ELABORAZIONE AUTOMATICA DEL LINGUAGGIO PARLATO

Fabio Tamburini

Tell me and I forget,
show me and I remember,
involve me and I understand.[1]
- Confucio -

1. INTRODUZIONE

Considerata la difficoltà di definire un impianto teorico/metodologico di riferimento in grado di guidare il docente nel processo di creazione di materiali didattici da fruirsi in modalità e-learning, il paradigma maggiormente adottato, sul quale sono state fatte la maggior parte delle esperienze, resta quello costruttivista. Questo modello si basa su una concezione dell'apprendimento nella quale la partecipazione attiva dello studente, la possibilità di rielaborare la conoscenza in un rapporto paritario, orizzontale nei confronti degli altri partecipanti, e la capacità comunicativa diretta assumono ruoli fondamentali nel processo di apprendimento stesso, lasciando presupporre un'organizzazione dei saperi non gerarchizzata ma distribuita e reticolare. Il fulcro stesso dei vari processi è il discente, che controlla l'intero processo di apprendimento e interagisce con gli attori o gli strumenti a sua disposizione in modo autonomo e responsabile, benché in ambienti, per quanto possibile, regolamentati e controllati.

A livello comunicativo, l'interazione faccia a faccia domina la vita quotidiana di ogni essere umano, in termini di quantità e di qualità, ed è la base fondante

[1] *Se ascolto, dimentico, se vedo, ricordo, se faccio, capisco.*

della didattica tradizionale in aula, sia per quanto riguarda l'interazione docente-discente sia nelle attività collettive. Per sua stessa definizione questa forma comunicativa richiede la compresenza fisica e temporale degli interlocutori, ma consente un flusso di informazioni molto superiore rispetto alle altre forme di comunicazione. In questi ambienti il docente è in grado di progettare e adattare i materiali agli interessi specifici del discente, osservare e analizzare il comportamento dello studente e fornire *feedback* individualizzato e specifico adottando tutta una serie di strategie comunicative in modo da motivare, interessare e focalizzare l'attenzione del discente.

Nella comunicazione faccia a faccia assistiamo allo scambio di molteplici tipi di messaggi, ascrivibili sostanzialmente a due differenti categorie: messaggi *verbali* e *non verbali*. Nei primi riconosciamo gli elementi tipici del linguaggio parlato, ma anche tutti quei messaggi paralinguistici come l'intonazione, la qualità della voce, il ritmo, che contribuiscono marcatamente alla caratterizzazione della comunicazione. Nella seconda categoria raccogliamo tutti quegli elementi cinesici come la mimica facciale, la postura, lo sguardo e i cenni del corpo che completano il messaggio globale e il contenuto informativo di questo tipo di comunicazione.

Questo doppio asse di sviluppo della comunicazione faccia a faccia ha come importante conseguenza che la sua tipologia prevalente sia di tipo *dialogico* e che la quantità di informazione scambiata tra i partecipanti sia massima, rispetto a tutte le altre forme comunicative.

Spostando la nostra attenzione sulla comunicazione mediata dal computer (CMC) riconosciamo una prevalenza netta di due modalità precise: quella testuale e quella grafica. Nella prima si possono distinguere due categorie distinte in base alla compresenza o meno dei partecipanti: avremo quindi da un lato strumenti intrinsecamente *asincroni* come la posta elettronica, i *newsgroup*, le *mailing list* e, in senso più lato, le pagine

HTML statiche, e dall'altro strumenti *sincroni* come le *chat* o i *Multi User Domain*. Nella seconda categoria, riguardante la modalità grafica, riconosciamo perlopiù materiali asincroni come immagini statiche, animazioni o filmati.

Confrontando le due modalità comunicative sembra emergere un'apparente discrasia tra di esse, o meglio, nonostante la comunicazione faccia a faccia sia quella più naturale e predominante nella vita di ogni essere umano, essa sembra essere completamente esclusa, o nettamente sottoutilizzata, all'interno della CMC.

In particolare, nell'ambito della didattica in modalità e-learning si nota ancora una netta predominanza di materiali e procedure fondate su modelli comunicativi simili alla CMC ove le potenzialità della comunicazione faccia a faccia non vengono utilizzate appieno. Le motivazioni sono certamente da ricercare in una gamma di problematiche che hanno ostacolato l'adozione di schemi comunicativi più vicini alla interazione quotidiana tra esseri umani. Certamente la mancanza di opportuni modelli tecnologici che consentissero una corretta gestione delle interazioni dialogiche e le intrinseche difficoltà nella gestione e manipolazione di tali tecnologie hanno fortemente limitato il loro impatto nella progettazione di corsi da fruirsi in formato elettronico.

Il panorama attuale sta però rapidamente cambiando, o meglio è già cambiato, e si assiste a un notevole sviluppo di strumenti tecnologici in grado di supportare, in maniera relativamente semplice e immediata, la definizione di applicazioni specifiche che consentono la creazione di materiali didattici basati su forme di comunicazione tra studenti e computer molto più simili a quelle tra esseri umani (studenti-studenti o studenti-docenti). Il recupero, seppur parziale e simulato, di una comunicazione più vicina all'interazione faccia a faccia, può consentire di alleviare, almeno in parte, il disorientamento cognitivo che ogni studente sperimenta allorché posto in un

contesto didattico che lo vede come unico protagonista del suo processo di apprendimento. I corsi erogati in modalità e-learning, in special modo se non prevedono alcuna forma di tutoraggio, tendono a richiedere un grosso sforzo da parte del discente nel maneggiare, gestire e sistematizzare una grande quantità di informazioni nonché nell'organizzazione e nella pianificazione autonoma del suo processo di apprendimento. La presenza di tutor virtuali può agevolare questo processo creando un collegamento inconscio con una situazione didattica più tradizionale, quando, per ragioni progettuali, viene a mancare il supporto che può fornire un tutor umano.

2. TUTOR INTERATTIVI VIRTUALI E AGENTI INTELLIGENTI

Uno degli strumenti tecnologici che sta ricevendo la maggiore attenzione da parte degli studiosi del settore riguarda senz'altro lo sviluppo di *tutor interattivi virtuali* che consentano un'interazione dialogica col discente e che presentino comportamenti e risposte agli stimoli intelligenti (*agenti intelligenti*). Senza voler divagare citando i miti che caratterizzano questo tipo di impresa, come il calcolatore HAL9000 del famosissimo film '2001 odissea nello spazio', esistono attualmente numerosi progetti di ricerca volti alla costruzione di questo tipo di strumenti. Si vedano per esempio i tutor virtuali sviluppati al KTH di Stoccolma (Granström, 2004), all'Università di Memphis (Graesser *et al.* 2001), al CSLU dell'Università del Colorado (Jiyong, Jie, Ronald 2002) o al CNR di Padova (Cosi *et al.* 2004) per citarne alcuni.

Questo tipo di sistemi risulta essere una complessa e articolata combinazione di moduli che svolgono attività specifiche; la figura 1 mostra uno schema a moduli di un generico tutor virtuale.

Sono presenti sostanzialmente due fasi distinte riguardanti l'elaborazione automatica del linguaggio

parlato: la prima concernente il processo di generazione vocale dei messaggi del tutor sintetico (ramo di sinistra) e la seconda riguardante il riconoscimento dei messaggi prodotti dall'utente che interagisce col sistema (ramo di destra). Queste due fasi, opportunamente interfacciate da una sezione che realizza l'interpretazione semantica e dialogica dei messaggi e crea le opportune risposte del tutor virtuale, implementano il circolo virtuoso riguardante la comunicazione dialogica tra uomo e macchina.

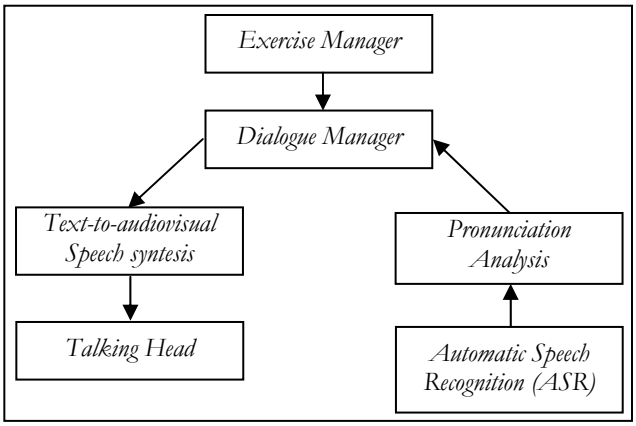

Figura 1: Schema a moduli di un generico tutor virtuale.

Dal punto di vista didattico le esperienze maturate finora nell'uso di tutor virtuali hanno evidenziato aspetti che necessitano di ulteriori fasi di sviluppo, ma anche situazioni nelle quali questi strumenti possono già offrire un valido contributo. Il confronto di questi strumenti automatici con un docente umano li vede ancora evidentemente perdenti, ma vi sono alcune caratteristiche che li rendono particolarmente vantaggiosi:

- possono dedicare al discente un *tempo illimitato* per esercitare ed acquisire le abilità richieste;
- consentono un alto livello di *privacy*, rimuovendo dal processo di apprendimento l'imbarazzo dello studente di fronte a un docente umano;

- nel caso specifico dell'apprendimento della pronuncia di una lingua sono in grado di mostrare particolari articolatori che nessun umano è in grado di riprodurre.

Dato il tema di questo contributo verranno approfonditi gli aspetti principalmente collegati con il riconoscimento e la generazione del messaggio acustico/visuale, tralasciando tutte le problematiche connesse con la gestione simbolica, semantica e dialogica del messaggio, realizzata dai moduli centrali dello schema in figura 1.

Automatic Speech Recognition (ASR)

E' possibile individuare tre grandi classi di applicazioni riguardanti l'ASR: (a) la prima comprende il riconoscimento di parole isolate, circondate quindi da opportune pause; (b) la seconda racchiude tutte quelle problematiche relative ad applicazioni che richiedono il riconoscimento di linguaggio parlato continuo, ma utilizzano vocabolari estremamente ridotti e impongono restrizioni molto rigide sul linguaggio utilizzato (di solito descrivibile precisamente con un modello formale); (c) infine, la terza comprende applicazioni che si basano su dizionari contenenti parecchie decine di migliaia di parole e non pongono alcun vincolo alla complessità del linguaggio utilizzato (*large vocabulary continuous speech recognition system*).

La terza di queste classi affronta il problema nella sua forma più generale e tipicamente richiede lo sviluppo di modelli altamente sofisticati. In questo contesto le difficoltà da affrontare sono molteplici, basti pensare che nel parlato continuo non vi è alcuna indicazione riguardante i confini delle parole (sequenze di parole come "*I scream*" e "*ice cream*" risultano essere esattamente identiche dal punto di vista acustico). Inoltre gli effetti di co-articolazione, generati dalla evidente finitezza dei movimenti articolatori umani,

producono sovrapposizioni di suoni durante le fasi transienti, rendendo ogni suono strettamente dipendente dal contesto. I modelli su cui si fondano tali sistemi risultano quindi estremamente complessi.

In questo contesto non è pensabile di utilizzare la parola come unità minima nella fase del riconoscimento, il modello risultante sarebbe difficilmente gestibile da svariati punti di vista. Si preferisce quindi basare i modelli su cui si fonda l'intero processo su unità segmentali più piccole della parola; la scelta ormai universalmente accettata dagli studiosi del settore ricade quindi sul *fono*, o meglio su sequenze di tre foni – *trifoni* – in modo da poter meglio modellizzare e gestire i fenomeni di co-articolazione.

Attualmente sono due gli approcci modellistici a questo tipo di problemi: metodi che utilizzano *Hidden Markov Model* (*HMM*) – *Sphinx* (Lee, Hon, Reddy, 1990), *HTK* (Woodland *et al.* 1995), *Julius* (Lee, Kawahara, Shikano 2001) – e metodi basati su *reti neurali* o ibridi – *CSLU Toolkit* (Sutton *et al.* 1998). I primi sono generalmente preferiti in quanto richiedono una minor quantità di tempo durante la fase di apprendimento dei parametri utili al riconoscimento; per questo motivo e per ragioni di economia, delineeremo brevemente solo la struttura dei sistemi basati su modelli *HMM*.

La modellizzazione del problema del riconoscimento per quanto riguarda questi sistemi si divide sostanzialmente in due grandi sezioni: il *modello acustico*, costruito propriamente utilizzando reti HMM, che si occupa del riconoscimento della sequenza di foni che compongono l'enunciato esaminato e il *modello della lingua*, solitamente formalizzato con n-grammi (Manning, Schütze 1999), che individua le sequenze di parole (solitamente trigrammi) permesse nella lingua in esame mediante l'analisi di un opportuno *corpus* di testi (si veda per es. Clarkson, Rosenfeld 1997).

Durante la fase di riconoscimento di un enunciato i due modelli vengono fusi per costituire un'unica rete

HMM contenente tutte le possibili ipotesi riguardanti la trascrizione dell'enunciato. Utilizzando opportuni algoritmi di ricerca in questa rete – *Viterbi decoding* (Rabiner 1989) – si individua la sequenza di parole che presenta la probabilità più alta di risultare la trascrizione dell'enunciato in esame.

Pronunciation Analysis

Questo modulo, non sempre presente nelle implementazioni di sistemi reali, affronta il problema dell'identificazione del complesso insieme di fenomeni che esulano dalla semplice trascrizione effettuata dai moduli ASR. Questi fenomeni possono comprendere ad esempio l'analisi di tutta la gamma di sfumature prosodiche – intonazione, prominenza, ritmo, pause, ecc., – fornendo fondamentali informazioni al sistema di interpretazione e gestione del dialogo. Risulta tuttavia opportuno sottolineare che, per la natura e la complessità dei fenomeni, ben pochi sistemi realizzano attualmente elaborazioni di questo tipo.

Un secondo compito ascrivibile a questo modulo potrebbe riguardare l'analisi della corretta pronuncia da parte dell'utente, specialmente se il sistema è inserito in applicazioni che riguardano la didattica delle lingue, ove il controllo della correttezza della pronuncia è uno degli obiettivi dell'intero sistema.

Text-to-audiovisual Speech Syntesis

Speculare al sistema di analisi e riconoscimento dei messaggi vocali provenienti dall'utente, realizzato dai due moduli appena descritti, si trova la catena di moduli progettati per generare i messaggi provenienti dal sistema. I moduli relativi alla gestione del dialogo e del *task* specifico del modulo didattico, che non abbiamo analizzato in questa sede, forniscono gli opportuni messaggi testuali che è necessario comunicare al discente; possono comprendere la correzione di errori,

la descrizione dei compiti e dei passi necessari al completamento del *task* nonché tutta una serie di *feedback* che possiamo definire 'emozionali' che rendono l'interazione un vero e proprio dialogo.

Questo modulo si occupa quindi specificamente di trasformare il messaggio generato dal sistema in opportuni messaggi da veicolare acusticamente (*text-to-speech syntesis*) o visivamente (attraverso *talking heads*). Descriveremo la generazione degli stimoli visuali nella prossima sezione che si occupa precisamente della descrizione di questo compito.

Il problema della generazione di messaggi vocali artificiali ha ricevuto una grande attenzione da parte degli studiosi fin dalla nascita delle scienze informatiche. Inutile sottolineare come sistemi di questo tipo siano di importanza strategica in vari settori, come nelle telecomunicazioni, nella progettazione di interfacce uomo-macchina, nella creazione degli ausili per disabili e, non ultimo, nella didattica in modalità e-learning.

Allo stato dell'arte sono numerosissimi i contributi in questo settore; rimandiamo pertanto all'interessante rassegna di Dutoit (1997) per un quadro più completo della materia.

Un sistema per la generazione del linguaggio parlato è composto essenzialmente da due moduli fondamentali (si veda la figura 2):

- un modulo di elaborazione del testo per l'estrazione di tutte quelle informazioni linguistiche necessarie a una corretta generazione dell'enunciato sintetico;
- un modulo che utilizza le informazioni precedenti per generare un modello acustico dell'enunciato.

Il modulo per l'elaborazione automatica del testo è ulteriormente suddiviso in un numero di blocchi elaborativi che realizzano ciascuno un compito altamente specifico:

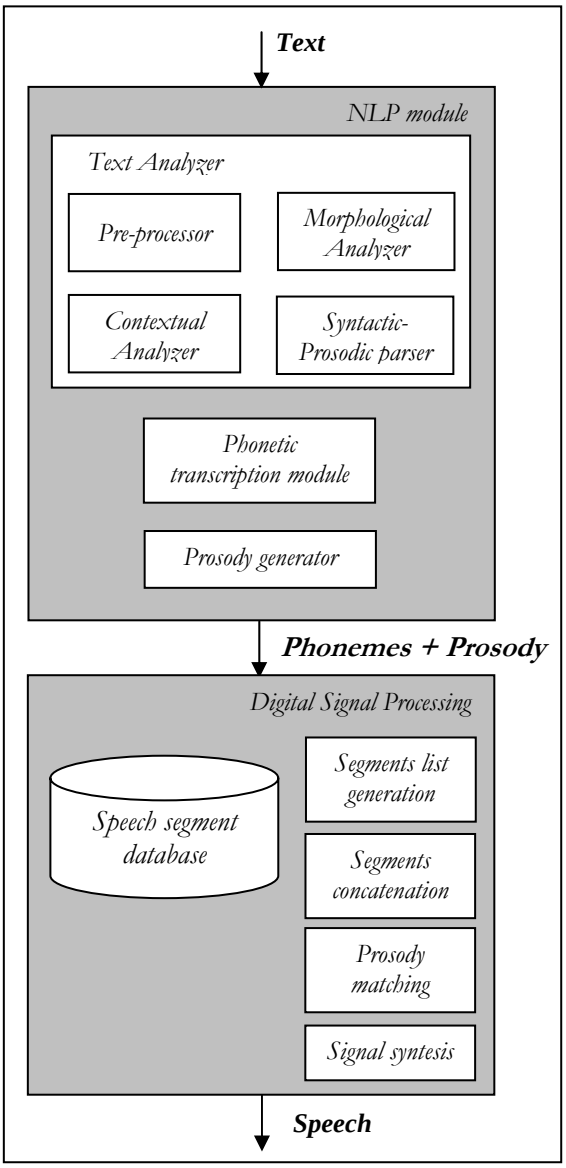

Figura 2: Schema funzionale di un generico sistema per la generazione del linguaggio parlato.

- il blocco di *pre-elaborazione* identifica opportune sequenze di parole isolando date, numeri, abbreviazioni, ecc., e le trasforma, ove necessario, in testo esteso (es. 52 – cinquantadue);
- i due blocchi di analisi morfologica e contestuale si occupano dell'estrazione delle informazioni morfologiche e della disambiguazione delle parti del discorso delle parole per ottenere la corretta pronuncia di ogni termine;
- l'ultimo blocco, riguardante l'analisi del testo, si occupa dell'analisi sintattica della frase per ottenere un'ipotesi di informazioni prosodiche da sovra-imporre all'enunciato. I fenomeni prosodici del linguaggio parlato dipendono da fattori sintattici, ma soprattutto dalla semantica della frase e da fattori legati al livello pragmatico. Un'analisi approfondita di questi ultimi due livelli risulta essere estremamente complessa, e viene tipicamente non trattata in sistemi di questo tipo;
- sulla base delle informazioni ricavate nelle elaborazioni precedenti due ulteriori moduli si occupano rispettivamente della trascrizione fonetica del testo da generare – tipicamente utilizzando metodi basati su regole – e dell'identificazione delle posizioni ove inserire fenomeni prosodici specifici – *pitch accent* o *stress* frasali – o della intonazione globale che è opportuno far assumere all'enunciato.

Il secondo dei due moduli in figura 2, che si occupa dell'effettiva generazione dell'enunciato, può essere basato su svariati modelli di riferimento. Attualmente si tende ad utilizzare modelli *concatenativi* in grado di costruire un intero enunciato combinando opportunamente segmenti più piccoli. I segmenti che vengono tipicamente utilizzati in sistemi concatenativi comprendono *difoni* – che modellizzano la transizione tra due foni consecutivi – e *trifoni* – che modellizzano due transizioni immediatamente precedenti e seguenti ad un fono. Con opportuni algoritmi, basati su

complesse fasi di apprendimento, si individuano i parametri caratteristici dei segmenti selezionati utilizzando *corpora* di linguaggio parlato disponibili, accuratamente segmentati ed etichettati rispetto alla loro composizione fonetica. L'opportuna concatenazione di questi segmenti, unita a procedure di alterazione del segnale – in termini di durata temporale di intensità e di cambiamenti delle componenti in frequenza – per realizzare i fenomeni prosodici individuati nelle fasi precedenti, termina il processo di generazione dell'enunciato a partire dalla sua trascrizione testuale.

Sistemi molto noti per l'implementazione parziale – per es. *MBROLA* (Dutoit *et al.* 1996) – o totale – per es. *Festival* (Taylor *et al.* 1998) – delle procedure descritte sono disponibili sotto forma di *open-source* software, e possono essere adattati a specifici utilizzi e a particolari lingue attraverso opportune procedure di *training*.

Talking Heads

Nella comunicazione faccia a faccia il canale di comunicazione visuale assume un'importanza fondamentale; la possibilità di vedere il viso della persona con cui si sta dialogando, oltre a veicolare un gran numero di informazioni extralinguistiche come emozioni, enfasi e di supporto al dialogo, migliora notevolmente la comprensione del messaggio parlato, specialmente in condizioni ambientali avverse per la presenza di disturbi. Questo fa dello sviluppo di strumenti visuali per la sintesi di visi virtuali il naturale complemento di ogni sistema per la generazione di linguaggio parlato (Beskow 1995) e un importante supporto per lo sviluppo di materiali didattici da fruirsi in modalità e-learning (Cosi, Magno Caldognetto, 2004).

Il modello facciale consiste tipicamente di una griglia tridimensionale che forma la superficie del viso;

per la creazione di questa griglia è possibile procedere disegnandola artisticamente o rilevando le posizioni di opportuni punti di riferimento direttamente da un viso umano (si veda il modello di viso sintetico mostrato in figura 3). Quest'ultima modalità, adottata dalla maggior parte degli studi nel settore, consente inoltre un'approfondita analisi dei movimenti facciali umani e una conseguente riproduzione fedele dei movimenti della faccia sintetica attraverso la modulazione di decine di parametri che contribuiscono a definire la forma e il comportamento dinamico dei lineamenti (Beskow *et al.* 2003).

Per una corretta riproduzione dei messaggi facciali sono di grande importanza i modelli tridimensionali della lingua, della mandibola e delle labbra del viso sintetico: è infatti il movimento di questi strumenti articolatori primari che veicola la maggior parte delle informazioni visive legate alla produzione del parlato. La figura 3 mostra come i tre modelli tridimensionali sono integrati per formare un insieme capace di fornire il corretto completamento visivo delle informazioni fornite acusticamente.

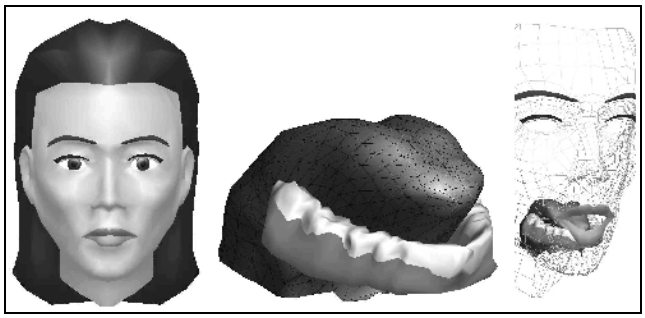

Figura 3: Il modello del viso (sinistra), della lingua (al centro) e la combinazione dei due (destra) nella faccia parlante *SynFace* (riprodotto con autorizzazione da Beskow *et al.* 2003)

La ricchezza dei parametri del modello tridimensionale (movimenti della lingua, delle labbra, della mandibola, degli occhi, delle guance, delle sopracciglia, ecc.) consente, attraverso complessi modelli della dinamica facciale, la creazione di una gamma enorme di movimenti ed espressioni. Ad ogni fono della lingua considerata viene associato uno specifico *visema* che rappresenta i movimenti necessari all'articolazione del suono; successivi visemi vengono concatenati uno di seguito all'altro per produrre la sequenza di movimenti necessari alla riproduzione di un enunciato e, sovrimposti ad essi, le espressioni facciali e le emozioni, strettamente legate al livello prosodico, vengono riprodotte dalle altre parti del viso sintetico. Movimenti spontanei degli occhi e del viso, come il battito delle palpebre, vengono aggiunti in modo casuale alla testa artificiale, specialmente durante le fasi di silenzio, per rendere ancor più realistico il risultato finale.

LE SFIDE DEL FUTURO

Volare sulle ali dell'immaginazione seguendo fino all'orizzonte le linee che le nuove tecnologie tracciano ogni giorno porta a immaginare scenari fantascientifici ove l'interazione uomo-macchina raggiunge livelli tali da considerare plausibile la costruzione di "compagni artificiali" (*Artificial Companions* come li definisce Wilks, 2004), macchine in grado di interagire con gli esseri umani in modo così performante da essere considerate veri e propri compagni per la vita, per anziani, bambini o semplicemente per memorizzare l'intera quantità di informazioni di una vita completa (si pensi al progetto "*Memories for Life*" di cui si parla tanto nel Regno Unito).

Anche se proiezioni immaginarie di questo tipo non rappresentano certamente lo stato dell'arte della tecnologia odierna, non bisogna nemmeno credere che simili imprese siano completamente al di fuori della portata delle moderne tecnologie; giocattoli come il

piccolo *Tamagochi*, l'animaletto *Furby* e i recenti cuccioli robot della Sony, hanno tracciato un percorso preciso che mostra come l'intelligenza artificiale e la robotica possano, in un futuro non necessariamente remoto raggiungere asintoticamente quegli obiettivi che ora ci sembrano assolutamente lontani o irrealizzabili.

Abbiamo già descritto numerosi progetti volti alla creazione di ambienti di apprendimento mediati da tutor artificiali in grado di realizzare modelli comunicativi molto più simili a quelli propri della didattica tradizionale, capaci cioè di trarre i massimi vantaggi dall'ibridazione delle due modalità comunicative (faccia a faccia *vs* CMC) descritte in precedenza. Questo consente la creazione di ambienti di apprendimento nei quali il disorientamento cognitivo del discente viene ridotto proprio grazie ad un recupero di modalità comunicative e di interazioni più naturali; si consideri ad esempio il livello di naturalezza dell'interazione dialogica con un sistema automatico in confronto all'uso di interfacce "manuali" quali la tastiera o il mouse.

Il concetto stesso di attività di rinforzo all'interno di modelli "*learning by doing*" viene completamente ridefinito dalle possibilità fornite dagli ambienti immersivi, in grado di proiettare il discente in attività di simulazione capaci di coinvolgere molti flussi comunicativi. Un esempio emblematico di questo tipo di studi è il progetto TLTS (*Tactical Language Training System*) avviato dall'Università della California del Sud in collaborazione con l'Accademia Militare degli Stati Uniti (Johnson *et al.* 2004), per lo sviluppo di un sistema immersivo per l'apprendimento della lingua e della cultura di un paese straniero mediante simulazioni di attività reali. L'interazione tra il sistema e il discente è fondamentalmente dialogica, ma sono previste altre modalità comunicative paralinguistiche come le espressioni facciali e la postura e i movimenti del corpo (non necessariamente in entrambe le direzioni comuicative).

Di estremo interesse appare anche lo studio condotto nell'ambito del progetto ARTUR (*ARticulation TUtoR*) rivolto alla progettazione e implementazione di un tutor virtuale in grado di analizzare, attraverso opportuni dispositivi di ripresa e metodologie di elaborazione di materiali audiovisivi, i movimenti articolatori del discente al fine di una analisi della corretta pronuncia di una lingua (Engwall *et al.* 2004).

C'è ancora molta strada da percorrere per il raggiungimento dell'obiettivo della creazione di tutor virtuali capaci di gestire correttamente informazioni comunicative complesse, ma la strada è ormai tracciata e il futuro, per questo tipo di strumenti didattici, è quantomai prossimo.

BIBLIOGRAFIA

Beskow J. 1995, "Rule-based Visual Speech Syntesis", *Proc. of Eurospeech '95 Conference*, pp. 299-302.

Beskow J., Engwall O., Granström B. 2003, "Resyntesis of Facial and Intraoral Articulation from Simultaneous Measurements", *Proc. of ICPhS 2003 Conference*, pp. 57-60.

Clarkson P., Rosenfeld R. 1997, "Statistical Language Modeling using the CMU-Cambridge toolkit", *Proc. of Eurospeech '97 Conference*, pp. 2707-2710.

Cosi P. et al. 2004, "Italian literacy tutor, tools and technologies for individuals with cognitive disabilities", *Proc. of InSTIL/ICALL Symposium on Computer Assisted Language Learning*, Venezia, pp. 207-215.

Cosi P., Magno Caldognetto E. 2004, "E-learning e facce parlanti: nuove applicazioni e prospettive" , *Atti delle XIV Giornate del GFS*, pp. 247-252.

Engwall O. et al. 2004, "Design strategies for a virtual language tutor", *Proc. of ICSLP 2004 Conference*, pp. 1693-1696.

Dutoit T. et al. 1996, "The MBROLA Project: Towards a Set of High-Quality Speech Synthesizers Free of

Use for Non-Commercial Purposes", *Proc. of ICSLP '96 Conference*, pp. 1393-1396.

Dutoit T. 1997, "High-Quality Text-to-Speech Synthesis: an Overview", *Journal of Electrical & Electronics Engineering*, 17, pp. 25-37.

Graesser A.C. et al. 2001, "Teaching tactics and dialog in AutoTutor", *International Journal of Artificial Intelligence in Education*, 12, pp 257-279.

Granström B. 2004, "Towards a virtual language tutor", *Proc. of InSTIL/ICALL Symposium on Computer Assisted Language Learning*, Venezia, pp. 1-8.

Jelinek F. 1999, *Statistical Methods for Speech Recognition*, Cambridge (MA): MIT Press.

Jiyong M., Jie Y., Ronald C. 2002, "CU animate tools for enabling conversations with animated characters", *Proc. of ICSLP 2002 Conference*, pp. 197-200.

Johnson W.L. et al. 2004. « Tactical Language Training System: Supporting the Rapid Acquisition of Foreign Language and Cultural Skills", *Proc. of InSTIL/ICALL Symposium on Computer Assisted Language Learning*, Venezia, pp. 21-24.

Lee A., Kawahara T., Shikano K. 2001, "Julius - an open source real-time large vocabulary recognition engine", *Proc. Eurospeech 2001 Conference*, pp. 1691-1694.

Lee K., Hon H., Reddy R. 1990, "An Overview of the SPHINX Speech Recognition System", *IEEE Transaction on Acoustics, Speech and Signal Processing*, 38, pp. 35-45.

Manning C.D., Schütze H. 1999, *Foundations of statistical natural language processing*, Cambridge (MA): MIT Press.

Rabiner L.R. 1989. "A tutorial on Hidden Markov Models and selected applications in speech recognition", *Proc. of the IEEE*, 77, pp. 257-286.

Rabiner L.R., Juang B.H. 1993, *Fundamentals of Speech Recognition. Englewood Cliffs*, New Jersey: Prentice Hall.

Roversi A. 2004, *Introduzione alla comunicazione mediata dal computer*, Bologna: Il Mulino.

Sutton S. et al. 1998, "Universal Speech Tools: the CSLU Toolkit", *Proc. of ICSLP '98 Conference*, pp. 3221-3224.

Taylor P.A., Black A., Caley R. 1998, "The architecture of the festival speech synthesis system", *Proc of the Third ESCA Workshop in Speech Synthesis*, pp. 147-151.

Wilks Y. 2004, "Artifical Companions", *Proc. of InSTIL/ICALL Symposium on Computer Assisted Language Learning*, Venezia, pp. 155-159.

Woodland P.C. et al. 1995, "The HTK Large Vocabulary Continuous Speech Recognition System: An Overview", *Proc. of ARPA Spoken Language System Technology Workshop*, Austin (TX), pp. 104-109.