di Raffaello De Masi

Un esame comparato di tre prodotti per il riconoscimento vocale

Io parlo, tu scrivi

Quanta strada, in mezzo secolo! Agli inizi degli anni cinquanta frequentavo le elementari, e i miei ricordi di allora sono quelli della signora Laura Chiarelli, una maestra già canuta ed esile come una filigrana, le aule con il braciere, i banchi di legno con l'appoggiaschiena separato (banchi che avevano servito, probabilmente, anche mio padre, che avevano scritte d'antichi alunni passati da tempo immemorabile, e cigolavano come vecchi legni di porto). L'accesso ai libri (un libro di lettura e un sussidiario che, insieme, spesso non raggiungevano le trecento pagine) era preceduto dal cigolio delle bandelle delle levate del piano di scrittura, vero scrigno dove veniva custodita la borsa (altro che gli zainetti firmati di oggi). I banchi avevano al centro un calamaio di bachelite, incastrato nel piano, dove s'intingeva il pennino (i compagni più ricchi avevano il pennino Cavallotti, invenzione della genialità italiana che ormai dorme nella tomba dell'oblio), con un inchiostro che, oggi, in qualche residuo "amarcord" conservato gelosamente da mia madre è divenuto, negli anni, color seppia.

Il tutto era basato su un ritmo lento, misurato, regolare, e scarsamente soggetto alla fretta. Allora, scrivere una lettera significava impiegare anche qualche ora, visto che si cercava, sovente, attraverso il foglio, di comunicare non solo notizie, ma anche stati d'animo, sensazioni, e presenze. Comporre un tema (anzi, come si chiamava allora, svolgere una composizione), sotto il vigile occhio della signora Laura, era bagnare il pennino nel calamaio, lisciarlo sul bordo per eliminare l'inchiostro superfluo, riempire i righi con lettere e parole tutte della stessa altezza, ben allineate e coperte, 'sì che alla fine la pagina offrisse un colpo d'occhio preciso e puntuale, ordinato e gradevole. E non dava fastidio certo l'Ottimo finale, scritto dalla maestra con la sua grafia sottile e gelida, soprattutto rispetto alla nostra panciuta e "calda".

Le cose si complicavano un poco con il dettato. Qui occorreva tenere il ritmo di chi leggeva, e un errore poteva essere fatale, visto che poi ci costringeva a correre appresso alla dettatura fino almeno alla fine del paragrafo, quando c'era una sosta ristoratrice. Oggi, a diversi decenni da quelle mattine, mi sembra di parlare di un mondo alieno, quasi visitato in un'altra vita; lettere per-

IBM ViaVoice Millennium Pro versione 7

IBM Italia
http://www-4.ibm.com/software/speech/millennium/professional.html
disponibile nei migliori negozi di informatica

Prezzo: (IVA compresa) L. 349.000

Philips FreeSpeech 2000 versione con SpeechMike

Philips Speech Processing
P:O: Box 138
1120 Vienna – Austria

Distribuito in Italia da:
Italsel srl
Via Lugo 1 - 40128 Bologna - Italy
tel. +39-051-320409

Prezzi: (IVA compresa) £. 359.000
versione senza Speech Mike L. 200.000

Dragon Dictate Naturally Speaking 3.5 versione standard.

Dragon System Inc.
320 Nevada Str.
Newton , Massachussets 02460
USA

Distribuito in Italia da:
Italsel srl
Via Lugo 1 - 40128 Bologna - Italy
tel. +39-051-320409

Prezzo: (IVA compresa) L. 159.000

sonali, magari scritte a mano, non ne mandano più. Chi ne avrebbe il tempo? Un messaggio d'e-mail si prepara e si spedisce in un momento e, se si è fortunati, si riceve la risposta magari prima di disconnettersi.

E il dettato della signora Chiarelli? Forse si fa ancora, nelle scuole elementari, a meno che qualche pedagogo infallibile non abbia scoperto che è alienante e diseducativo! E la dattilografia, croce e delizia delle scuole di ragioneria, sulle vecchie Olivetti, 'ché , per essere bravi, occorreva imparare a battere con l'anulare e il mignolo, tanto da sviluppare un muscolo da farci il sollevamento pesi! Inutile, oggi si sfiorano i tasti di un computer, sensibili e confortevoli, e domani la tastiera diverrà obsoleta, visto che già oggi molti utenti più innovativi prendono il microfono e dettano; e il computer ascolta e scrive!

Il sogno del millennio

Intorno al 1989 le tecniche d'utilizzo del software ad attivazione vocale, successivamente definite come riconoscimento vocale (anche se, a rigore, tale definizione è restrittiva) uscirono dai laboratori di sperimentazione, per approdare nell'area dei personal computer. Fino ad allora tale tecnica era rimasta confinata all'area dell'intelligenza artificiale (ricordo , nella rubrica a lungo tenuta su MC, di averne diverse volte parlato), anche perché l'hardware disponibile era di potenza tanto modesta da risultare praticamente inutilizzabile. Ad esempio, la Kurzweill A.I., cui spetta il merito di aver introdotto sul mercato queste tecniche, aveva già in catalogo l'ambiente VoiceRAD, ma l'hardware disponibile allora, con punta nel 386, era assolutamente inadeguato a supportarlo; i risultati erano modesti, l'uso era sfibrante, frustrante e fastidioso, e ben raramente si ottenevano risultati appena simili a quelli che si potevano vedere nelle dimostrazioni, che si nascondevano sotto il trucco, spudoratamente commerciale e non onesto, di utilizzare un ambiente utente-macchina ben coordinato, che oltre tutto utilizzava forme, script, sequenze fortemente collaudate. I risultati, sul campo, erano scoraggianti , in termini di velocità e qualità, e questo lancio forse eccessivamente prematuro portò a un "rebound" del mercato, creando una più o meno vera fama d'inaffidabilità e d'inutilizzabilità finale.

Ma già qualche anno dopo l'avvento del Pentium e la disponibilità di memoria a più basso costo (vi racconto un'esperienza diretta: nel 1982 64K di RAM per

Cosa è una pausa discreta?. Parlare di essa significa accennare alle due grandi difficoltà insite nel riconoscimento vocale.

Partiamo dal principio, assiomatico, che il cervello umano è estremamente più versatile, se non più potente di qualunque macchina. Immaginiamo di ascoltare un'orchestra; sebbene noi siamo investiti contemporaneamente dal suono di tutti gli strumenti miscelati assieme, è sufficiente adottare un minimo di attenzione per distinguere e "ascoltare" solo i violini, i corni o individuare subito il tonfo della grancassa o dei piatti. Un altro esempio; immaginiamo di essere in un ambiente rumoroso, dove è accesa una radio, diverse persone parlano contemporaneamente, magari la finestra è aperta e si sente il rumore del traffico, il clacson delle auto e il fischio di un aereo che passa. Ciononostante, se un interlocutore ci parla riusciamo ad ascoltare senza problemi ciò che ci dice, magari senza neppure un grosso sforzo.

Il fatto è che il cervello umano riesce, in ambedue i casi, istintivamente a "filtrare" e scartare quello che non gli interessa, recuperando solo quello che desidera davvero sentire. Immaginiamo invece cosa avviene all'ingresso di un sistema di input vocale; se non siamo in un ambiente silenzioso arriverà, "all'orecchio" della macchina, un coacervo di rumori tra cui essa dovrà distinguere un "parlato". Compito strenuo, davvero!

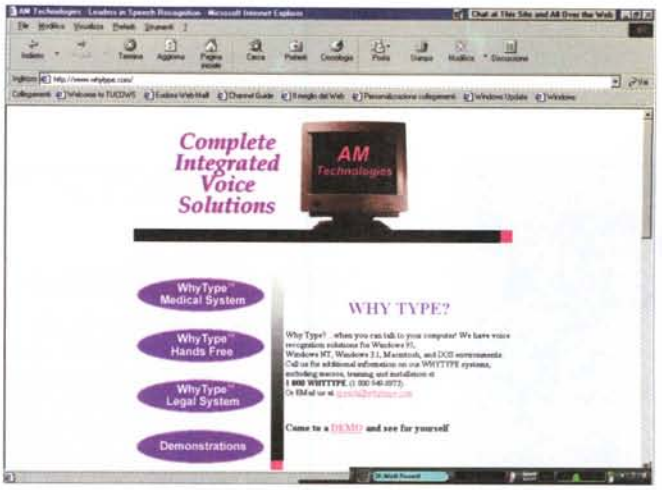
La seconda difficoltà è, forse, ancora più complessa da superare, visto che le funzionalità del nostro cervello sono qui ancora più spinte. E , paradossalmente, è intorno a questo problema che si sono concentrati gli sforzi e si sono raggiunti i migliori risultati. Immaginiamo una semplice frase: "Buongiorno, come sta, spero bene . E i suoi?" . Quando pronunciamo questa frase noi diciamo "Buongiornocomestasperobeneeisuoi". La frase è brevissima e già si fa una certa fatica a leggerla. Immaginiamo cosa succede a recitare una filastrocca, una preghiera o a leggere un discorso.

La potenza del PC che abbiamo in testa è tanta che, senza neppure accorgercene, il cervello scinde l'immensa frase nelle parole che la compongono. Ma per una macchina è dura riconoscere e tagliare in frammenti logici una frase del tipo "Sopralapancalacrapracampasottolapancalacapracrepa" Ihhh! Immaginate solo il lavoro , eseguito per forza a tentativi, di selezione dei fonemi e di riconoscimento nell'ambito della traslazione in una parola logicamente valida e così via. Insomma, è proprio dura! Per questo, all'inizio , si leggeva separando le parole con una pausa di almeno 1/5 di secondo, cosa innaturale e che mal si adattava con l'esigenza di attenzione necessaria per la dettatura "a braccio".

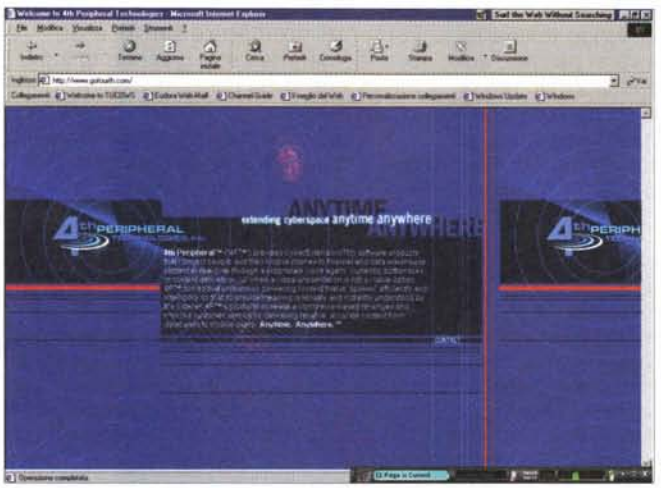
Il primo problema fu, in tempi pionieristici, risolto utilizzando ambienti silenziosi (anche oggi, per quanto possa essere divenuto sofisticato il nostro software, lavorare lontano da rumori permette di ricavare risultati migliori); oggi si utilizzano microfoni direzionali e a soppressione di rumore, oltre a implementare nel software routine di eliminazione dello "sporco". Per il secondo problema si può affermare che il tutto è profondamente legato alla potenza del microprocessore, alla quantità di RAM disponibile e alla bontà della scheda audio, nell'ordine. Gli algoritmi di scansione del parlato e di riconoscimento sintattico e logico dello stesso sono stati continuamente migliorati nel tempo, con risultati di tutto punto, visto che si raggiungono, in condizioni ideali, riconoscimenti del 95-98% con velocità fino a 150 parole al minuto (ben oltre quindi le capacità di un'esperta dattilografa).

E siamo solo agli inizi!

l'HP 95 costavano 250.000 lire. Nell'88, per un 8086, 256K costavano duecentomila, nel 94 un MB costava 70 biglietti da mille, qualche mese fa, prima degli eventi tellurici sudasiatici, mille x mega, più o meno) riproponeva la ricognizione vocale come alternativa interessante, almeno per il word processing, alla tastiera. I prezzi erano ancora alti (un sistema a hoc, compreso l'hardware, era passato , in un anno, dai 30.000 $ a prezzi inferiori della metà) ma già confrontabili e sovente inferiori allo stipendio annuale di



Alcuni dei numerosi siti che offrono prodotti e tecniche di riconoscimento vocale. Molti prodotti sono orientati verso l'area medica e verso il recupero dei disabili.



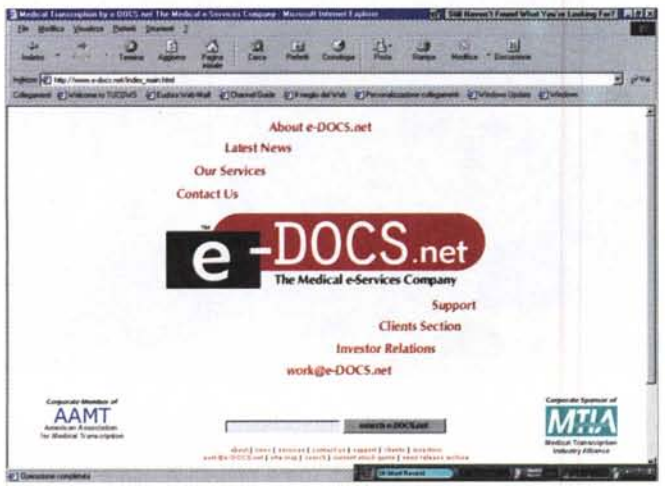
un dattilografo; inoltre la qualità delle schede audio era fortemente migliorata. Insomma il miglioramento delle tecniche di database, la velocità e la qualità degli algoritmi di riconoscimento, il forte decremento di costo dell'hardware concorrevano tutti a creare un nuovo favorevole ambiente più propizio all'uso e allo sfruttamento delle tecniche di ricognizione vocale. Il raggiungimento del traguardo dei 100 e oltre MHz nel 1995, e la standardizzazione di alte velocità contribuì notevolmente a incrementare qualità e velocità delle tecniche di riconoscimento vocale. E' di quella data la drastica riduzione delle pause discrete , che passavano da 1/5 a 1/10 di secondo e anche meno, in determinate occasioni. Il passo immediatamente successivo fu quello dei programmi che offrivano il riconoscimento del parlato continuo per numeri e piccole frasi generiche.

Fu a questo punto che i pacchetti di riconoscimento vocale si divisero in due direttrici principali: quelli "speaker dipendenti" e quelli "speaker indipendenti". La differenza essenziale tra i due è che questi possono essere usati direttamente dopo l'installazione, senza alcuna customizzazione. Gli altri sono, per così dire, "affezionabili" ai singoli utenti; dopo l'installazione, occorre organizzare una seduta di training per "insegnare" al programma a riconoscere il nostro parlato, il nostro timbro vocale, il nostro inevitabile accento dialettale. Insomma, occorre educare la macchina a riconoscere il nostro "profilo vocale". Questa via è, ovviamente più semplice e più disponibile a fornire risultati ottimali, ma ha come (piccola) contropartita la necessità di sottoporsi a una seduta di training, durante la quale il computer propone una lunga serie di frasi da leggere, e che la macchina userà per confrontarle con un suo modello vocale interno. Inutile dire che , in questo modo, l'ambiente si "affezionerà" a noi come un cane e , man mano che lo addestreremo (anche in sedute successive) diverrà sempre più ubbidiente e "comprensivo". Altra contropartita è anche quella che il modello è strettamente legato alla singola persona, e che ben difficilmente un nuovo utente potrà adottarne uno già esistente per il proprio uso.

Fortunatamente i tempi di training si sono, continuamente, ridotti. Una volta occorrevano almeno sei ore per raggiungere un livello accettabile di collaborazione, a fronte dell'ora di oggi. I costi sono proporzionalmente diminuiti (i prezzi sono tutti livellati intorno alle 200-300.000 lire) e le esigenze, in fatto di hardware, sono senz'altro accettabili (Dragon Dictate, il meno esigente in fatto di hardware, si accontenta di un Pentium a 166 MHz., 32 MB di RAM, e qualunque scheda audio a 16 bit) . Tutti i pacchetti , poi, offrono, compreso nel prezzo, un microfono direzionale a soppressione di rumore di fondo (pregevolissimo lo SpeechMike di Philips) e numerose facility per aggiornare il vocabolario di base, che già gode di almeno 50-60.000 vocaboli (si tenga presente che nel parlato comune non si usano più di 3-4.000 vocaboli), e un vocabolario esteso di diverse centinaia di migliaia di vocaboli.

Guarda, mamma, senza mani!

Sviluppato, all'inizio come aiuto per i disabili, l'input vocale toccherà, nel 2000, il tetto dei tre miliardi di dollari. Nuove applicazioni di tecnica di riconoscimento della voce vengono implementate ogni giorno e non è raro ritrovarsi a rispondere , a telefono, a una voce preregistrata che ci guida all'interno di menu di scelta. Le più grandi marche mon-



diali concorrono nel cooperare allo sviluppo, ancorché concorrenzialmente, della tecnologia. IBM, WordPerfect, Kurzweill, Dragon e, più recentemente, Philips si affrontano nella grande area con mezzi altamente sofisticati, potenti, disponibili e addirittura impensabili dieci anni fa. Curiosamente, tutte queste compagnie hanno articolato i loro sforzi in due direzioni.

Esiste una fiorente area di utilizzo verticale dell'input vocale, che è orientata, essenzialmente, al mercato legale e medico. La cosa ci può interessare fino a un certo punto; l'altra direzione è quella dell'utilizzo orizzontale, vale a dire la disponibilità di un software da addestrare e customizzare per poter soddisfare esigenze multiple, anche se , nella maggior parte dei casi, orientate al word processing generico. Il potenziale client qui è enorme, se si pensa che questo tipo di input potrà rappresentare una nuova pietra miliare nel campo dell'informatica, come non si vedeva da quindici anni circa, quando fu introdotto il mouse e l'interfaccia a desktop. L'utenza è fortemente motivata verso questo ambiente, visto che usare tastiera e mouse è cosa talora fastidiosa e perfino inagevole, mentre tutti sono familiari con la potenza insita nell'uso della voce. E, particolare di non poco conto, un buon ambiente di input vocale potrebbe recuperare grosse schiere di persone che , per pregiudizio o per approccio incorretto, hanno rinunciato all'uso di un PC.

Abbiamo già detto dei sistemi speaker-indipendent. La potenza crescente del software li sta mettendo rapidamente fuori gioco, a favore di pacchetti legati a ambienti vocali utente che, a fronte della richiesta di lettura di un brano de "L'isola del tesoro" o di una serie di frasi commerciali, regala un ambiente personale in cui ci si muoverà a proprio agio. In ottica di ulteriore riduzione dei tempi di training, alcuni produttori offrono modelli acustici del tipo "maschio-femmina", "alto-basso", "età", "accento" e così via. In base al modello, il software costruisce un profilo dell'utente , che contribuisce ancora di più a ridurre i tempi di allenamento, con livelli di riconoscimento, come accennavamo, del 95% o più, e velocità di almeno 150 parole (contro le 200 parole del parlato naturale veloce, senza pause).

Ma come funziona un input vocale? Ogni produttore usa una tecnica lievemente differente per l'interpretazione del parlato. Al contrario di quel che avveniva nei primi prodotti, oggi, grazie alla maggiore potenza dell'hardware, le parole vengono identificate nel contesto d'uso in una particolare frase. IBM ha sviluppato una tecnica proprietaria chiamata "trigrams". Un trigram è una combinazione caratteristica di tre parole (le combinazioni , ovviamente, sono differenti da lingua e lingua) che è più o meno frequente in una certa lingua. Con questo sistema, il programma può scegliere non solo tra parole che si assomigliano nel suono (si immagini una diversa persona in un verbo) ma anche tra omofoni, parole differenti che hanno lo stesso suono. All'atto pratico, il sistema analizza la frase tentando di individuare, in essa, la maggior quantità possibile di sillabe, che verranno poi organizzate per cercare di mettere insieme quante più parole e combinazioni di senso logico possibile.

I sistemi più avanzati di ricognizione vocale utilizzano, oggi, il metodo cosiddetto discreto. Nato all'inizio per funzionare inserendo una pausa tra parola e parola, oggi, grazie alla potenza più avanzata dei microprocessori, permette di gestire la parlata continua senza difficoltà. Altra difficoltà, che è divenuta di fatto oggi superata, era la mole del dizionario fornito e, di conseguenza, la parte di esso caricata in RAM e quello che veniva tenuto fuori e veniva caricato alla bisogna. Infine, altro termine importante nel gioco delle parti è il DSP, acronimo di Digital Signal Processor, strettamente riferito alla scheda sonora utilizzata. La disponibilità del DSP permette di affidare alla scheda audio parte dell'elaborazione in altri casi affidata al microprocessore; attualmente molti sistemi di ricognizione vocale funzionano solo se abbinati a schede DSP.

Conclusioni

Uscita dieci anni fa dai laboratori di ricerca e giunta oggi a livelli avanzati grazie alla cresciuta potenza delle macchine, la tecnica del riconoscimento vocale è oggi affidabile e capace di produrre, in mano a persone allenate, risultati di grande qualità. Paradossalmente, il problema nell'utilizzo non sta nei programmi, oggi quasi del tutto trasparenti, ma nella persona che li usa; chi ha avuto a che fare con questi pacchetti si accorge immediatamente, a meno di non esserci già abituato, di come sia poco familiare e, talvolta, innaturale dettare. Già, proprio così, forse il vero limite sta proprio nell'acquisire nuove tecniche di comportamento, ed è inutile negare che, per raggiungere risultati soddisfacenti, occorre allenare se stessi a questo nuovo modo di confrontarsi con la macchina. Ognuno userà la sua tecnica, ma poche regole basteranno. Oltre a pronunciare chiaramente le parole (e vedremo come è incredibilmente strano quante sillabe "ci mangiamo" nel discorrere), sarà necessario, almeno all'inizio, non preoccuparsi degli errori, anzi è consigliabile non guardare proprio lo schermo. Ma, come un cane fedele, qualunque programma sceglieremo si affezionerà sempre di più a noi e alla nostra parlata. Gran bel gioco, da portare avanti!