## INTERAZIONE VOCALE UOMO-COMPUTER

**Il parlato umano.** È composto da "fonemi", attraverso i quali si possono costruire tutte le parole di una data lingua per mezzo di specifiche regole. Un fonema può essere individuato solo se analizzato nel contesto della parola; la concatenazione di più fonemi dà origine al fenomeno della "coarticolazione", cioè della alterazione introdotta dal fonema che precede e da quello che segue. Il significato di una frase dipende anche dalla "prosodia", cioè dall'intonazione, dalle variazioni di timbro, dall'accentuazione di particolari sillabe, dalla durata dei fonemi.

**Sintesi artificiale del parlato.** *Sintesi per registrazione.* Le parole pronunciate da un operatore vengono digitalizzate e inserite nel vocabolario in memoria del computer; risposte articolate vengono costruite accedendo ripetutamente a tale archivio e concatenando le varie parole. Infine per conferire capacità prosodiche alla voce si memorizza più volte la stessa parola con diverse intonazioni.

*Sintesi attraverso segmenti.* Si memorizzano solo i cosiddetti "difoni", cioè le coppie di fonemi. In questo modo si inglobano nei suoni elementari tutte le possibili varianti indotte dal passaggio da un fonema al successivo.

*Sintesi attraverso regole.* Si ricavano da un suono le zone di frequenza più energetiche dette "formanti" e si individuano alcuni parametri fondamentali che, successivamente, permettono di ricostruire con buona approssimazione lo spettro originario. La tecnica utilizzata è quella predittiva lineare che prevede: a) la scelta fra una sorgente di tipo periodico (per suoni vocalizzati) e una sorgente casuale (per suoni occlusivi e fricativi); b) un filtro numerico con guadagno e caratteristiche impostabili in funzione dei suoni da produrre; c) un convertitore digitale/analogico che produce il segnale da inviare all'altoparlante. Tale metodologia necessita di una quantità di memoria inferiore a quella delle precedenti tecniche ma è necessario un sistema con maggiore potenza di calcolo. In figura A è indicato lo schema di principio per un sintetizzatore di voce partendo da un testo scritto.

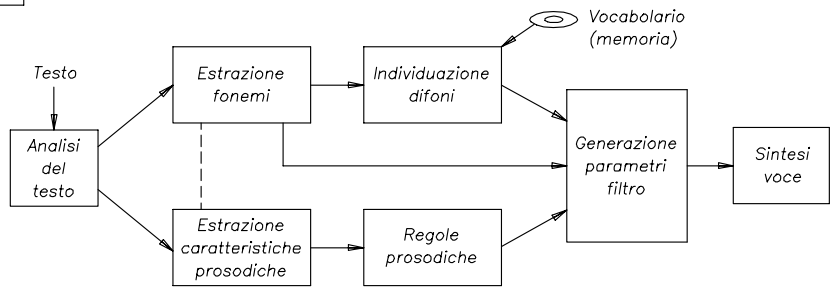
**Riconoscimento della voce.** Viene anche indicato con l'acronimo ASR (Automatic Speech Recognition). È un processo più complesso rispetto alla sintesi in quanto la pronuncia di una parola varia da persona a persona. Inoltre anche lo stesso soggetto introduce modifiche nel timbro e nell'intonazione. I sistemi di riconoscimento della voce possono essere di tipo dipendente o indipendente dal parlatore; nel secondo caso si è in presenza di macchine più flessibili che per contro presentano un inferiore grado di accuratezza (percentuale media di parole riconosciute). Il processo di riconoscimento prevede (fig. B): la digitalizzazione del segnale microfonico; l'individuazione delle singole parole; l'analisi delle relative caratteristiche (spettro, energia, durata ecc.). Ottenuto l'insieme dei parametri identificativi dei vari fonemi in una data parola, si passa all'inserimento in archivio se si è nella fase di apprendimento, oppure al confronto con quelli già in memoria se si è in fase di riconoscimento. È in genere necessario anche un processo di allineamento, in quanto le parole possono essere pronunciate con fonemi di durata più o meno lunga.

Il riconoscimento completo del parlato continuo richiede infine la separazione delle varie parole e l'analisi lessicale, sintattica e semantica del testo. A titolo esemplificativo in figura C vengono confrontate le caratteristiche di alcuni sistemi commerciali di riconoscimento vocale.

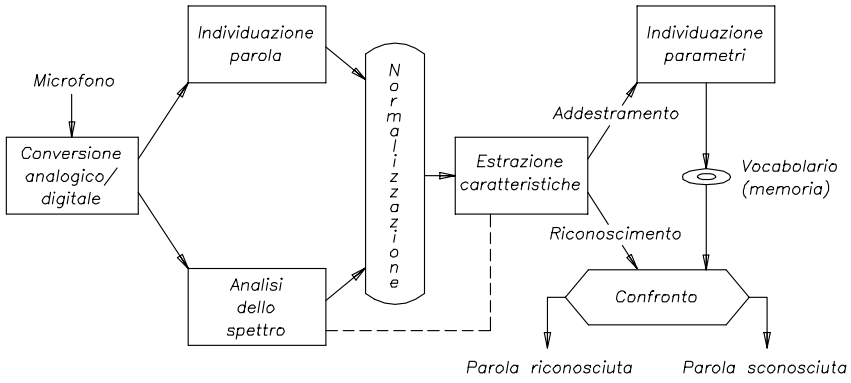
I costi variano da alcune centinaia a qualche migliaia di €, comprensivi delle schede HW per l'interfacciamento con il computer.



A *Sintetizzatore di voce*




B *Riconoscimento delle parole*




C *Confronto fra alcuni sistemi di riconoscimento vocale*

| Sistema | Casa costruttrice | N.° parole | Percentuale riconoscimento | Tipo di riconoscimento |
|---|---|---|---|---|
| Phonetic nengine | Speech system | >20.000 | 90% | – dipendente<br>– parlato continuo |
| Introvoice V | Voice connection | >400 | 98% | – dipendente<br>– parlato continuo |
| SAR–10 | NEC America | >250 | 98% | – dipendente<br>– parole singole |
| Voice command | IBM | >100 | 98% | – dipendente<br>– parole singole |
| SSB–1000 | AUDEC | >100 | 95% | – dipendente<br>– parole singole |
| VC–1000 | Voice control system | >50 | 99% | – indipendente<br>– parlato continuo |

*Dario Abastanotti*