Roberto Pieraccini. The Voice in the Machine

Paolo Baggia
24 Luglio 2012

Che cosa accomuna Wolfang von Kempelen, alla corte di Maria Teresa d'Austria, i film di culto di sci-fi come *2001: Odissea nello Spazio* o *Star Wars* e il lavoro di centinaia di ricercatori nei laboratori in tutto il mondo? Il tentativo di dare voce a una macchina, di permetterle di comunicare con noi. Dalla macchina parlante di fine '700, ad HAL 9000, fino alle applicazioni odierne delle tecnologie vocali, più propriamente il riconoscimento della voce, la sintesi vocale, le biometriche vocali ed altro ancora. Se volete entrare in questo mondo affascinante, potete leggere il nuovo libro di Roberto Pieraccini, *The Voice in the Machine: Building Computers that Understand Speech*, appena uscito per MIT Press in inglese, non ancora disponibile in italiano.

Roberto Pieraccini, viareggino, diviene ingegnere a Pisa, per poi spostarsi allo CSELT (Centro Studi E Laboratori Telecomunicazioni) di Torino, culla delle prime ricerche italiane sulle tecnologie vocali. Siamo nei primi anni '80 e lo CSELT è un luogo unico per innamorarsi di una settore di ricerca così particolare. Passano gli anni, alla soglia dei '90, Pieraccini trascorre un anno ai Bell Labs in New Jersey dove decide di fermarsi. Dopoi 10 anni si sposta prima in SpeechWorks (un precursore della attuale Nuance), poi al centro ricerche di IBM a Yorktown Heights, infine quale CTO (direttore tecnico) in SpeechCycle, una start-up con sede davanti al toro di Lower Manhattan. Da ultimo è direttore e presidente all'ICSI (International Computer Science Institute) di Berkeley in California. Il libro descrive la nascita, i successi e le speranze legati alle tecnologie vocali, illustrando le tappe principali della loro evoluzione.

Il primo capitolo descrive il paradosso di capacità così naturali per l'uomo, usate apparentemente senza sforzo, ma molto complesse e difficili da replicare in modo automatico in un computer. Illustra come le idee sono strutturate in concetti e parole e poi articolate in un'onda sonora che giunta all'orecchio di un ascoltatore

viene ricostruita nel suo significato. Semplice a parole, ma difficile da riprodurre. I tentativi di fine ‘700 e ‘800 sono sulla scia degli automi, cercano di creare una macchina che possa parlare. Nel ‘900, l’elettricità prima, l’elettronica poi, ed infine i computer saranno gli strumenti per continuare la ricerca.

Il libro prosegue con una carrellata dei principali attori degli albori e dei loro successi. Nel dopoguerra la nascita del computer attrae la ricerca soprattutto negli Stati Uniti, ma poi anche in Europa e nel resto del mondo. In quegli anni nasce l’Intelligenza Artificiale (AI) da un piccolo seminario di ricerca a Dartmouth nel 1956, organizzato da John McCarthy a cui hanno partecipano Shannon, il padre della teoria dell’informazione, Allen Newell, Herbert Simon e altri. La nuova disciplina cerca di indagare il ragionamento umano e produrre macchine capaci di fare inferenze, come i sistemi esperti, di apprendere o competere con l’uomo in giochi complessi come gli scacchi. Anche la lingua e la voce sono un campo di applicazione. L’idea è di codificare in regole il maggior numero di conoscenze così come sembra fare il cervello umano. Nascono dalle speranze di quel periodo le macchine parlanti e pensanti come HAL 9000 del film di Kubrick.

Peccato che la voce e il linguaggio resistano a questo approccio. Gli eleganti sistemi a regole, sempre più complessi e raffinati, non riescono a garantire risultati concreti ed affidabili. Dagli anni ‘70 si vedrà un ritorno alla probabilità ed alla statistica, tramite approcci denominati di *forza bruta,* cioè basati quasi unicamente sulla memoria e la potenza di calcolo, invece di raffinati modelli cognitivi. Il riconoscimento della voce sarà riformulato in modo statistico facendo uso di catene di Markov, gli “hidden Markov models” (HMM) divenuti da allora lo strumento vincente. Un’altra limitazione del periodo precedente era stata la mancanza dei meccanismi di valutazione che permettessero ai ricercatori di confrontare i risultati e riprodurre i successi altrui. Inoltre le limitate capacità di calcolo non permettevano di sfruttare le potenzialità di grandi database vocali per promuovere la ricerca. Di questo periodo la frase storica di Frederick Jelinek, uno dei fautori dei metodi statistici in IBM: “Ogni volta che licenzio un linguista, le prestazioni del riconoscimento aumentano”. Sintomo di una frattura tra tecniche statistiche rifiutate da altri ricercatori, questa frattura sarà sanata negli anni successivi ibridando le tecniche diverse.

In questo periodo in USA il DARPA (Defense Advanced Research Project Agency) finanzia progetti di ricerca sulla voce basati su competizioni affidate al NIST (National Institute of Standards and Technology), il quale fornisce grandi mole di file audio per addestramento e poi organizza una vera e propria gara con materiale ignoto e poco tempo per fornire i risultati da confrontare con gli altri partecipanti. In queste competizioni sono presenti tutti i maggiori centri di ricerca statunitensi e poi via via anche alcuni centri europei. Sull'altro lato dell'oceano la Comunità Europea finanzia progetti di ricerca che spingono le università a collaborare con l'industria anche nel settore vocale. Forse il meccanismo americano è più efficace perché permette un confronto alla pari dove via via le tecnologie si affinano. Ed ecco passare il compito del riconoscimento da poche parole lette in ambiente silenzioso, a frasi, con vocabolario che cresce da decine di parole a migliaia, fino a quasi centomila. Anche l'ambiente, da silenzioso e controllato, passa ad essere rumoroso come gli ambienti quotidiani. Infine da un riconoscimento su singolo parlatore, si passa a un riconoscimento indipendente dal parlatore. Solo i meccanismi statistici riescono, con potenze di calcolo sempre maggiori, ad adattarsi alle nuove condizioni e a produrre risultati sempre migliori.

I capitoli centrali del libro sono di rara maestria, senza formule, con esempi semplici sono illustrati i fondamenti che stanno alla base di tutte le tecniche utilizzate nei sistemi di riconoscimento. Il panorama si arricchisce di dettagli e di questioni che permettono al lettore di entrare, anche se superficialmente, nel cuore dei problemi che sono affrontati dalla ricerca.

Giungiamo alla fine degli anni '90, le prestazioni sono aumentate in modo marcato e le potenze di calcolo permettono su normali PC di eseguire dei compiti proibitivi negli anni precedenti. Ecco allora come una fresca ventata che ha scosso tutti i maggiori centri di ricerca mondiali. I tempi erano maturi per dare alla luce quella che sarà chiamata di lì a poco la *speech industry*, l'industria della voce. Sulla costa ovest degli Stati Uniti, lo SRI (Stanford Research Institute) genera Nuance Communications, sulla costa est l'MIT di Boston e AT&T forniscono la tecnologia a SpeechWorks queste aziende si contendono il mercato americano. In Europa lo stesso, ma in forma meno marcata, forse il maggior successo è italiano: lo CSELT nel 2001 dà i natali a Loquendo che presto diventerà una delle prime aziende mondiali di questo settore, a cui ho preso parte anch'io.

Sempre in USA, un imprenditore cresciuto presso lo XEROX PARC, Paul Ricci, diviene CEO di un’azienda di SW per scanner e riconoscimento di caratteri (OCR) la ScanSoft di Boston. Quest’azienda in pochi anni diventa un gigante, acquisendo decine di altre aziende. Come primo passo rileva parte del fallimento della belga Lernaut & Hauspie, poi acquisisce il settore riconoscimento della Philips di Aachen, poi SpeechWorks, per poi fondersi con Nuance e mantenerne il nome. Anche Loquendo, nell’agosto del 2011 è venduta da Telecom Italia a Nuance ed io con lei.

Pieraccini descrive questo periodo soffermandosi principalmente sui tipi di applicazioni realizzate, sui sistemi di dialogo che passano dall’automazione di semplici a menù, ad applicazioni informative, come la richiesta di orari dei treni o dei voli, in seguito ad applicazioni di tipo transazionale, quali le applicazioni nei settori bancari o finanziari e infine le applicazioni di *problem solving* che aiutano un utente a risolvere un problema ad esempio con il modem e la tv via cavo. Le tecnologie vocali uscite dai laboratori sono diventate parte della nostra vita.

Il libro termina dopo aver illustrato in modo magistrale questi cinquant’anni di ricerca e poi di utilizzo dei risultati migliori con la domanda su quale sarà il futuro prossimo, se si giungerà ad HAL 9000. Probabilmente no, perlomeno nei prossimi anni; viceversa le tecnologie vocali saranno sempre più presenti nella nostra vita, dall’automobile, dove serviranno a ridurre i rischi per il guidatore nel controllare l’abitacolo, ma allo stesso tempo a ricevere ed inviare SMS o e-mail e comunicare con il mondo esterno. Anche in casa si assisterà alla diffusione di apparecchiature che oltre ai telecomandi potranno ricevere comandi vocali, “registra NCIS domani alle 21 su RAI 2”. E poi gli smartphone; in Italia si attenderà fino ad ottobre per poter avere su un iPhone l’assistente virtuale Siri, già disponibile dallo scorso autunno in USA. Siri, nato da uno spin-off dello SRI, è stato acquisito da Apple nel 2010 e lanciato come applicazione pre-installata su iPhone 4S nell’autunno 2011. Presto, la risposta di Google e degli altri operatori.

Le tecnologie vocali sono tra noi, e il libro di Pieraccini aiuta a capire come funzionano, i loro limiti e le loro potenzialità.

rszd-2.jpg
cover.jpeg

Se continuiamo a tenere vivo questo spazio è grazie a te. Anche un solo euro per noi significa molto. Torna presto a leggerci e SOSTIENI DOPPIOZERO