deve essere ridotto all'analisi di un modello statistico: ciò può essere eseguito utilizzando il teorema di Nyquist: questo teorema ammette che, nell'analisi di un campione continuo, come la voce, è possibile eseguire un campionamento regolare di esso in modo abbastanza ristretto in modo che il risultato possa univocamente rappresentare, ed essere interpretato, come il segnale di partenza. L'intervallo tra i campioni prelevati non può, però essere inferiore (nel caso di un segnale sonoro come la voce) al semiperiodo della più alta frequenza presente nel campione stesso (ciò è ovvio, in quanto una analisi ad intervalli più lunghi porterebbe alla perdita di segnali significativi). Esperimenti in tal senso hanno però dato scarsi risultati, visto che pur non tenendo conto delle frequenze più alte della voce umana, frequenze necessarie da analizzare, come quelle intorno ai 5.000 Hz, darebbero modelli ancora troppo complessi anche per macchine dedicate.

L'altra soluzione, classica, del problema è rappresentata da un aggiramento dell'ostacolo dell'analisi di tutto lo spettro. Esistono studi pregevoli in tal senso, uno per tutti quello di Roberto Pieraccini, pubblicato in diverse tranche su un notissimo periodico scientifico. In tali memorie l'autore, impegnato nei laboratori del CSELT (Centro Studi e Laboratori Telecomunicazioni) di Torino dimostra come l'orecchio umano sia sintonizzato su un certo numero di bande di frequenza, dette bande critiche, sufficienti alla comprensione, anche abbastanza raffinata, della parola stessa. La tecnica mostrata dall'autore è piuttosto semplice ed intuitiva, almeno a livello di analisi iniziale; il segnale vocale (ed il suo spettro) vengono suddivisi (si lavora su sonogrammi tridimensionali, in cui l'asse orizzontale rappresenta il tempo quello verticale la frequenza, e quello perpendicolare al foglio, evidenziato da un maggiore o minore annerimento delle linee del sonogramma stesso, l'energia per una data frequenza in un dato istante) in intervalli consecutivi, diciamo di alcuni millisecondi; questi intervalli, regolari, detti «finestre» possono essere quantizzati calcolandone lo spettro totale di energia. Ogni finestra, quindi, può essere indi-

viduata univocamente in base ai contributi specifici di ogni banda critica.

Il Pieraccini nelle stesse note, fa un esempio piuttosto chiaro: nell'intervallo tra 300 e 3.400 Hz, che rappresenta lo spettro normalmente usato nelle comunicazioni telefoniche, e che garantisce una discreta qualità del messaggio sonoro (anche in termini di intonazione, sensibilità, rispetto dell'inflessione dell'interlocutore), sono individuabili 13 bande. L'analisi di tali bande porta ad una quantizzazione numerica delle stesse, e, alla fine, alla rappresentazione delle stesse tramite un numero (è a ciò che si voleva arrivare). Di qui la via è facile (si fa per dire). Ogni finestra viene rappresentata da una array di 13 numeri. L'insieme di tali array rappresenterà il discorso.

L'analisi di tanti dati, e la conservazione di tanti numeri, non possono essere svolte da un calcolatore tradizionale: a tale compito vanno dedicate macchine particolari, chiamate array processor, che svolgono, per conto di un più tradizionale sistema, il compito della manipolazione di tanti e tali valori.

L'acquisizione di dati è forse la parte meno complessa dell'operazione; il tutto viene eseguito trasformando, attraverso sistemi A/D-D/A i segnali elettrici ottenuti da un tradizionale microfono in valori numerici destinati poi all'analisi ed all'immagazzinamento.

Gli esperimenti eseguiti al CSELT in tal senso hanno fornito buoni risultati iniziali: ma lo stesso autore individua subito un problema, forse non evidente immediatamente. Gli esperimenti, eseguiti su una macchina Digital VAX 11/780 diedero buoni risultati a patto che, ovviamente, le parole fossero perfettamente scandite: ma non basta. Il problema è che la lingua parlata, anche la più primitiva e rudimentale non ammette pause tra parola e parola. Anzi, addirittura, molto spesso complica le cose il fenomeno del legame tra parole di cui la prima terminante e la seconda iniziante con vocale. In questo caso, addirittura, in una parlata abbastanza fluente si può avere la scomparsa di una vocale, generalmente quella finale della parola precedente. Occorre, quindi, pronunciare le parole facendole intervallare da pause sufficientemente lunghe. Si tratta, quindi, di un modo piuttosto innaturale di pronuncia, ma che può, come fa notare Pieraccini, ancora essere utilizzato efficacemente per impartire ordini semplici, come, ad esempio, l'avvio e l'interruzione della lavorazione di una macchina, inoltre, l'analisi di singole parole (e non di un significato di una intera frase) risulta ovviamente piuttosto facile, specie se le pa-

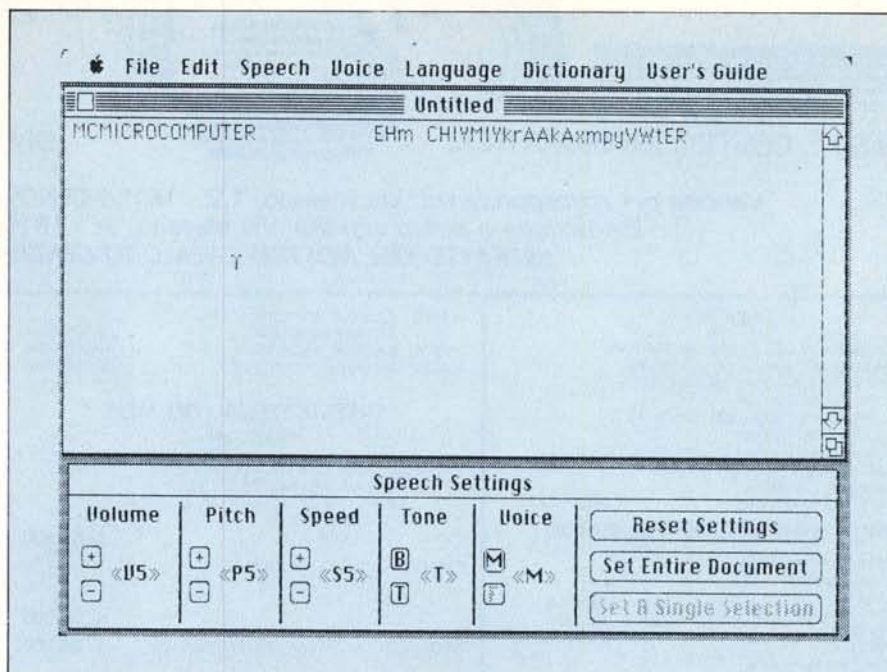


Figura 2 - Lo stesso pacchetto, nella scelta delle modalità dei parametri della voce: si noti la possibilità di variazione del tono e del timbro (maschile o femminile) della voce sintetizzata dal piccolo altoparlante del Mac; ancora, in alto a destra, si osservi la rappresentazione convenzionale dei fonemi della parola presente a sinistra.

role chiave, comprensibili dalla macchina ed a cui essa reagisce, possiedono spettri sonori molto diversi e non equivocabili tra loro.

Il grosso problema, insito nel riconoscimento di parole successive (e, in certa misura, anche nel riconoscimento di una parola singola) è, però quello dell'individuazione del punto iniziale e finale di una parola. Tanto per banalizzare il problema, cosa succederebbe se si lavorasse in un ambiente rumoroso, o se l'operatore avesse l'asma, o se tossisse, o se accanto a lui ci fossero due persone in discussione? Ipotesi indesiderabile, ma non peregrina, visto che è pensabile che sistemi, destinati a controllo di macchine, e quindi destinati a riconoscimento semplificato di ordini dati a voce, operano, prevedibilmente, in ambienti non proprio privi di rumore.

Come è possibile aggirare il problema? La soluzione ci viene ancora data da una tecnica a campionamento, anzi, per essere più precisi, da una analisi statistica di confronto. La macchina viene dotata di un vocabolario di base (in forma di array, di cui parlavamo prima). La macchina, anche in presenza di rumore (è impensabile poter lavorare in un ambiente ideale, senza suono), eseguirà una analisi continua delle finestre di parsing e verificherà, volta per volta, la corrispondenza dei suoni a lei pervenienti con il suo vocabolario di base. È ovvio che la corrispondenza non potrà, in ogni caso, essere perfetta, ma la definizione di un

marginale di errore (che, lo si noti, è rappresentato da una vera e propria differenza numerica) di tipo probabilistico, e che può essere definito, anche sperimentalmente, in base al livello di interferenze (rumore, rumori accidentali) presenti nell'ambiente di lavoro, porterebbe ad un soddisfacente livello di comprensione tra uomo e macchina.

Il tutto può essere eseguito «allineando» le due rappresentazioni, vale a dire quella della parola campione e quella della parola analizzata: è intuitivo che il parsing, l'analisi, ha maggiori probabilità di successo se il confronto viene effettuato finestra per finestra, affidando poi la decisione ad un processo di valutazione statistico (ovvio che la corrispondenza di 80% delle finestre rappresenta, con notevoli probabilità, l'identità dei due segnali), piuttosto che ad una valutazione globale del totale «peso» delle finestre stesse.

La valutazione delle differenze dovute a diversità di timbro, ad esempio, o di frequenza (l'ordine potrebbe essere inviato da un uomo, magari affetto da raffreddore, e da una donna dotata di ugola da mezzosoprano) sono ben misera cosa (e peraltro ancora agevolmente risolvibili) nei confronti di ben più complessi problemi di riconoscimento, legati a diverse altre variabili.

Ma di ciò parleremo la prossima volta.



CEIN S.r.l. CENTRO INFORMATICA

DIV. VENDITA PER CORRISPONDENZA

Vendita per corrispondenza: Via Merano, 1/2 - 16154 GENOVA - Tel. 010-673522

Esposizione e punto vendita: Via Merano, 3r - 16154 GENOVA

ESTRATTO DAL NOSTRO CATALOGO GENERALE

LINEA PC

PC2001 Computer MS DOS compatibile IBM 100%
Contenitore metallico con ventilazione
256 Kbyte RAM espandibili a 640 Kbyte
Scheda grafica colore
Scheda stampante parallela CENTRONICS
Controller per 4 drivers
2 floppy disk drivers incorporati da 360 Kbyte
Tastiera ascii standard 84 tasti Cherry
Garanzia 1 anno.
Manuale e schemi elettrici 2.345.000 + IVA

OFFERTA HOBBY & STUDIO (Cod. OHS001)

- 1 Personal computer PC2001
1 Monitor monocromatico 12"
ROLAND MA121
1 Stampante Panasonic KX P 1091
1 Cavo monitor
1 Cavo Stampante
1 Risma carta per stampante
1 Confezione dischi 5" 1/4

OFFERTA LIMITATA NEL TEMPO

Prezzo speciale L. 3.299.000 + IVA

OMAGGIO 1 Package software con splendidi games ed utilities

SCHEDE PER PC

Table with 2 columns: Product Name and Price. Includes items like Main Board 256 KB, RAM card 384 KB, Multi function card, etc.

ESPANSIONI DI MEMORIA DI MASSA

Table with 2 columns: Product Name and Price. Includes items like Hard disk 10 Mbyte, Hard disk 20 Mbyte, etc.

ACCESSORI PER PC

Table with 2 columns: Product Name and Price. Includes Coprastiera, Bus converter per M24.

Table with 3 columns: Product Code, Description, Price. Includes PCA005, PCA006, PCA002.

OFFERTE SPECIALI DEL MESE AMSTRAD

Table with 3 columns: Product Code, Description, Price. Includes PS5040, PS5041, PS5042.

COMMODORE

Table with 3 columns: Product Code, Description, Price. Includes PS5030, PS5031, PS5032, GIOCHI SU CASSETTA.

C64

Table with 3 columns: Product Code, Description, Price. Includes PS5001, PS5002, PS5005, etc.

C16

Table with 3 columns: Product Code, Description, Price. Includes PS5013, PS5014, PS5015, etc.

Table with 3 columns: Product Code, Description, Price. Includes SUPER ZAXXON, SLAPSHOT, etc.

LE ISTRUZIONI SONO SUL VIDEO IN ITALIANO

MSX

Table with 3 columns: Product Code, Description, Price. Includes GIOCHI SU CASSETTA, PS5022, etc.

N.B. LE OFFERTE SONO LIMITATE ALLA DISPONIBILITÀ

Table with 3 columns: Product Code, Description, Price. Includes NOVITÀ SOFTWARE SPECTRUM 48K, SSA134, etc.

NOVITÀ ESCLUSIVA PER AMSTRAD CPC 464/664/6128

RISIKO

Table with 3 columns: Product Code, Description, Price. Includes ASA295, ASA296.

COMMODORE C64

CHB010 KIT SPEEDDOS: CON QUESTO SPLENDDO KIT POTRAI TRASFORMARE IL COLLEGAMENTO DEL DRIVE 1541 CON IL C64 DA SERIALE A PARALLELO, AUMENTANDONE 20 VOLTE LA VELOCITÀ.

CHB009 SPROTECT: FAVOLOSO CARTRIDGE CHE TI PERMETTERÀ LA COPIA DI PROGRAMMI PROTETTI DA DISCO A DISCO, DISCO A CASSETTA O CASSETTA A DISCO. FORNITO CON SOFTWARE E MANUALE IN ITALIANO.

CHB002 COPIANASTRI: DISPOSITIVO HARDWARE CHE PERMETTE DI EFFETTUARE COPIE DI PROGRAMMI SU NASTRO ANCHE PROTETTI COLLEGANDO SIMULTANEAMENTE DUE REGISTRATORI AL TUO COMPUTER.

TELEFONI

Table with 3 columns: Product Code, Description, Price. Includes MONOCORPO, TLA101, TLA102, etc.

TLA201 TELEFONO A FORMA DI BANCONOTA DA 200 DOLLARI SIMPATICISSIMO COMPLETAMENTE ELETTRONICO L. 59.000

MY FLOWER TELEFONO A FORMA DI FIORE L. 69.000 UN MERAVIGLIOSO REGALO PER SIGNORA

TLA301 BIANCO TLA302 OCRA TLA303 ROSSO

TELEFONO DA SCRIVANIA «VIVA VOCE» L. 139.000

SHERITONE AT9 TELEFONO SENZA FILI L. 249.000 PORTATA OLTRE 200 METRI TRASMISSIONE FM FULL DUPLEX AUTO RICARICA DELLE BATTERIE MEMORIZZAZIONE ULTIMO NUMERO CHIAMATO BLOCCO DI SICUREZZA

GARANZIA CEINPOST - La CEIN S.r.l. si impegna a sostituire o a rimborsare la merce venduta per corrispondenza, purché rispedita entro e non oltre OTTO giorni dalla data di consegna.

ORDINAZIONE - Per una migliore gestione dei Vostri ordini Vi preghiamo, per quanto possibile, di utilizzare gli allegati moduli d'ordine indicando chiaramente: Nome, Cognome, Indirizzo, Telefono e Codice, Quantità, Descrizione e Prezzo degli articoli da Voi desiderati.

Si accettano ordini telefonici.

PREZZI - I prezzi indicati si intendono, se non diversamente specificato, comprensivi di IVA. Non sono vincolanti per la CEIN S.r.l. e potranno subire variazioni per motivi indipendenti dalla nostra volontà.

PAGAMENTO - Esclusivamente contrassegno al ricevimento della merce. Per importi inferiori a L. 300.000 saranno addebitate L. 5.000 per importo spese di imballo e spedizione.

FATTURAZIONE - Le aziende che necessitano emissione di fattura dovranno richiederla al momento dell'ordine indicando il numero di partita IVA.

INOLTRE DISPONIBILITÀ SOFTWARE PER:

AMSTRAD - COMMODORE - SPECTRUM MSX - CP/M - MS DOS

CHIEDETE IL NOSTRO CATALOGO GENERALE INVIANDO L. 3.000 RIMBORSABILI AL PRIMO ACQUISTO

CENTRO PILOTA AMSTRAD