I quaderni di Telèma

a cura di Alberto Mucci

Le sfide 2006 della tecnologia della lingua

Non molti anni fa, l'obiettivo era dare (e ricevere) informazioni in quantità sufficiente alle necessità di una società evoluta. Oggi di informazione ce ne è troppa. Spesso si trasforma in rumore. Ed ecco che la tecnologia è chiamata a trovare soluzioni, a mettere a punto strumenti per gestire l'informazione, per poterla utilizzare in funzione di specifiche esigenze.

I progressi delle tecnologie della comunicazione, del trattamento automatico della lingua, parlata e scritta, sono importanti. Il 2005 è stato un anno di transizione, ma significativo nei risultati conseguiti, come dimostrano gli articoli di questo "Quaderno". Ed ora il 2004 affida al 2006 fra le altre, due sfide importanti: l'intelligence (legata ai pressanti problemi della sicurezza) un tema che tratteremo in un prossimo numero; la medicina, nella diagnosi e nella cura delle patologie del linguaggio.

Il TAL (trattamento automatico della lingua) ha nell'intelligence l'area di maggior sviluppo. Per molte ragioni. Innanzitutto per ragioni di sicurezza, per individuare cioè nel mare delle informazioni che circolano, quelle che possono avere riflessi nella lotta al terrorismo. Ma anche per individuare, sempre nel mare di informazioni, quelle che possono avere risvolti nelle scelte di politica industriale (da parte di governi e/o di singole aziende). Non si tratta di fare operazioni di spionaggio industriale, ma di ricevere e selezionare informazioni che esistono, attraverso metodi che la tecnologia riesce a definire.

La sfida 2006 nella medicina riguarda la diagnosi e la cura delle patologie del linguaggio, Anche qui passi avanti sono stati fatti, altri sono possibili a breve. Si possono citare l'orecchio bionico per i sordi (un punto di eccellenza è l'Università di Verona) e l'uso delle tecnologie TAL con i non vedenti.

"Comunicare con le macchine" abbiamo titolato il "Quaderno di Telèma", pubblicato a dicembre 2004: macchine sempre più sofisticate, tecnologicamente avanzate e sempre più utili per selezionare e rendere prontamente utilizzabili le molteplici informazioni che caratterizzano la società della comunicazione.

Supplemento al numero 232 di dicembre 2005/gennaio 2006 di

MEDIA DUEMILA

Indice

Il quaderno di Telèma è stato realizzato dalla Fondazione Ugo Bordoni (Presidente il Prof. Giordano Bruno Guerri, Direttore Generale il Consigliere Guido Salerno, Direttore delle Ricerche l'ing. Mario Frullone). Coordinatore del Quaderno l'ing. Andrea Paoloni. Hanno collaborato: Michele Mignano, IRCCS S. Raffaele Roma; Silvia Mosso, Sheyla Militello, Loquendo; Roberto Pieraccini, Tell-Eureka Corporation.

Sono usciti nel 2004/2005:

TAL e medicina

Gli ambiti applicativi delle tecnologie della lingua sono molto vari, vanno dalla criminalistica all'editoria, dalla ricerca nelle basi di dati alle applicazioni d'ufficio, dalla traduzione automatica alla telefonia.

La criminalistica utilizza il riconoscimento del parlante per identificare, sulla base della traccia fornita dalla loro voce, gli autori di sequestri di persona o di estorsioni oppure utilizza i sistemi di trascrizione automatica plurilingue per "distillare", da un flusso di comunicazioni, un sottoinsieme che contiene informazioni di interesse per gli investigatori; nelle applicazioni di ufficio oltre al correttore ortografico e grammaticale, inserito in ogni programma di scrittura computerizzata, si possono utilizzare sistemi in grado di verificare la leggibilità dei testi prodotti al fine di migliorarla e sistemi di ricerca delle informazioni basati su principi linguistici. Tra le applicazioni di ufficio si può anche annoverare la traduzione automatica da lingua a lingua, sebbene a questo tema sia opportuno dedicare, stante l'importanza applicativa, un ambito specifico; vi sono poi le numerose applicazioni relative alla telefonia fissa e mobile, che vanno dalla codifica della voce, utilizzata nella telefonia cellulare per ridurre la banda occupata da una comunicazione, ai "call center" con tecnologia vocale in cui un automa consente, ad esempio, all'utente di prenotare un biglietto aereo o ferroviario a qualsiasi ora del giorno e della notte. Tra le applicazioni delle tecnologie della lingua, vogliamo qui approfondire quelle legate alla medicina e all'ausilio ai disabili. In questo campo le due principali aree di intervento sono quelle relative ai difetti dell'udito ai difetti della fonazione. Nel presente numero dei Quaderni di Telèma altri due articoli illustreranno in maggior dettaglio sia gli ausili tecnologici a disposizione dei disabili per leggere e comunicare, sia le tecniche di analisi della voce che vengono impiegate, e sempre più verranno impiegate, nella cura dei difetti della parola.

Nel presente lavoro si fornirà una panoramica di queste aree e verranno date alcune indicazioni sugli sviluppi futuri delle tecnologie TAL applicate alla medicina.

Sui difetti della fonazione

I disturbi della comunicazione parlata e scritta, sono più frequenti di quanto si creda mentre il loro studio può ritenersi ancora in una fase primitiva. I disturbi più frequenti riguardano la dislessia che colpisce circa il 4% dei bambini. Con il termine dislessia si denota l'incapacità di leggere e ricordare le parole scritte principalmente a causa della trasposizione delle sillabe durante la lettura. Secondo una recente teoria l'origine di questo inconveniente sarebbe un difetto della comunicazione nei due emisferi celebrali. Nelle prime fasi dell'apprendimento verrebbe utilizzato l'emisfero destro, mentre crescendo il controllo passerebbe all'emisfero sinistro che consente l'attivazione di regole sintattiche e semantiche. Nel caso della dislessia questo passaggio avverrebbe o troppo presto, con conseguente lettura veloce con molti errori, o troppo tardi, con conseguente lettura lenta e frammentaria. Un'altra teoria ritiene che la dislessia sia dovuta ad un'anomalia dei meccanismi visivi in particolare per un campo visivo troppo ampio che causa una dispersione dell'attenzione. La cura in questo caso consiste nel limitare il campo visivo del paziente. Il disturbo può essere diagnosticato in maniera più precisa se non ci si basa esclusivamente sui sintomi ma si fa uso delle tecniche di tomografia celebrale. Queste tecniche consentono di mettere a rilievo le aree del cervello attive nel momento che vengono eseguiti dei compiti che vengono assegnati.

Non meno frequenti sono i disturbi della parola che sono conseguenza di traumi o alterazioni funzionali dell'apparato articolatorio. Non entreremo qui nel dettaglio dei diversi sintomi che affliggono il parlato a seconda delle alterazioni funzionali presenti, vogliamo tuttavia rilevare l'importanza che uno studio oggettivo della qualità della voce riveste, sia ai fini diagnostici, sia ai fini di valutare l'efficacia delle cure somministrate.

Le tecniche del Trattamento Automatico della Lingua (TAL) possono anche essere utilizzate a fine riabilitativo, a supporto degli esercizi somministrati dal logopedista.

Il logopedista è un operatore sanitario specializzato nell'educazione e nella rieducazione dei disturbi della voce, della parola. Questa professione richiede non solo competenza, professionalità e preparazione, ma doti e qualità umane quali sensibilità, intuito, pazienza, creatività. Il logopedista ha un ruolo importante in moltissimi settori; nel campo dell'età evolutiva collabora alla diagnosi di disturbi specifici di linguaggio, di disturbi d'apprendimento, di disfluenze (balbuzie) e si pone come riferimento nella loro rieducazione. Nel campo della voce il logopedista rieduca i disturbi vocali da disfunzione e post intervento; nel campo della neuropsicologia cognitiva formula diagnosi funzionali in soggetti cerebrolesi e programma iter riabilitativi personalizzati, coerenti e razionali.

Sui disturbi dell'udito

Se la foniatria e la logopedia si occupano della "sintesi della voce" ovvero della capacità dell'uomo di generare il parlato, l'audiologia si occupa dello studio e della cura dei difetti auditivi ovvero della capacità che ha l'uomo di "riconoscere il parlato". Anche in questo caso le patologie possibili possono essere suddivise in quelle relative al funzionamento del cervello o neurologiche e quelle che debbono essere addebitate più propriamente al funzionamento dell'apparato auditivo e del nervo acustico che collega tale apparato al

Figura 1. Una moderna protesi acustica.

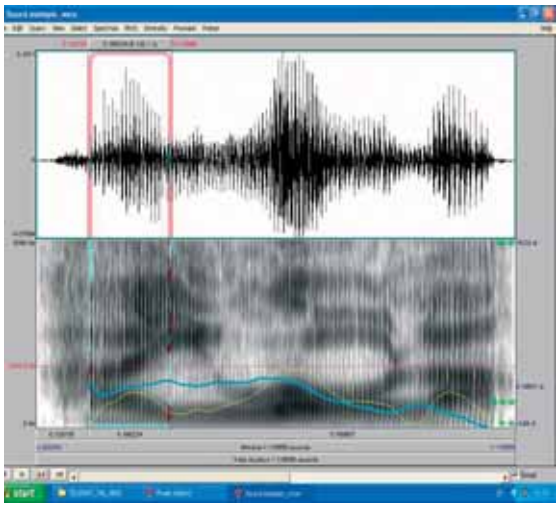

Figura 2. L'analisi elettroacustica della parola "aiuole".

cervello. Questi disturbi vengono per lo più risolti utilizzando degli amplificatori in grado di rinforzare il suono a sufficienza da stimolare un udito carente, tuttavia quando ad essere danneggiati è la coclea il sistema di stimolazione deve agire direttamente al livello del nervo acustico, stimolando quel trasduttore suono-impulsi nervosi che è costituito dalla coclea. Questi impianti che operano il miracolo di far udire i sordi profondi, sono chiamati impianti cocleari; si tratta del primo organo di senso artificiale creato dall'uomo e può essere considerato, a ragione, una delle più importanti conquiste della moderna medicina. L'Impianto Cocleare, altrimenti definito Protesi Cocleare, è un dispositivo finalizzato alla stimolazione elettrica della coclea in grado di sostituire la funzione delle cellule ciliate cocleari e di trasformare il suono in impulsi elettrici che stimolano direttamente il nervo acustico, in modo da determinare sensazioni uditive a livello delle aree corticali in soggetti con sordità totale o profonda.

Gli impianti cocleari multicanale attualmente utilizzati sono costituiti da una parte esterna e da una parte interna che viene impiantata chirurgicamente. Il componente interno è impiantato sotto la pelle, in un letto scavato nell'osso dietro l'orecchio. C'è un filo porta elettrodi inserito nella coclea, la sua "coda" è un elettrodo che permette all'impianto di utilizzare differenti metodi di stimolazione. All'interno dell'impianto si

trova un magnete che tiene l'antenna trasmittente nella corretta posizione.

Il componente esterno è costituito da un microfono direzionale e da un processore del linguaggio. Si tratta di sofisticati microcomputer che velocemente processano il suono in informazioni digitali, mandando milioni di istruzioni per secondo. Ciascun processore è programmato digitalmente, garantendo la capacità di un fine aggiustamento del processore stesso alle necessità di udito di ciascun paziente.

Il loro funzionamento attualmente è assai rudimentale e non consente certamente ai sordi di sostenere un dialogo normalmente a causa del numero ridotto di canali attivi resi disponibili. Il progresso della tecnologia permetterà tuttavia di aumentare il numero delle connessioni con il nervo acustico sino a giungere a valori sufficienti a garantire una precisione di analisi sufficiente alla completa comprensione del parlato.

L'ausilio ai disabili

Come abbiamo visto nei precedenti paragrafi il trattamento automatico della lingua fornisce strumenti assai utili per diagnosticare problemi comunicativi dovuti sia a difetti uditivi sia a difetti fonatori. Queste tecnologie inoltre sono in grado di fornire strumenti per la cura delle patologie come gli impianti cocleari e gli strumenti per la riabilitazione.

Le tecnologie dell'elaborazione della lingua intervengono anche in ausilio dei non vedenti e degli ipovedenti, grazie ai sistemi per la lettura automatica degli schermi dei computer ma anche gli schermi dei cellulari e di altri dispositivi di comunicazione. Altre applicazioni di questa tecnologia intervengono nella la domotica, ovvero la possibilità di comunicare verbalmente con gli elettrodomestici; questa tecnologia risulta di importanza fondamentale per persone affette da handicap motori, e inoltre è di grande ausilio alle persone anziane.

Guardando al futuro

Le tecnologie TAL hanno raggiunto un grado di maturità sufficiente a rappresentare la soluzione per un numero sempre più grande di problemi diversi. Certamente un ruolo sempre più

Figura 3. Il logo della conferenza TAL 2006 (9-10 marzo).

importante, come abbiamo visto, lo svolgono e lo svolgeranno nel campo della medicina e dell'ausilio ai disabili. sia nel campo della riabilitazione dei soggetti affetti da disturbi della comunicazione, sia nel campo della comunicazione uomo macchina per rendere più semplice all'uomo il controllo dei dispositivi che ci circondano, dall'automobile con comandi vocali, agli elettrodomestici della nostra casa. Tuttavia le prospettive di sviluppo sono assai più vaste sia nel campo affine dell'accessibilità, sia in campi meno usuali come i giochi, dove personaggi robotici avranno spazi sempre più interattivi, sia nella traduzione automatica che almeno nella versione testuale trova sempre più applicazioni. La tecnologia TAL sta dando risultati sempre più interessanti nel campo dell'annotazione automatica dei segnali consentendo l'analisi automatica dei contenuti.

Nel prossimo anno avranno inizio alcuni progetti interessanti nell'ambito del TAL e più precisamente della trascrizione automatica. Sempre nel 2006 si svolgerà in primavera una conferenza denominata "TAL 2006" che farà il punto sulle applicazioni TAL dal punto di vista dell'utente. Ci auguriamo che tutte queste iniziative possano favorire un ulteriore importante sviluppo delle tecnologie della lingua.

Andrea Paoloni Fondazione Ugo Bordoni

Il computer nei disturbi della voce

Nell'ambito della comunicazione umana la voce costituisce l'elemento fondamentale dell'espressione verbale-fonatoria. Ma cosa si intende con il termine voce e che cos'è una bella voce? La voce è una caratteristica della persona e dal suo timbro possiamo identificare colui che parla o, se si tratta di un estraneo, indovinare la sua età, cultura, regione di provenienza, stato emotivo.

La qualità della voce è fondamentale per i professionisti della parola e ancor più per i cantanti. Quando un vero professionista della voce, esperto della tecnica, ci trascina con lui noi viviamo le sue stesse sensazioni, noi respiriamo insieme a lui, la nostra laringe funziona senza contrazioni.

Con il termine voce si indica un prodotto acustico costituito da un suono complesso provocato attivamente dalla laringe in presenza di una corrente respiratoria, normalmente espiratoria. In altri termini "la voce" costituisce quella parte del linguaggio articolato realizzata con la vibrazione delle corde vocali.

La voce è un fenomeno multidimensionale, un insieme di componente armonica, rumore ed energia variabile e variamente distribuita per ogni sua componente. Dal punto di vista fisico ogni singola componente può essere scomposta ed analizzata anche numericamente. Esaminando la voce con gli analizzatori, ovvero con apparecchi che permettono di scomporla nei suoi valori di frequenza, si scoprono caratteristiche molto specifiche. Il suono emesso è una vibrazione acustica caratterizzata da un volume, una forma, e una coloritura. Il volume viene chiamato intensità e la forma è determinata dal tono. Il più grave, viene chiamato frequenza fondamentale e rappresenta la frequenza inferiore tra quelle contenute nel suono emesso; è la frequenza di base di vibrazione delle corde vocali. La coloritura infine è legata alla tessitura delle frequenze sovrapposte, alla presenza e all'intensità delle armoniche che si vanno formando nei risuonatori, cioè nelle cavità poste al disopra della laringe, la principale delle quali è l'orofaringe. Dalla composizione delle armoniche risulta l'oggetto sonoro che sarà più o meno colorito, più o meno brillante, più o meno scuro. La valutazione del suono, della sua qualità, dipende quindi dai parametri che abbiamo appena elencato; un suono di qualità, una volta eseguita l'analisi, si rivela un suono molto ricco di coloritura. La valutazione della qualità resta in ogni caso qualcosa di molto soggettivo e molto individuale, legata alla perce-

Figura 4. Confronto tra una laringe normale ed una affetta da noduli.

zione acustica del risultato finale del fenomeno sonoro. Inoltre la voce è un fenomeno assolutamente individuale, in relazione anche alla caratteristiche morfologiche degli organi deputati alla sua produzione, le corde vocali e le cavità di risonanza (tratto vocale).

Le malattie della voce

La qualità della voce può essere perduta a causa di malattie di varia natura o, in particolare nei professionisti della parola e del canto, a causa dell'usura a cui vanno incontro gli organi vocali se sottoposti a sforzi eccessivi. Questi sforzi possono essere causati da un abuso dell'utilizzo del mezzo fonatorio o, più frequentemente, da un suo cattivo impiego. Il sintomo delle alterazioni dell'apparato vocale è costituito dalla disfonia, ovvero da un'alterazione della voce parlata in senso quantitativo e qualitativo o dall'assenza di voce, definita afonia. I disturbi della voce cantata vengono denominati disodie.

Poiché la voce è un fenomeno globale, e nella sua produzione coinvolge l'individuo totalmente, possono influire su di essa tutti quei fenomeni che alterano l'equilibrio psico-fisico dell'individuo stesso, quali alterazioni dello stato emotivo, stress, stanchezza, alterazioni del ritmo sonno-veglia, alterazioni posturali.

Secondo Segre le disfonie vanno classificate in disfonie da lesioni organiche e disfonie da alterazioni funzionali.

Nell'ambito delle disfonia da **lesioni organiche** ricordiamo le forme infiammatorie acute e croniche (laringiti acute e croniche), le disfonie post traumi laringo - cervicali, le disfonie da alterazioni embrionali (glottide palmata, solco cordale, ipoplasie cordali congenite), le disfonie da paralisi laringee (nella miastenia gravis, nelle lesioni nervose), le disfonie da neoformazioni laringee benigne (polipi, cisti, papillomi, etc.) e maligne. Inoltre vanno ricordale le disfonie di origine ormonale o causate da alterazioni del sistema nervoso centrale.

Nell'ambito delle disfonie da **alterazioni funzionali** ricordiamo: la disfonia cronica infantile, le difonie psicogene, la disfonie ipocinetiche, ipercinetiche, la disfonia spastica, le disfonie professionali, le disfonie sopravvenute nel corso di disturbi della muta vocale, etc.

Esistono inoltre le disfonie originate da cause miste tra le quali ricordiamo i noduli.

La valutazione della qualità della voce

La valutazione percettiva della qualità vocale è uno degli argomenti più largamente trattati in campo foniatrico ed ha dato luogo a studi in tutto il mondo, non sempre univoci.

Tali approcci valutativi morfologici o percettivo-soggettivi, sono spesso imprecisi, variabili da soggetto a soggetto, da esaminatore ad esaminatore, sulla base delle singole interpretazioni o dell'esperienza raggiunta nel campo specifico. Pertanto appare necessario, nell'ambito dello studio dell'analisi vocale in campo clinico, integrare la valutazione percettiva con analisi oggettive.

Attualmente solo la "voce" è studiata in maniera abbastanza completa, con valutazione integrata percettiva/oggettiva effettuata con l'ausilio della spettrografia vocale digitale associata allo studio multiparametrico MDVP (multi dimentional voice program). Il programma MDVP produce una rappresentazione grafica di 22 parametri della voce (frequenza fondamentale e sue derivate, energia del rumore e delle armoniche, variazioni di ampiezza e sue derivate, numerosità dei tratti sordi, etc.). Il grafico radiale, normalizzato con i valori di una base di dati di voci normali, avrà un aspetto regolare e comunque sarà contenuto all'interno di un cerchio di soglia, mentre valori patologici si porranno all'esterno del cerchio citato.

Tutta la produzione verbale (linguaggio parlato, lettura, etc.) è per lo più valutata quasi esclusivamente in maniera percettiva con tutti i limiti che abbiamo già rappresentato. Vi è inoltre da segnalare che i metodi di valutazione e i parametri oggettivi presi in esame variano da una clinica all'altra in quanto non esistono a tutt'oggi metodi stan-

dard universalmente riconosciuti per la valutazione obiettiva della qualità della voce. Dopo aver effettuato una diagnosi della patologia in atto si procede, quando necessario, agli opportuni interventi medico chirurgici e successivamente agli adeguati trattamenti riabilitativi.

I moderni trattamenti riabilitativi dei disturbi della voce prevedono innanzitutto un'osservazione dell'atteggiamento posturale, durante la fase fonatoria, valutazione della respirazione prevalente a riposo ed in fonazione spontanea, motilità della mandibola e grado di apertura della bocca, presenza di eventuali contratture a carico del collo, uso delle modalità comunicative non verbali ed inoltre, con l'ausilio della spettrografia vocale, si può effettuare un'analisi percettiva-oggettiva della produzione verbale. Si può valutare l'andamento temporale (ritmo dell'eloquio o velocità di fonazione, le pause di rifornimento, durata della frase), valori inerenti l'intensità e la frequenza, qualità del timbro laringeo, aspetti relativi alla fono articolazione (mobilità della mandibola, grado di apertura della bocca, mobilità linguali, presenza di rumori, tipo di attacco vocale (dolce, duro, soffiato, etc.), qualità generale della voce (soffiata, tesa, intubata, pressata, strozzata, etc.).

Figura 5. Un sistema di fibro-laringoscopia elettronica (naso).

Pertanto la riabilitazione dei disturbi della voce comporta un intervento integrato sulla postura, sulla respirazione (con eventuale correzione del pattern respiratorio), sulle fono articolazioni (motilità delle lingua e della mandibola) e quindi sull'emissione vocale e sulla produzione verbale.

Attualmente il trattamento riabilitativo dei disturbi della voce è facilitato da software in grado di fornire in tempo reale dati relativi a tutti gli aspetti elettroacustici dell'emissione vocale (presenza del suono, estensione ed intensità sonora, presenza di emissione sonora, estensione altezza tonale, controllo altezza tonale, etc.). Tali software permettono inoltre al paziente, guidato dal terapista, di utilizzare un feed-back visivo piuttosto che acustico, acquisendo una propiocezione corporea maggiore a supporto della propria tecnica vocale.

Nuove tecniche diagnostiche

La tecnica diagnostica tradizionale prevede la ripetizione di alcune vocali e l'analisi percettiva e/o oggettiva dei suoni prodotti. Al fine di un significativo miglioramento sia degli aspetti diagnostici, sia della valutazione dell'efficacia dei trattamenti riabilitativi, è necessario procedere alla standardizzazione dei metodi che consentono di quantificare l'entità della disfonia e di valutare in modo obiettivo i risultati conseguiti con un trattamento riabilitativo.

Un altro elemento da tenere in dovuta con-

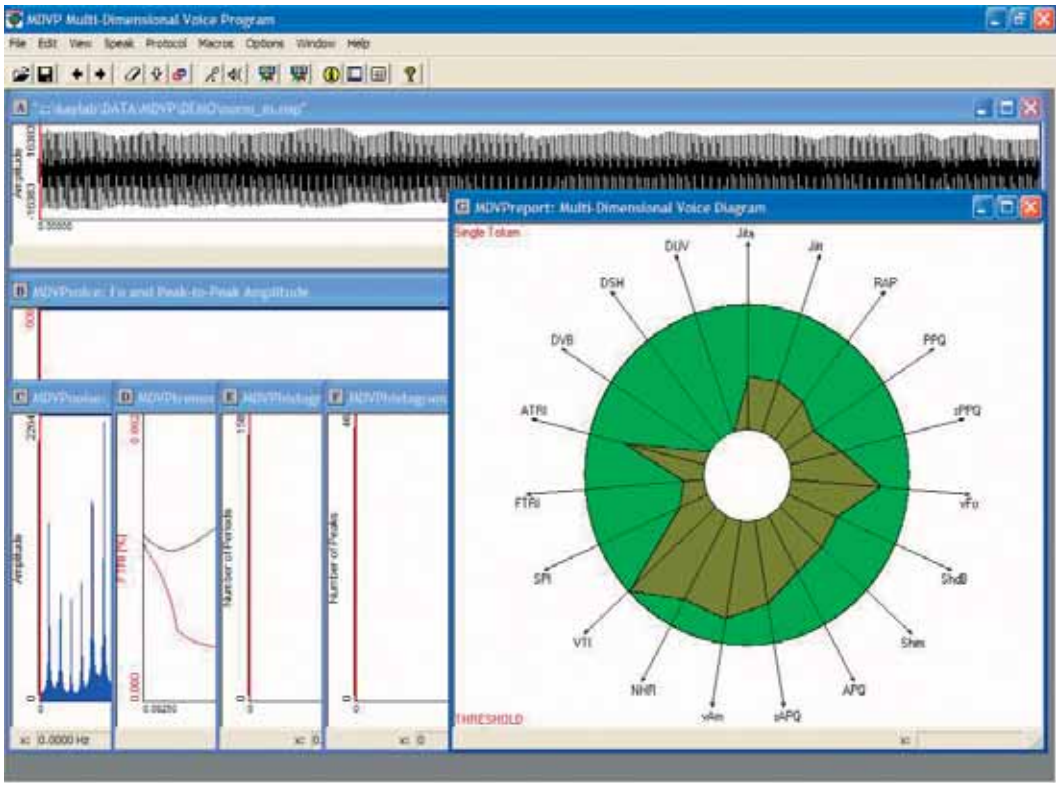

Figura 6. MDVP report: analisi della voce eseguita con la tecnica multiparametrica.

siderazione è la misurazione dei parametri elettroacustici non in condizioni statiche, sui suoni sostenuti, ma nelle condizioni dinamiche del parlato spontaneo. Infatti in questa situazione le vocali si combinano variamente con le consonanti e sono prodotte con grande variabilità di volume e di frequenza. La difficoltà di condurre correttamente un intervento riabilitativo è proprio quella di far sì che il paziente sia in grado di gestire la dinamica dei volumi e delle frequenze nel loro andamento temporale, soprattutto in situazioni in cui sia emotivamente coinvolto. È chiaro pertanto che la voce può subire grosse variazioni se impegnata in lettura o utilizzata nel parlato spontaneo e che una riabilitazione ben riuscita si può considerare tale solo quando il paziente riesce a gestire bene entrambe le situazioni.

La ricerca svolta presso numerose cliniche è pertanto diretta allo sviluppo di nuove tecniche di diagnosi, più strettamente connesse con l'effettiva meccanica fono-articolatoria e di nuovi parametri, in grado di caratterizzare in modo oggettivo le diverse disfonie.

Dott. Michele Mignano
Specialista in Otorinolaringoiatrie e Foniatria
Dirigente del Servizio di Foniatria
IRCCS S. Raffaele Roma

La tecnologia TAL in aiuto ai disabili

L'accesso alle informazioni e la possibilità di comunicare rappresentano oggi dei diritti primari per tutti i cittadini. Le nuove tecnologie dell'informazione hanno permesso di superare le limitazioni di tempo e di spazio, offrendo nuove soluzioni impensabili fino a pochi anni fa.

Tali tecnologie potenzialmente potrebbero offrire nuove e straordinarie opportunità anche ai diversamente abili ed agli anziani, ma di fatto le categorie più deboli sono in gran parte rimaste escluse a causa della mancanza di accessibilità. Troppo spesso le tecnologie e gli strumenti di comunicazione sono studiati per i clienti "normodotati" e non prendono in considerazione i bisogni degli utenti con esigenze speciali, creando all'atto pratico nuove barriere. Le tecnologie di trattamento automatico del linguaggio (TAL) possono contribuire alla rimozione di queste barriere, ma possono fare molto di più offrendo agli individui con disabilità nuovi mezzi per migliorare la propria indipendenza e favorendone la partecipazione nella vita sociale e lavorativa.

Figura 7. Esempio di interfaccia di uno screen reader per telefono cellulare.

In particolare, le tecnologie di sintesi vocale (TTS) e riconoscimento del parlato (ASR) possono creare le condizioni per una maggiore accessibilità. Una tecnologia accessibile permette di adattare i supporti informativi per venire incontro a bisogni specifici di tipo visivo, uditivo, motorio, cognitivo e verbale. Essa include sia le opzioni di accessibilità inserite nativamente nel prodotto sia specializzazioni hardware e software del prodotto stesso (assistive technology products) che aiutano gli individui ad interagire con il computer.

Alcune esemplificazioni, di seguito riportate, ci forniscono un quadro per identificare la gamma di abilità fisiche e cognitive tra la popolazione adulta in età lavorativa e gli attuali utilizzatori del computer, il tipo di difficoltà e impedimenti che ne limitano gli scopi di attività ed il loro grado di severità, ed il numero di persone che possono potenzialmente beneficiare dall'utilizzo delle tecnologie assistite.

Queste informazioni, comparate con la tendenza all'invecchiamento della popolazione, possono aiutare a spiegare l'impatto dell'invecchiamento della popolazione sulle modalità di utilizzo dei supporti informatici ed il bisogno di tecnologie assistite.

Quando si pensa ai disabili sensoriali che possono beneficiare della sintesi vocale, la prima e più immediata applicazione che viene in mente è quella legata ai non vedenti. Questo dato appare di maggiore rilevanza se si pensa che l'incidenza dei disturbi ed impedimenti nell'utilizzo del computer tra i diversamente abili, vede al primo posto i disturbi visivi. Vista, destrezza manuale e disturbi dell'udito sono da considerarsi i più comuni impedimenti. Secondo uno studio commissionato da Microsoft e condotto da Forrester Research, Inc (2003), approssimativamente un utilizzatore su quattro (25%) ha un disturbo di tipo visivo, il 24% di disabili che utilizzano il computer ha un disturbo legato alla destrezza manuale, circa uno su cinque (20%) ha disturbi di tipo uditivo, una percentuale minore ha disturbi di tipo cognitivo (16%) mentre una percentuale minore presenta disturbi del linguaggio (3%).

I più noti ausili per disabili visivi sono gli screen reader, prodotti che, grazie alla sintesi

Figura 8. IdeoVoice Mobile.

vocale, leggono le informazioni che compaiono sullo schermo e le ripetono ad alta voce; questi prodotti permettono la lettura di testi, descrizioni di grafici, tabelle e campi in input, consentendo anche lo scorrimento del testo in avanti e all'indietro.

Oggi gli screen reader sono disponibili, oltre che per i personal computer, anche per PDA e telefoni cellulari.

Mobile Speak di Code Factory (www.codefactory.es) è uno dei primi esempi di screen reader per telefoni cellulari. Consente ai non vedenti ed agli ipovedenti di ricevere feedback vocali sul contenuto dello schermo e sul menù del telefono, garantendo un utilizzo semplice e in totale autonomia dello strumento. È possibile variare i parametri della voce, ad esempio velocità e volume, e scegliere la modalità di lettura dei contenuti, identificando il livello di dettaglio ideale che si vuole ascoltare sulle voci scelte (Figura 7).

Altri prodotti pensati per non vedenti e ipovedenti sono gli applicativi che integrano la sintesi vocale con software di tipo OCR (optical character recognition), oppure con tastiere Braille, offrendo una maggiore versatilità nell'utilizzo dello strumento.

E ancora a proposito di telefoni cellulari, in Italia TIM ha lanciato già da qualche anno il servizio "SMS non vedenti", che permette ai clienti disabili di ascoltare gli SMS ricevuti e di rispondere in voce. Il servizio, evolutosi in un'applicazione residente sui cellulari tipo smartphone, permette oggi di trasformare in voce tutte le funzioni del telefono cellulare, grazie a un'offerta completa di funzionalità, tra cui la variazione delle impostazioni, la gestione della rubrica e del calendario degli appuntamenti, l'opportunità di scrivere e leggere email.

Oltre che per i non vedenti, la tecnologia di sintesi vocale serve anche per dare voce a chi non ce l'ha, consentendo di "parlare" con un vocabolario illimitato. Le tecnologie assistive di questo tipo possono essere presenti su dispositivi dedicati oppure software integrabili in un comune PC.

Esempi di applicazioni dedicate a questa tipologia di disabilità sono rappresentati dalle diverse soluzioni di "comunicatore simbolico" offerte dalla società canadese Oralys (www.oralys.ca), che offrono l'opportunità alle persone impossibilitate a parlare di esprimersi in modo indipendente, svincolandole dal linguaggio dei segni.

Queste applicazioni sono disponibili anche su palmare e consentono di comunicare in modo immediato mediante la scelta degli ideogrammi presenti sullo schermo. La selezione delle icone, possibile sia tramite penna sia con la pressione del dito, permette di comporre frasi che vengono visualizzate e lette dal software di sintesi vocale, in diverse lingue. Gli ideogrammi disponibili, e la possibi-

Figura 9. Icona PDA eVALUES.

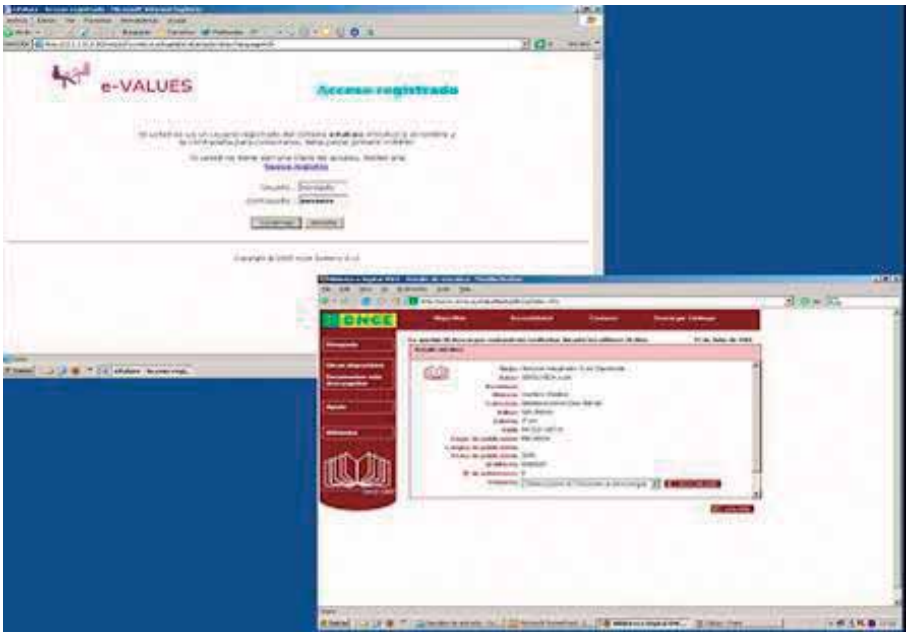

Figura 10. Sito associazione ONCE.

lità di definirne di personalizzati, consentono la creazione di un numero illimitato di frasi, adatte a qualunque tipo di situazione. In alternativa al testo scritto/vocalizzato, è possibile con alcune soluzioni ottenere come risultato una rappresentazione nel linguaggio dei segni (Figura 8).

Un altro esempio di utilizzo delle tecnologie del TAL è l'impiego di sintesi e riconoscimento vocale negli apparecchi telefonici dedicati alle persone che soffrono di limitazione della funzione uditiva.

Esistono telefoni per sordomuti, dotati di tastiera e touch screen, che trasformano in voce sintetica i messaggi scritti dall'utente; le risposte che arrivano dall'altro capo del telefono in voce, grazie al software di riconoscimento vocale vengono convertite in un messaggio grafico sul video del telefono.

È importante sottolineare come le persone che hanno difficoltà nella sfera della comunicazione richiedano caratteristiche precise alla tecnologia: il principale requisito per gli individui che non possono parlare con la propria voce è la scelta della voce sintetica più adatta (ad esempio in base all'età), ma anche la possibilità di variazione del tono, della velocità di eloquio e così via.

Dai risultati dei sondaggi effettuati nei confronti di persone che presentano gravi deficit vocali, risulta che la tecnologia di sintesi vocale utilizzata in tali ambiti deve presentare una voce accurata, deve prevedere la possibilità di variare la velocità di eloquio, ed essere rapida nella reazione ai comandi.

Nell'area delle disabilità sensoriali, il progetto finanziato dall'Unione Europea eVALUES costituisce un valido esempio di come le tecnologie vocali possano essere poste al servizio di utenti con difficoltà.

Figura 11. Accesso eVALUES da postazione PC.

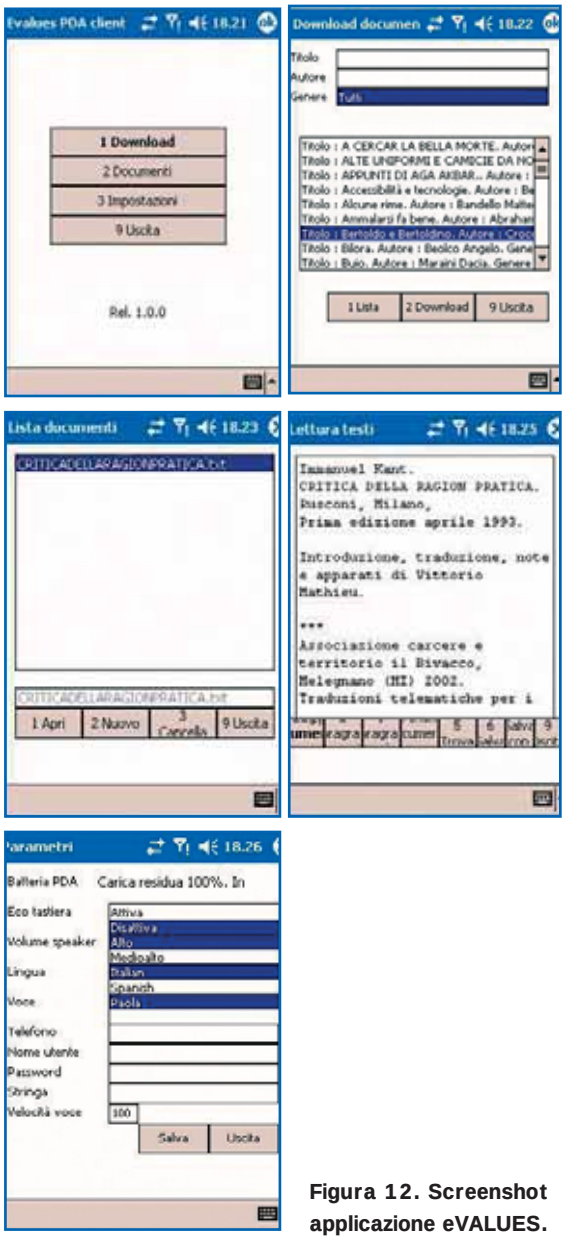

Figura 12. Screenshot applicazione eVALUES.

Obiettivi progetto market validation eVALUES

Il progetto eVALUES si propone la fornitura di un servizio trans-europeo, basato su Internet, per la lettura, tramite voce sintetica, di libri e altri documenti scritti, nonché delle notizie. Basandosi su soluzioni vocali esistenti il progetto intende fornire agli utenti (ciechi o ipovedenti, anziani, dislessici e altri utenti con problemi di lettura), la straordinaria possibilità di accedere a ogni informazione scritta ed effettuarne il download, attraverso Personal Computer, ma soprattutto device portatili come PDA, in grado di garantire la fruizione dei testi anche in mobilità e al di fuori delle mura domestiche (in autobus, a scuola, al parco, dagli amici).

Il servizio è attualmente fornito come prototipo dall'Unione Italiana dei Ciechi (UIC). Nell'ambito del progetto eVALUES l'evoluzione di tale servizio è stata testata con due trial pilota, in Italia ed in Spagna.

Obiettivo finale è la validazione di mercato, per la successiva introduzione del servizio eVALUES, in diversi Paesi europei, per i cittadini con difficoltà di lettura.

A tal fine, un gruppo principale di utenti formato e coordinato dall'Unione Italiana dei Ciechi, con un secondo gruppo di utenti spagnoli di ONCE, Organizzazione Nazionale dei Ciechi di Spagna, ha testato il servizio pilota per la validazione di mercato.

L'analisi dei requisiti necessari per soddisfare le esigenze degli utenti ed ottenere il successo del servizio sul mercato, condotta nell'ambito del progetto, è stata cruciale per una migliore conoscenza del mercato di destinazione e degli inerenti rischi nella fase iniziale di introduzione sul mercato, e anche per determinare la direzione futura da seguire in termini di soddisfacimento, da parte del servizio e dell'interfaccia vocale in particolare, dei requisiti utente.

Il progetto ha consentito di fornire un servizio evoluto, attento alla domanda ed ai futuri bisogni dei cittadini europei, basato sull'uso delle tecnologie vocali frutto di lunghi anni di ricerca e sviluppo in Loquendo, sull'esperienza di settore di Voice Systems e di Moviquity per la componente ICT, e soprattutto sul diretto contributo delle Associazioni degli utenti (UIC e ONCE).

I grafici seguenti sintetizzano i risultati delle analisi condotte con gli utenti, da cui emerge un generale apprezzamento per il servizio, e una netta propensione all'utilizzo anche in mobilità, resa possibile dalla compattezza del dispositivo.

Figura 13. Sessione training per uso servizio eVALUES.

Si è accennato prima alla possibilità di sfruttare al meglio le tecnologie vocali anche per forme di disabilità diverse da quelle che interessano la sfera della comunicazione.

Grandi passi avanti si stanno facendo per supportare i disabili motori; per loro esistono sistemi di input vocale che, emulando i comandi della tastiera e/o del mouse, consentono di interagire con il computer semplicemente con l'uso della voce. A questi si possono aggiungere i software di dettatura del testo, non pensati specificamente per i disabili, ma che possono rivelarsi un valido aiuto per chi ha difficoltà nel movimento o nella coordinazione.

Una grande sfida è rappresentata dalle potenzialità delle tecnologie del TAL come ausili informatici nel contesto educativo e riabilitativo: le tecnologie vocali possono essere validamente impiegate per supportare gli studenti con difficoltà di lettura o di apprendimento, grazie all'aggiunta di un feedback vocale a quello visuale. La sintesi vocale può, in questo senso, favorire il processo di apprendimento.

È possibile pensare anche a sistemi di supporto per attività di logopedia che, attraverso l'analisi della voce, permettano la diagnosi e la riabilitazione delle diverse patologie.

In aggiunta a quanto detto sinora, è importante ricordare che esistono anche altri ambiti applicativi che possono beneficiare delle tecnologie del TAL. La sintesi e, soprattutto, il riconoscimento del parlato, possono supportare disabili motori e anziani nello svolgimento dei compiti quotidiani per raggiungere una maggiore autonomia. Chi di noi non sogna la casa del futuro in cui sia possibile il controllo dell'ambiente con la voce? Perché non immaginare una casa automatizzata in cui si possa "parlare" con il televisore e con la lavatrice?

Figura 14. Presentazione servizio eVALUES HANDYMATICA.

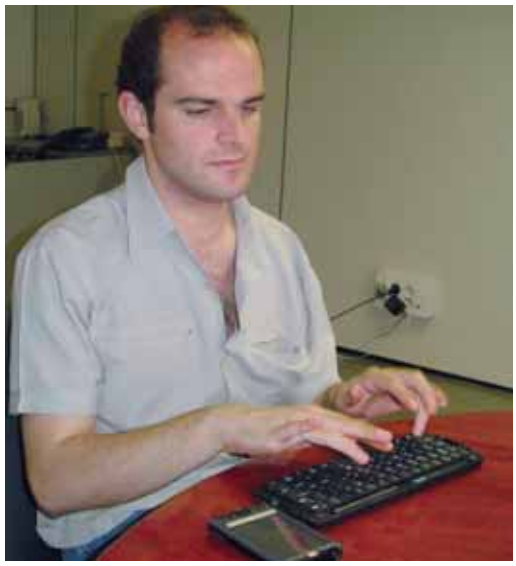

Figura 15. Non vedente in contesto d'uso eVALUES.

Un futuro che riguarda le Tecnologie per l'Accessibilità

Gli avanzamenti tecnologici incrociano spesso le loro tracce con i temi dell'accessibilità per soggetti con disabilità di vario genere. Molti progressi hanno in generale reso possibile per i diversamente abili trovare un lavoro, mentre nello stesso tempo altri lo hanno reso più difficile, soprattutto per alcuni soggetti con disabilità nell'utilizzo del computer. Sebbene i prodotti di "assistive technology" abbiano seguito in larga parte questi avanzamenti per garantire l'accessibilità dove sarebbe stata in altro modo impossibile ed abbiano reso possibile che soggetti con disabilità potessero divenire produttivi in un contesto "business", questi stessi prodotti hanno richiesto per il loro sviluppo una reingegnerizzazione, spesso in senso inverso così da porsi indietro rispetto a quegli stessi prodotti che supplivano. Oggi i produttori di hardware e software devono lavorare insieme per fornire le necessarie informazioni richieste da interfacce specializzate indispensabili per venire incontro ai bisogni dei soggetti disabili, come ad esempio il text-to-speech e le descrizioni vocalizzate delle interfacce utente (GUI).

Oggi siamo sul confine di un rivoluzionario mutamento della tecnologia. Gli utilizzatori di supporti informatici, primo tra essi il computer, richiedono interfacce utente che non siano solo "semplici da usare", ma anche più flessibili e facilmente personalizzabili. Essi sentono che utilizzare un computer è sempre più difficile di quanto dovrebbe essere. Al crescere dell'età media della popolazione, tale esigenza sarà sentita in modo sempre più forte; gli utilizzatori di supporti informatici vorranno anche accessi remoti intelligenti da ogni luogo. Più importante, i consumatori vorranno scegliere come interagire con il computer. La ricerca e sviluppo in cui si è investito per

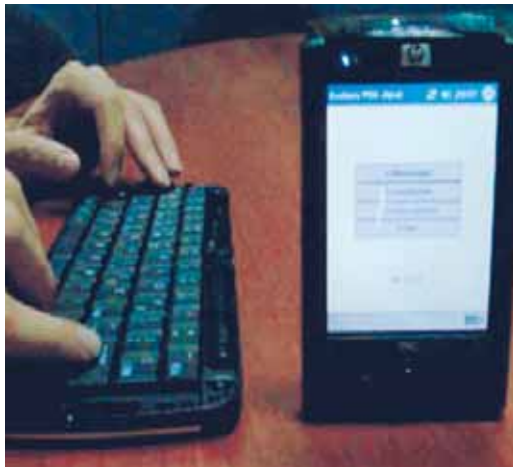

Figura 16. Dettaglio uso eVALUES con tastiera bluetooth.

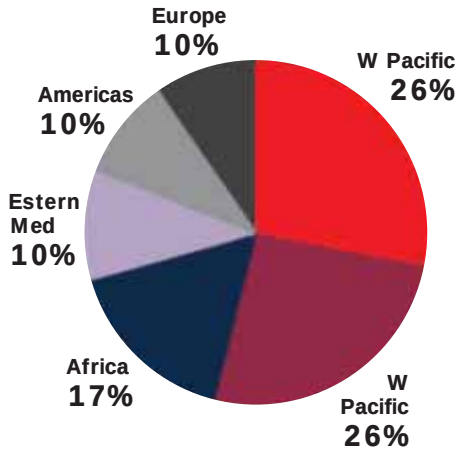

Figura 17. Distribuzione disturbi visivi per regioni WHO.

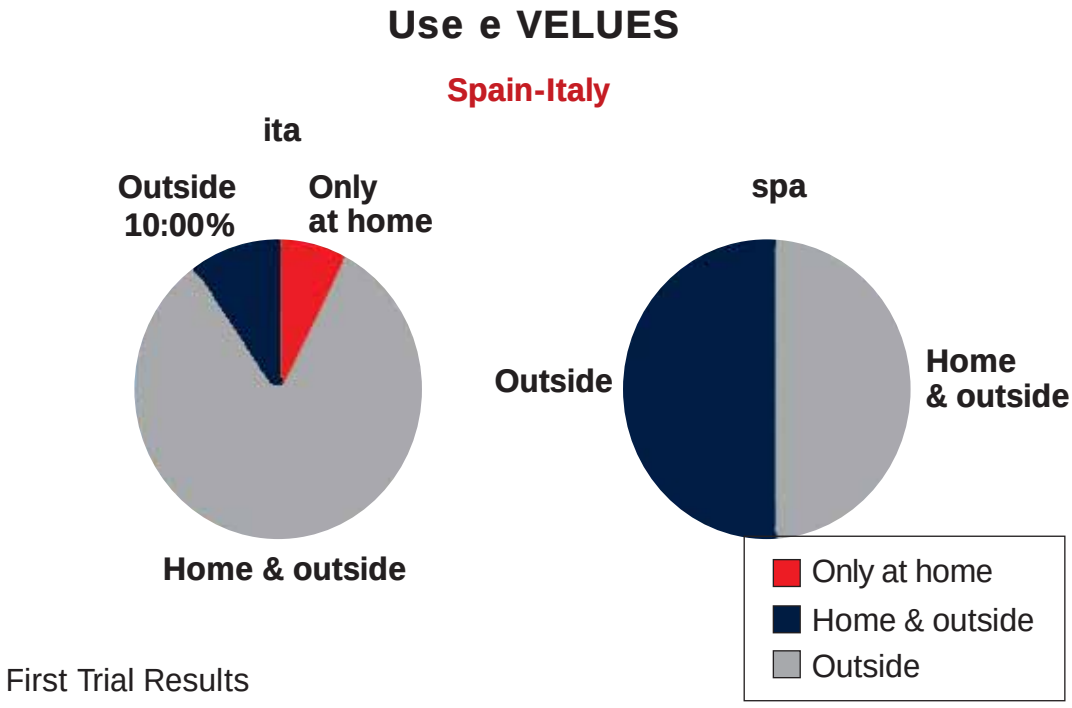

Figura 18. Grafico utilizzo eVALUES (Mobilità vs. Home).

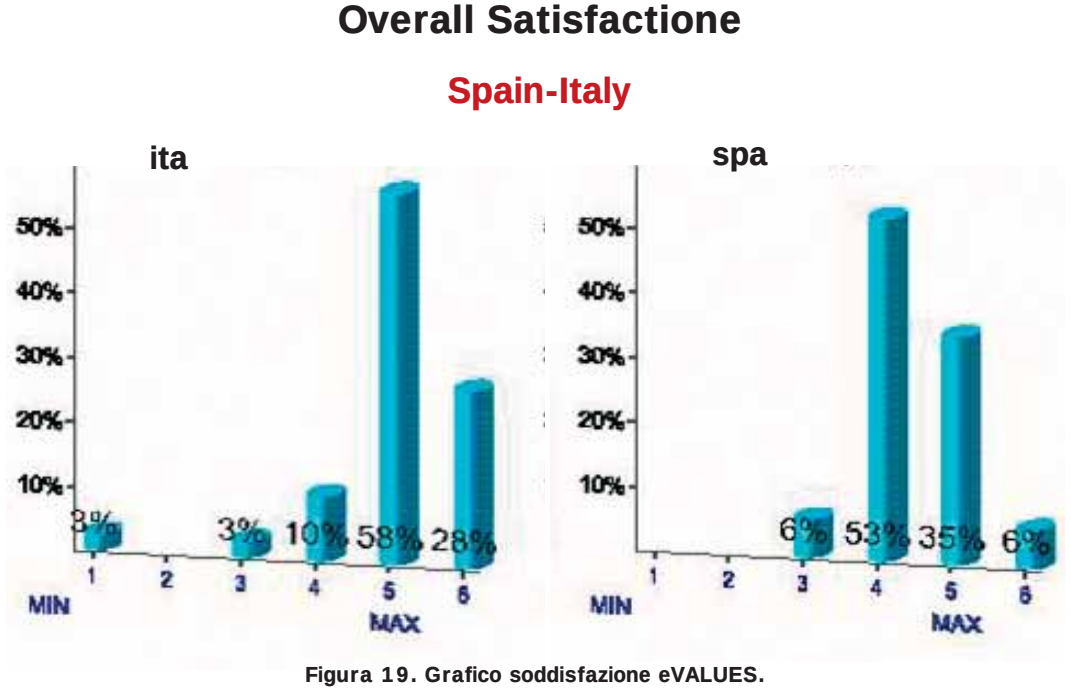

Figura 19. Grafico soddisfazione eVALUES.

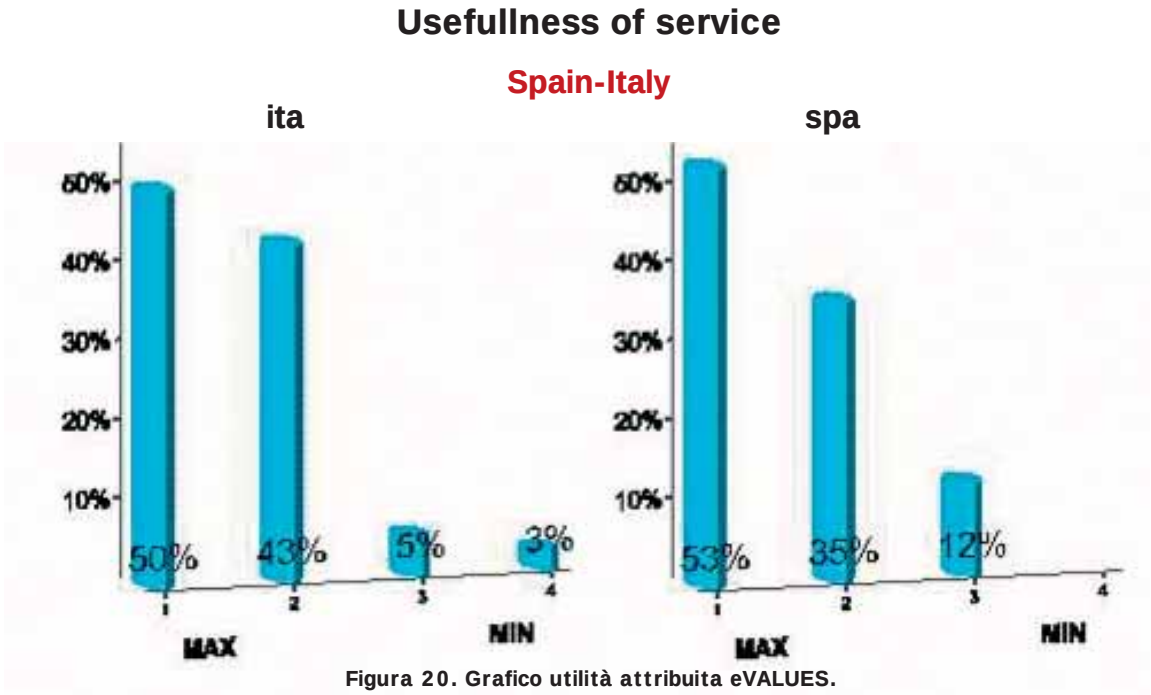

Figura 20. Grafico utilità attribuita eVALUES.

realizzare tali mutamenti, si fonda sulle lezioni imparate quando le tecnologie assistite hanno per prime dovuto far fronte proprio ad alcune di queste stesse necessità. Se i mutamenti tecnologici della prossima decade mantenessero gli stessi obiettivi in comune, offrirebbero significativi aumenti nella produttività ed accrescerebbero la flessibilità per tutti gli utilizzatori di supporti informatici, specialmente i soggetti portatori di disabilità. Questi stessi mutamenti offrono anche la promessa di accrescere il potere della capacità d'uso degli strumenti informatici ad una più ampia fascia di utilizzatori finali, con o senza disabilità.

Sebbene sia difficile predire attendibilmente il futuro, si possono tuttavia utilizzare i progetti di ricerca e sviluppo come indicatori di dove la tecnologia sia diretta. Si può così anticipare che alcuni dispositivi che paiono oggi sogni irrealizzabili saranno solamente un punto di partenza per tecnologie oggi inimmaginabili.

Silvia Mosso Business Analyst
Business Development area
Sheyla Militello
Ergonomist Business Development area
Loquendo Vocal Technologies and Services

Nuove tecniche per parlare con i computer

Il primo sistema di riconoscimento vocale di cui si abbia notizia è sicuramente quello costruito nel 1952 ai Bell Laboratories da Davis, Biddulph, e Balashek [1]. AUDREY era un sistema analogico - un *rack* alto oltre due metri pieno di valvole termoioniche e relais - in grado di capire sequenze di numeri separati da pause. Per ogni numero pronunciato, AUDREY faceva illuminare la lampadina corrispondente sul pannello frontale. AUDREY era in grado di raggiungere una precisione compresa fra il 97% e il 99%, purché tarato sulla voce di chi lo utilizzava. In realtà, AUDREY non fu mai un successo commerciale. Anche se il gigante delle telecomunicazioni AT&T, che ne aveva sponsorizzato la costruzione, avrebbe potuto estenderne l'utilizzo alla composizione di numeri telefonici da parte delle centinaia di miglia di operatori telefonici o per clienti particolarmente benestanti, AUDREY non fu mai in grado di competere, per costo ed efficienza, con i sistemi elettromeccanici rimasti in uso fino a pochi decenni fa. Comporre numeri premendo tasti o facendo girare rotelle era, per quei tempi, decisamente più veloce, affidabile ed economico rispetto all'uso della voce.

Ci sono voluti più di trent'anni dopo AUDREY perché, dopo l'introduzione dei computer e l'evoluzione della scienza dei segnali digitali, si definisse una tecnologia per il riconoscimento vocale basata su teorie statistiche e matematiche. Ci sono poi voluti altri vent'anni di ricerca e di continuo miglioramento perché questa tecnologia raggiungesse un livello di prestazioni accettabili. Solamente all'inizio di questo secolo, quasi cinquant'anni dopo AUDREY, si è visto emergere un mercato e un'industria maturi per lo sviluppo e l'utilizzo commerciale della tecnologia della voce. Ciononostante, i principi su cui sono basati i sistemi di comprensione vocale odierni sono fondamentalmente gli stessi di quelli inventati verso la fine degli anni '70. La differenza sostanziale nelle prestazioni attuali rispetto a quelle di allora è essenzialmente dovuta alla disponibilità di computer di diversi ordini di grandezza più potenti, alla quantità enorme di dati vocali trascritti e opportunamente trattati per l'addestramento automatico dei modelli statistici e al continuo processo di raffinamento delle tecnologie.

La comprensione automatica della voce, ben lontana da prestazioni anche vagamente confrontabili con quelle umane, è oggi comunque in grado di offrire benefici commerciali sostanziali per un considerevole numero di applicazioni estese a diversi settori del mercato: dalla dettatura di documenti alla trascrizione di conversazioni fra agenti e utenti, alla traduzione automatica della voce. Ma l'applicazione che sta dimostrando lo sbocco commerciale più esteso è quella dei sistemi di dialogo, conosciuti anche come *sistemi conversazionali,* cioè sistemi che integrano le tecnologie vocali con la gestione di una conversazione atta a fornire un servizio, dare informazioni, o risolvere un problema specifico.

Le tecnologie della voce oggi

Nonostante che nel passato si sia cercato di sviluppare sistemi di riconoscimento della voce basati sulla conoscenza profonda dei fenomeni fonetici e linguistici, è ormai universalmente accettato che gli unici sistemi che dimostrano prestazioni ragionevoli siano quelli basati sulla "forza bruta" computazionale e sulla statistica. La situazione è simile a quella che si è osservata, per esempio, nei programmi per il gioco degli scacchi, dove la capacità di immagazzinamento, gestione e ricerca di un numero assai elevato di dati hanno determinato il successo rispetto a tecniche basate sulla inferenza e sull'intelligenza artificiale classica.

Alla fine degli anni '70, il gruppo di ricerca di IBM a Yorktown Heights, NY, guidato da Fred Jelinek, ha sviluppato tecniche statistiche, basate sulle teorie dei modelli Markoviani a stati nasco-

sti (Hidden Markov Models o HMM), che hanno dato prova di essere particolarmente adatte a modellare i fenomeni fonetici e linguistici. I vantaggi dei sistemi statistici di riconoscimento della voce [2] sono la loro base matematica, la loro semplicità di sviluppo - che non richiede profonde conoscenze linguistiche - e la capacità di essere addestrati automaticamente su quantità di dati vocali pressoché illimitate. Questi sistemi di riconoscimento della voce sono rimasti imbattuti fino a oggi, nonostante i vari tentativi del mondo della ricerca di muoversi verso altre tecnologie, come le reti neuronali (*neural networks*) e i sistemi basati sulla conoscenza (*knowledge based systems*). "Ogni volta che licenzio un linguista, le prestazioni del mio sistema di riconoscimento migliorano", era il motto di Fred Jelinek[1] che rinforzava una fede totale nei modelli puramente statistici. Praticamente, tutti i motori di riconoscimento della voce sul mercato mondiale sono oggi basati su modelli statistici HMM.

Il problema complementare al riconoscimento della voce - cioè la generazione di una rappresentazione vocale del testo, è la sintesi della voce (TTS-Text to Speech). I primi tentativi di sintesi automatica della voce risalgono al 1939, quando una macchina "parlante", il Voder (Figura 21), progettata e realizzata ai Bell Laboratories da Homer Dudley, fu presentata alla World Fair di New York.

Per diversi anni si sono visti approcci diversi alla sintesi della voce [3], come la sintesi per formanti, basata su regole per modellare le traiettorie dei picchi di risonanza dello spettro (formanti), la sintesi articolatoria, basata su un modello dell'apparato fonatorio umano dal punto di vista meccanico e acustico, e la sintesi per difoni.

[1] Si racconta che Fred Jelinek abbia detto questa frase, poi diventata famosa, a una conferenza nel 1988. Nel 2004, ha poi rivisitato la sua relazione con il mondo dei linguisti con un intervento, ad una conferenza internazionale, dal titolo *Alcuni dei miei migliori amici sono linguisti.*

SCIENCE NEWS LETTER, *for January 14, 1939* 19

ENGINEERING

Now a Machine That Talks With the Voice of Man

Voder Combines Electrical Currents to Produce Synthetic Speech; Converses at Signal From Keys

See Front Cover

NATURE took hundreds of thousands of years to teach man how to speak.

In two years, scientists have taught a machine how to talk, translating into real words and sentences signals punched into its controlling keyboard.

Controlled by a skilled operator who has learned how to mix the sounds the device's two electric discharge tubes produce, it combines varying electric currents that an amplifier turns into real speech. No phonograph records of any kind are ever used. It is the first device that actually creates human speech.

The name of this new robot is the Voder.

"Practice makes perfect," it told its first enthusiastic hearers at its debut before the Franklin Institute. It isn't perfect yet. But the Voder was good enough to convince its audience that the Fourth ment allez-vous?" it repeated "Hasenpfeffer" and "Comment allez-vous?" with perfect aplomb.

The machine resulted from efforts of Bell scientists H. W. Dudley and R. R. Riesz in fundamental telephone research. They developed an electrical speech analyzer and a speech synthesizer, both of which were demonstrated at the Harvard Tercentenary. The former machine fed an electrical control pattern into the synthesizer. With the exception of the organ-like keyboard, all its parts are in regular telephone use.

When the subject of the Bell exhibit at the New York and San Francisco fairs came up, it was suggested that the machine that talked, when the analyzer fed it the proper pattern, would be a fit display—particularly if it could be made to speak when an operator punched a keyboard instead of merely seeing that the analyzer continued feeding it the

Mr. Riesz is also known for his work in connection with the development of the artificial larynx. He and his associates are accustomed to calling the Voder "Pedro" after the Brazilian emperor,

Figura 21. Science News Letter del Gennaio 1939 con un articolo sul Voder, la prima "macchina parlante elettrica", costruita da Homer Dudley agli AT&T Bell Laboratories e presentata alla World Fair di New York. La foto mostra un'operatrice alla tastiera del Voder.

Per quanto i metodi di sintesi della voce per regole, articolatori o per difoni, abbiano impegnato i ricercatori per diverse decine di anni nella costruzione di sistemi complessi, verso la fine degli anni '90 è emerso un sistema di sintesi rivoluzionario, concettualmente molto più semplice dei precedenti, ma con prestazioni notevolmente superiori: la sintesi concatenativa. Il principio della sintesi concatenativa è basato sulla registrazione di una grande quantità di elementi fonetici di durata diversa: fonemi, difoni, sillabe, parole e frasi intere. La raccolta di questi elementi fonetici viene strutturata attraverso una base di dati che ne permetta la ricerca per mezzo di parametri, quali l'intonazione e il contesto fonetico. Al momento della sintesi, si effettua una ricerca nella base di dati al fine di trovare la sequenza ottimale di elementi fonetici per la costruzione della frase richiesta, tenendo in considerazione il contesto e l'intonazione desiderata.

Al fine di procedere alla selezione ottimale degli elementi fonetici e dell'intonazione, la sintesi della voce da testo richiede un'analisi sofisticata del messaggio. Si pensi, per esempio, alla parola "ancora", la cui intonazione dipende dal contesto ("Ho ancora un'ancora della vecchia nave"). Non solo, ma il testo da sintetizzare può contenere abbreviazioni e acronimi che richiedono conoscenze adeguate per ottenerne la pronuncia corretta (per esempio in "Gent.mo Sig. Rossi abitante in P.zza della Repubblica No. 6").

Mentre l'obiettivo originale dei primi sistemi di sintesi da testo era quello di garantire l'intelligibilità della voce negli anni successivi, una volta raggiunto un livello sufficiente al loro utilizzo, la ricerca si è focalizzata sulla naturalezza e quindi sulla

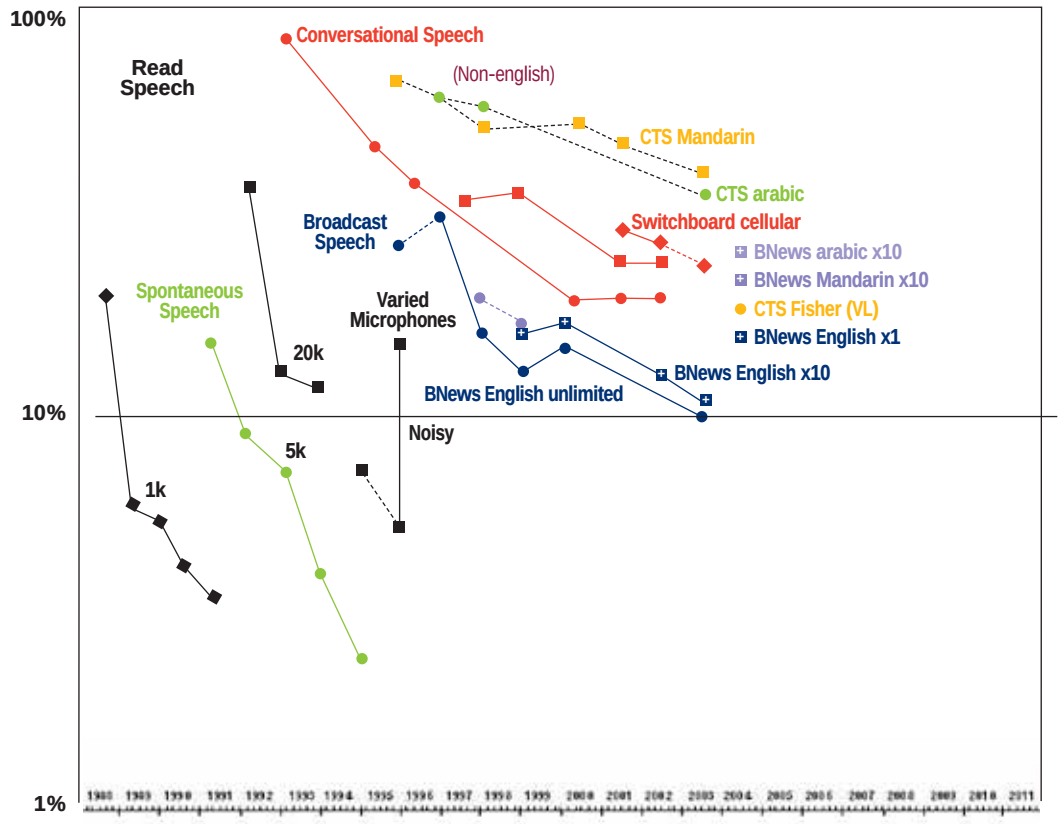

Figura 22. L'evoluzione delle prestazioni del riconoscimento della voce attraverso i vari progetti finanziati da DARPA dal 1988 a oggi. Ogni curva rappresenta la percentuale di errore per parola pronunciata in applicazioni di difficoltà crescente. Mentre le prime applicazioni alla fine degli anni '80 riguardavano il riconoscimento di poche migliaia di parole lette da testo, oggi si cerca di riconoscere il parlato con vocabolari pressoché illimitati da conversazioni spontanee e da programmi televisivi e radiofonici.

personalità, espressività, e riproduzione delle emozioni. I sistemi di sintesi di oggi, anche grazie alla possibilità di annotare il testo con elementi espressivi per mezzo dello standard SSML (Speech Synthesis Markup Language), permettono di generare voce quasi indistinguibile dalla voce umana.

Cambio di prospettiva e l'industria della voce

L'evoluzione delle tecnologie vocali è stata per molti anni prerogativa unica della ricerca accademica e dei centri di ricerca industriali, come, per esempio, AT&T Bell Laboratiores e IBM T.J.Watson Research negli Stati Uniti, CSELT (Torino), France Telecom e Philips in Europa, e Nippon Telephon and Telegraph (NTT) in Giappone. Tuttavia, fino alla metà degli anni'90, lo sfruttamento commerciale delle tecnologie della voce è sempre stato molto esiguo. La tecnologia, specialmente quella del riconoscimento della voce, non era pronta per il mercato.

Nella prima metà degli anni '90, la comunità scientifica nel settore del riconoscimento della voce era principalmente dedicata al continuo miglioramento delle prestazioni delle tecnologie vocali tramite progetti sponsorizzati da agenzie come DARPA (Defense Advanced Research Project Agency – Figura 22). Proprio in quegli anni, due piccole aziende fecero la loro comparsa su un mercato quasi inesistente. La prima fu Corona, successivamente chiamata Nuance, uno spin-off dello Stanford Research Institute (SRI). La seconda, uno spin-off del Massachussetts Institute of Technology (MIT), fu Altech, successivamente nota come SpeechWorks[2].

Mentre la comunità scientifica aveva come obiettivo la realizzazione di sistemi e algoritmi con capacità linguistiche simili a quelle umane, e quindi il riconoscimento e la comprensione del linguaggio naturale senza alcun vincolo sulle espressioni, Nuance e SpeechWorks adottavano una prospettiva diversa. L'idea era di quella di realizzare applicazioni che fossero abbastanza semplici da poter utilizzare con successo i sistemi di riconoscimento della voce disponibili e di ingegnerizzare l'interfaccia in modo da offrire agli utenti un'esperienza soddisfacente e senz'altro superiore a quella dei sistemi *touch-tone* (prema uno per… prema due per…), anche a costo di limitare la naturalezza e libertá di espressione. Dunque, i primi obiettivi commerciali sono stati applicazioni molto semplici, come il *tracking* dei colli basato sul codice di spedizione (FedEx e UPS) e applicazioni poco più complesse, come le informazioni in tempo reale del valore di mercato dei titoli di borsa (Schwab ed E-trade), e informazioni sullo stato dei voli (American Airlines e United Airlines).

Il principio guida era quella di garantire una esperienza positiva all'utente, il quale era quindi al centro dei criteri di progetto (*user-centered design)*. Infatti, sia che l'interfaccia permetta una conversazione libera, simile a quella umana, o che le richieste siano basate su menù consecutivi, il raggiungimento dell'obiettvo nel modo più veloce e più semplice - e senza errori - risulta comunque essere il fattore più importante per l'utente di un qualunque servizio.

A questo proposito occorre ricordare che nonostante gli sforzi della comunità scientifica internazionale, ancora oggi la comprensione di espressioni vocali libere da ogni vincolo linguistico è soggetta a un tasso di errore significativo. Al contrario, limitando le possibili espressioni degli utenti mediante interfacce e *prompts* progettati accuratamente (sistemi di *directed dialog*) si può ottenere un livello di accuratezza necessario al raggiungimento di livelli di automazione decenti.

Nuance e SpeechWorks hanno percorso un cammino a ritroso nella realizzazione del sogno di macchine in grado di comprendere il linguaggio umano (come HAL 9000 di 2001 Odissea nello Spazio) e si sono impegnate nella ricerca di obiettivi più realistici e realizzabili con la tecnologia disponibile. La limitazione delle espressioni dell'utente per mezzo delle interfacce vocali è stato lo strumento principale del loro successo. Il loro obiettivo era quello di costruire macchine utili e non necessariamente antropomorfe.

[2] SpeechWorks è stata poi acquistata da Scansoft nel 2003, la quale ha poi acquistato Nuance nel 2005. Dall'Ottobre 2005, Scansoft, che adesso si chiama Nuance, possiede più dell 80% del mercato dei motori di riconoscimento e sintesi della voce.

Le prime applicazioni "conversazionali" [4] della metà degli anni '90 hanno mostrato subito il loro valore commerciale e hanno cominciato ad attrarre l'interesse di altre aziende. Le prime industrie che hanno adottato i sistemi di *directed dialog* costruiti da SpeechWorks e Nuance sono quelle dei trasporti, delle spedizioni e della finanze,

Il concetto di applicazioni basate sul *directed dialog* è presto diventato un'alternativa efficiente alle applicazioni *touch-tone.* Nel frattempo è emersa una nuova figura professionale: il progettista di Voice User Interface (VUI). Il progettista VUI è responsabile per la specifica completa dell'interazione fra utente e sistema mediante il progetto delle frasi pronunciate dal computer (*prompts*) e del cosiddetto *call-flow* (Figura 25), cioè una descrizione accurata del dialogo mediante grafi arbitrariamente complessi. Contemporaneamente, la metodologia di sviluppo delle applicazioni ha assunto il rigore tipico dei progetti di ingegnerizzazione del software, con fasi di analisi dei requisiti, specifica, progetto, sviluppo, test, ecc. Il processo, codificato nei passi fondamentali dello sviluppo del software e comprensivo di elementi specifici delle applicazioni vocali (come per esempio lo sviluppo e il test delle grammatiche, e le prove di usabilità), ha reso possibile la realizzazione e la riproducibilità di applicazioni di qualità commerciale a costi contenuti e prevedibili.

La disponibilità di tecnologie sempre più sofisticate e con prestazioni sempre più alte ha spinto l'industria, agli inizi del 2000, a muoversi con cautela verso applicazioni più complesse. Un esempio è la tecnologia, sviluppata inizialmente agli AT&T Labs e conosciuta come HMIHY (How May I Help You) [5] o *call-routing*, che prevede la classificazione automatica di frasi in linguaggio naturale in un numero predefinito di categorie. Anche se lontana dal fornire una risposta al problema generale della comprensione del linguaggio naturale, la tecnologia HMIHY sta dando prova di essere estremamente utile ed efficace nello sviluppo di applicazioni avanzate.

Lo sviluppo dell'industria della voce

Alla metà degli anni '90, il maggior interesse commerciale delle aziende come Nuance e SpeechWorks era la commercializzazione di tecnologia di base. In realtà, per poter creare un mercato per la tecnologia, Nuance e SpeechWorks hanno dovuto assumere inizialmente ruoli industriali diversi, quali la realizzazione di applicazioni, e l'integrazione della tecnologia di base in piattaforme e strumenti di sviluppo. Con lo sviluppo del mercato, altre aziende sono entrate nell'industria assumendo ruoli specifici. Oggi, l'industria della voce ha assunto un struttura matura (Tabella 1) e basata su livelli di competenza diversi. Alla base ci sono le aziende che sviluppano tecnologie quali il riconoscimento e la sintesi della voce. Nuance (con circa l'80% del mercato), IBM e Microsoft sono i maggiori distributori negli Stati Uniti. Loquendo in

Settore dell'industria	**Prodotto**	**Esempi**
Hosting	Gestione di infrastrutture e risorse	Convergys, West
Applicazioni	Implementazione di applicazioni e servizi	Tuvox, Edify, Tell-Eureka, Citec-voce
Strumenti	Strumenti per la realizzazione, messa a punto, e gestione delle applicazioni	Audium, Fluency, Voice Objects
Piattaforme	Infrastrutture hardware e software, middleware	VoiceGenie, Voxeo, IBM, Loquendo
Tecnologia di base	Motori di riconoscimento e sintesi della voce	Nuace, IBM, Microsoft, Loquendo, Telisma

Tabella 1

Italia e Telisma in Francia hanno una discreta porzione del mercato europeo e internazionale. Al di sopra delle tecnologie di base ci sono le piattaforme, cioè l'hardware e il software che integrano i sistemi di riconoscimento e di sintesi con la telefonia. Esempi di aziende che commercializzano piattaforme per applicazioni conversazionali sono VoiceGenie, Voxeo, IBM e la stessa Loquendo in Italia. Aziende specializzate in software per lo sviluppo e la messa a punto di applicazioni (*tools*) sono, tra le altre, Audium, Fluency e Voice Objects. Ci sono poi gli integratori di applicazioni, quali Tuvox, Edify, Tell-Eureka negli Stati Uniti e Citec Voce in Italia. E infine, aziende, quali Convergys e West, forniscono e gestiscono servizi di infrastruttura e di *hosting* per la messa in linea delle applicazioni.

Con la crescita del numero di aziende coinvolte nella tecnologia della voce e con il numero di applicazioni sviluppate dell'ordine delle centinaia all'anno (negli USA), è emersa la necessità degli standard. Gli standard sono estremamente importanti per assicurare l'interoperabilità dei componenti indipendentemente dall'azienda che li fornisce. VoiceXML, per esempio, è oggi il linguaggio standard per la realizzazione di sistemi di dialogo basati su una architettura *browser-server* simile a quella adottata dalla Web (Figura 23). VoiceXML 1.0[3] standardizzato dal W3C (World Wide Web Consortium), migliorato ed esteso da VoiceXML 2.0[4]; nel 2004 è stato poi seguito da altri standard, come MRCP[5] (Media Resource Control Protocol), un protocollo per il controllo dei motori di sintesi e di riconoscimento della voce, SRGS[6] (Speech Recognition Grammar Specification), un linguaggio per la specifica delle grammatiche, CCXML[7] (Call Control Markup Language), un linguaggio pe il controllo della telefonia, SSML (Speech Synthesis Markup Language) per il controllo della sintesi della voce, e EMMA[8] (Extensible Multi Modal Annotation), per la rappresentazione del significato (semantica) nei sistemi vocali e multimodali.

3 http://www.w3.org/TR/voicexml/
4 http://www.w3.org/TR/voicexml20/
5 http://www.ietf.org/internet-drafts/draft-shanmugham-mrcp-06.txt
6 http://www.w3.org/TR/speech-grammar/
7 http://www.w3.org/TR/ccxml/
8 http://www.w3.org/TR/emma/

Modelli commerciali

I vari settori dell'industria della voce applicano modelli commerciali diversi. Per esempio, le tecnologie di base, le piattaforme e gli strumenti di sviluppo sono motivate da un profitto basato sulla licenza del software e degli aggiornamenti, mentre l'*hosting* è monetizzato con un costo associato all'uso (per esempio misurato in minuti di connessione).

Le applicazioni sono state, fin dall'inizio, tradizionalmente associate a costi di sviluppo e di manutenzione. In realtà, il modello commerciale delle applicazioni vocali basato sui costi di sviluppo sta lentamente cedendo il passo a quello basato su applicazioni pre-confezionate (*pre-packaged application*) in settori verticali specifici. Per esempio, esistono oggi pacchetti di applicazioni vocali per servizi bancari, finanziari, e medici. La commercializzazione di applicazioni pre-confezionate e configurabili per il cliente specifico (per esempio, cambiando la voce, la lingua o la logica dell'applicazione) è in grado di ridurre sia i costi di sviluppo - in quanto la stessa applicazione può essere riutilizzata con minime modifiche per clienti diversi - sia il rischio associato alla sua adozione - in quanto la stessa applicazione, già ottimizzata e messa a punto per massimizzarne le prestazioni, può essere utilizzata in altri contesti. Mentre un modello commerciale delle applicazioni pre-confezionate può essere quello della licenza d'uso, come la grande maggioranza dei prodotti software, alcune aziende hanno adottato con successo un modello basato sul valore apportato dal servizio. Questo è l'esempio di applicazioni per servizi di utente (*customer care*), dove sistemi conversazionali basati sulle tecnologie della voce possono aiutare le aziende a ridurre i costi di servizio. Infatti, il costo per transazione, o per minuto, di una applicazione conversazionale può essere notevolmente inferiore rispetto al corrispondente costo di un operatore umano.

Il futuro

Nonostante l'uso delle tecnologie vocali abbia avuto oggi un grande sviluppo nel settore dei sistemi telefonici di dialogo per la realizzazione di servizi d'utente, l'industria si sta muovendo

verso altri settori del mercato. Un settore che sta trovando sempre più impiego e uso nella vita quotidiana è quello dell'*embedded*. Oggi si possono realizzare sistemi di riconoscimento e di sintesi con esigenze di memoria e capacità di calcolo molto ridotte. Per esempio, è possibile includere sistemi di riconoscimento con diverse migliaia di parole nella memoria di un cellulare e utilizzarlo per comporre numeri a voce, cercare nomi e indirizzi e dettare messaggi SMS. Lo stesso può essere fatto con sistemi di sintesi della voce di alta qualità.

Un'evoluzione interessante della tecnologia della riconoscimento della voce per sistemi *embedded* è quella nota come DSR (Distributed Speech Recognition). I sistemi DSR sono basati sulla decomposizione del motore di riconoscimento vocale in due moduli. Il primo modulo, il *front-end*, è dedicato alla collezione dei campioni vocali e alla loro analisi al fine di ridurne la ridondanza ed estrarne le caratteristiche spettrali o *features*. Tipicamente, in un riconoscitore vocale, la voce viene campionata a 64 kb/sec e compressa, mediante analisi spettrale, in un insieme di *features* rappresentabili con un flusso di dati fra i 4 e 8 kb/sec. Il secondo modulo, posto in cascata al *front-end*, è la cosiddetta *search*, che generalmente richiede capacità di calcolo e di memoria molto più elevate rispetto al *front-end*. Mentre il *front-end* può essere comodamente ospitato dalle capacità di calcolo e di memoria dei dispositivi mobili oggi disponibili sul mercato, quali i telefoni cellulari e i computer palmari, il modulo di *search* può risiedere su un *server* remoto con capacità di calcolo e di prestazioni molto più elevate. Questa architettura DSR, standardizzata dal gruppo Aurora[9] dell'ETSI (European Telecommunications Standards Institute), presenta molti vantaggi rispetto all'architettura tradizionale (con il motore di riconoscimento completamente sul *server*), o all'architettura *embedded* pura (con il motore di riconoscimento completamente sul dispositivo cliente). Infatti, il bit rate di trasmissione del segnale vocale è notevolmente ridotto (4 kb/sec invece di 64 kb/sec) e la trasmissione delle *features* permet-

9 http://portal.etsi.org/stq/kta/DSR/dsr.asp

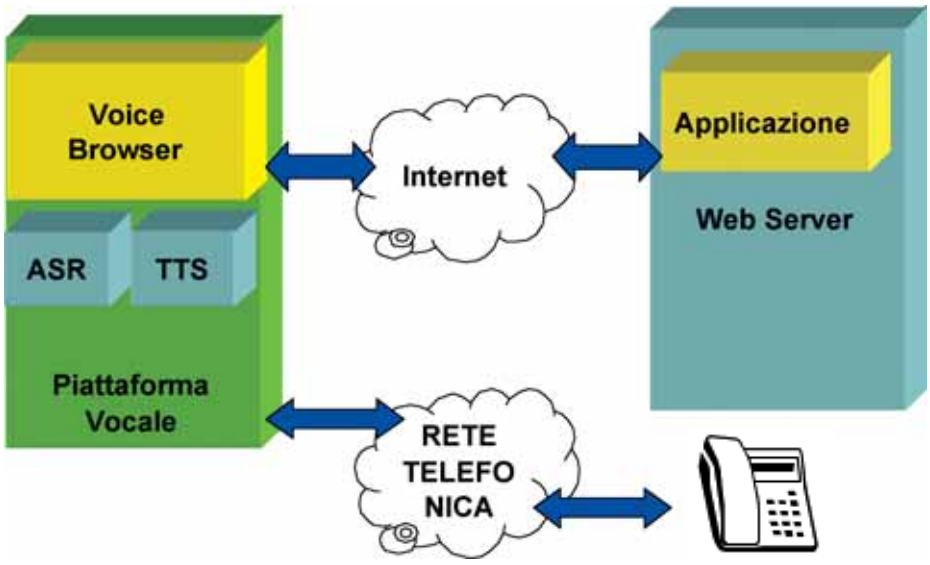

Figura 23. Architettura di un sistema conversazionale standard. Gli utenti si collegano alla piattaforma vocale mediante la rete telefonica tradizionale (oppure mediante VoIP, Voice over IP). La piattaforma vocale gestisce la telefonia di base e le risorse, quali il riconoscitore (ASR = Automatic Speech Recognizer) e il sistema di sintesi (TTS = Text To Speech). All'interno della piattaforma vocale, il riconoscitore e il sintetizzatore vengono controllati dal browser che riceve istruzioni dal programma applicativo sotto forma di documenti VoiceXML. Il programma applicativo risiede su un Web server tradizionale, per esempio come una servlet. Il Web server e il Voice browser comunicano con protocollo http.

te di eliminare completamente il rumore di canale, che costituisce un problema tipico nel riconoscimento vocale con telefonia tradizionale. Inoltre, si può aumentare la banda del segnale vocale al di sopra del limite dei 4 kHz del canale telefonico e quindi aumentare le prestazioni del riconoscitore (per esempio, i suoni con un contenuto spettrale al di sopra dei 4 kHz, come *f* e *s*, vengono spesso confusi su banda telefonica). Infine, si possono implementare sistemi di riduzione del rumore direttamente sul dispositivo cliente, tecnica non possibile con sistemi di riconoscimento completamente ospitati su un *server* remoto.

Il settore dove il riconoscimento della voce sta penetrando maggiormente è quello automobilistico, assieme a quello della telefonia mobile. La maggior parte delle grandi case automobilistiche statunitensi (per esempio, la Chrysler e la Ford), le europee (per esempio, la BMW), e giapponesi (per esempio, la Honda) hanno da qualche anno cominciato ad adottare sistemi di riconoscimento vocale per l'automazione di controlli all'interno dell'auto. Le stesse aziende hanno anche intrapreso progetti di ricerca per l'uso del riconoscimento della voce in attività multimodali avanzate per l'auto del futuro. La sicurezza è la motivazione più importante per il riconoscimento della voce nelle automobili. Si pensi, per esempio, all'uso dei cellulari ed alla possibilità non solo di comporre il numero a voce, ma anche di consultare la lista dei contatti, chiamare un contatto per nome e per luogo (per esempio "autista: telefona a Giovanni Bianchi in ufficio"), disambiguare gli omonimi (per esempio, "autista: telefona a Giovanni; sistema: Giovanni Bianchi o Giovanni Rossi?; autista: Giovanni Rossi; sistema: in ufficio, a casa, o al cellulare?") interagire con il sistema di controllo del veicolo e il sistema di navigazione (per esempio, "autista: guidami al ristorante "da Gino" a Castelnuovo; sistema: non c'è abbastanza benzina nel serbatoio per raggiungere la destinazione, vuoi che ti porti alla stazione di servizio più vicina sul tragitto?; autista: si, però vorrei evitare di prendere l'autostrada sistema: il percorso è di 47 kilometri. Gira a destra al prossimo incrocio…"). Un sistema di riconoscimento della voce ben integrato e con una interfaccia naturale e intuitiva può permettere all'autista di concentrare l'attenzione sulla guida ed evitare di distrarsi.

Sia per i dispositivi mobili come i telefono cellulari o computer palmari, sia per i sistemi di ausilio alla guida, il riconoscimento della voce deve essere integrato e associato a interfacce multimodali per ottenere la massima efficienza dell'interfaccia. I telefoni cellulari e i palmari hanno schermi sempre più sofisticati e ad alta risoluzione, ma l'ingresso dei dati è reso sempre più difficile dalle loro ridottissime dimensioni. Anche se dispositivi come *Blackberry* hanno una piccolissima tastiera QWERTY, il vantaggio di poter interagire sia con la voce sia con l'interazione tipo *touch screen* è evidente.

Come per le auto, l'uso di schermi integrati con il sistema vocale può portare indubbi vantaggi all'interfaccia. Per esempio, può aiutare l'utente a determinare lo stato dell'interazione, esplorare e scoprire i comandi possibili, correggere gli errori di riconoscimento della voce e visualizzare le opzioni disponibili in ogni momento dell'interazione. La multimodalità è dunque una delle frontiere nell'interazione uomo-macchina. L'integrazione avanzata degli input sia vocali sia provenienti da altre modalità di interazione, come quella tattile, o da localizzatori di presenza e sistemi di *eye-tracking* che consentono di determinare dove sta guardando l'utente è uno dei temi di ricerca più avanzati.

Mentre l'utilizzo del riconoscimento della voce per la dettatura (per esempio *Via-Voice* di IBM e *Dragon Dictate* di Scansoft) non è mai riuscito a trovare un volume commerciale sufficiente e si è limitato a piccole nicchie di mercato, come quella della dettatura dei referti medici, l'applicazione degli stessi sistemi a vocabolario molto grande sta incontrando un discreto interesse nel settore dell'automazione dei *call centers*. In questo settore ci si riferisce ad applizioni di cosiddetto *audio-mining*, cioè applicazioni in cui il riconoscimento della voce viene usato per trascrivere conversazioni (per esempio, fra operatore e utente) ed estrarre informazioni, che possono trovare un interesse da parte dell'azienda fornitrice del servizio. Per esempio, *l'audio-mining* può essere utilizzato per la scoperta automatica in tempo quasi rea-

le di situazioni di guasto o di problematiche comuni che sono comunicate agli operatori da diversi utenti anche geograficamente dispersi.

Sempre nel campo dell'*audio mining* è oggi di grande rilevanza l'estrazione di informazioni da lingue diverse. Per effetto degli eventi terroristici di questi ultimi anni, il governo degli Stati Uniti ha finanziato recentemente il progetto di ricerca GALE[10] gestito dalla agenzia DARPA (Defense Advanced Recearch Projects Agency). Gli obiettivi di GALE (Global Autonomous Language Exploitation) sono lo sviluppo e l'applicazione di tecnologie computazionali per l'analisi della voce e del testo in lingue diverse. Il raggiungimento di questo obiettivo potrebbe consentire l'eliminazione, almeno parziale, della necessità di impiegare linguisti e analisti per fornire (come avviene oggi) dati per la difesa militare e civile, in un processo chiamato *distillazione*.

Infine, occorre citare il grosso impegno di alcuni centri di ricerca nel settore dello *speech-to-speech translation*, cioè della traduzione automatica della voce in lingue diverse. Dopo un interesse iniziale agli inizi degli anni '90 e un lungo periodo di stasi, i sistemi di traduzione automatica sono di nuovo oggetto di ricerca di grande interesse.

Conclusioni

La scienza, l'industria e il mercato delle tecnologie vocali hanno avuto un'evoluzione lenta culminata negli anni '80 con l'introduzione di tecniche statistiche, le cui prestazioni sono ancora imbattute. Ci sono voluti altri 20 anni, caratterizzati di un lento miglioramento e una progressiva messa a punto, per portare le prestazioni a livelli utilizzabili nel campo commerciale. In particolare, è stato necessario rendersi conto dell'importanza dell'interfaccia vocale (VUI) come elemento fondamentale per il successo delle applicazioni. Questo è avvenuto alla metà degli anni '90, quando sono comparse le prime aziende che hanno trasformato le tecnologie vocali in successi commerciali. Da allora, l'industria della voce ha cominciato a strutturarsi in livelli *orizzontali* di competenza, con aziende specializzate in settori diversi, quali le tecnologie di base, le piattaforme, gli strumenti e l'integrazione di applicazioni. Contemporaneamente, si è osservata una corrispondente specializzazione *verticale* in applicazioni destinate a settori ben precisi, quali quello bancario, dei trasporti e ospedaliero.

Mentre l'industria della voce è oggi quasi totalmente concentrata sul settore dell'interazione telefonica, stanno emergendo grandi opportunità a livello di sistemi realizzati completamente o parzialmente sui terminali di utente (sistemi *embedded*) con applicazioni nel settore della telefonia mobile e nell'industria automobilistica. Una possibilità interessante è quella dei sistemi distribuiti di tipo DSR.

Infine, la ricerca nelle tecnologie vocali, oltre che a migliorare continuamente le prestazioni del riconoscimento e della sintesi della voce, si sta muovendo verso applicazioni avanzate, come l'estrazione automatica di informazioni da flussi vocali in lingue diverse, la multimodalità e la traduzione automatica.

Roberto Pieraccini Tell-Eureka Corporation

[10] http://www.darpa.mil/ipto/programs/gale/index.htm.

Bibliografia

[1] Davis, K., Biddulph, R., Balashek, S. (1952), "Automatic recognition of spoken digits", J. Acoust. Soc. Amer, V. 24, pp. 637-642.

[2] L. Rabiner, B.-H, Juang (1993), "Fundamentals of Speech Recognition," Prentice Hall.

[3] T. Dutoit (2001), "Introduction to Text-to-Speech Synthesis," Kluwer Academic Publishers.

[4] R. Pieraccini, J. Huerta (2005), "Where do we go from here? Research and commercial spoken dialog systems," Proceedings of 6th SIGDial Workshop on Discourse and Dialog, Lisbon, Portugal, 2-3 September, 2005.

[5] Gorin A. L., Riccardi G., Wright J. H., (1997) "How May I Help You?" Speech Communication, V. 23, pp. 113-127.