Riconoscimento della voce (II)

L'obiettivo principale della ricerca sul riconoscimento della voce è la creazione di sistemi che possano interpretare e rappresentare diverse voci parlando in modo naturale.

Questo è molto difficile, dato che la tecnologia attuale non risolve ancora le difficoltà che sorgono a causa della variabilità delle caratteristiche dei segnali acustici. Per questa ragione i dispositivi per il riconoscimento della voce sono classificati in funzione delle diverse restrizioni che presentano.

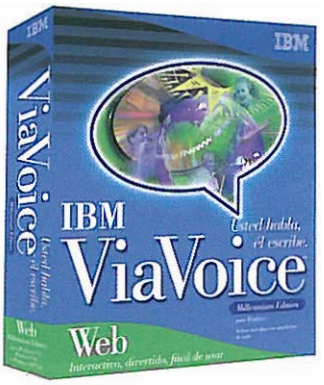
Dipendenza-Indipendenza dal soggetto che parla

Un sistema dipendente dal soggetto che parla realizza un maggior numero di riconoscimenti corretti, dato che in questo tipo di tecnica si lavora sempre con una persona specifica. I campioni che vengono utilizzati per il riconoscimento appartengono sempre allo stesso individuo e questo rende più facile il riconoscimento del vocabolario.

Quando parliamo di una applicazione indipendente, si tratta di un dispositivo capace di riconoscere la voce di qualsiasi persona indipendentemente se questa sia o non sia inclusa nel database dei dati durante il processo preliminare di apprendimento. Questo comporta un decremento delle probabilità di riconoscimento, in quanto le caratteristiche della voce di ogni persona che interviene possono essere completamente diverse.

Parole isolate - parlata continua

Chiamiamo parlata continua il naturale modo di espressione degli esseri umani che pronunciano le parole e realizzano le pause necessarie fra di esse. Il problema risiede nel fatto che non è facile identificare i limiti delle parole (inizio e fine), inoltre le parole possono essere pronunciate in una sequenza tale da essere confuse con una sola. Tuttavia, quando si tratta di riconoscere vocaboli isolati, chi parla lo fa più lentamente, e questo dà come risultato una probabilità di riconoscimento maggiore.

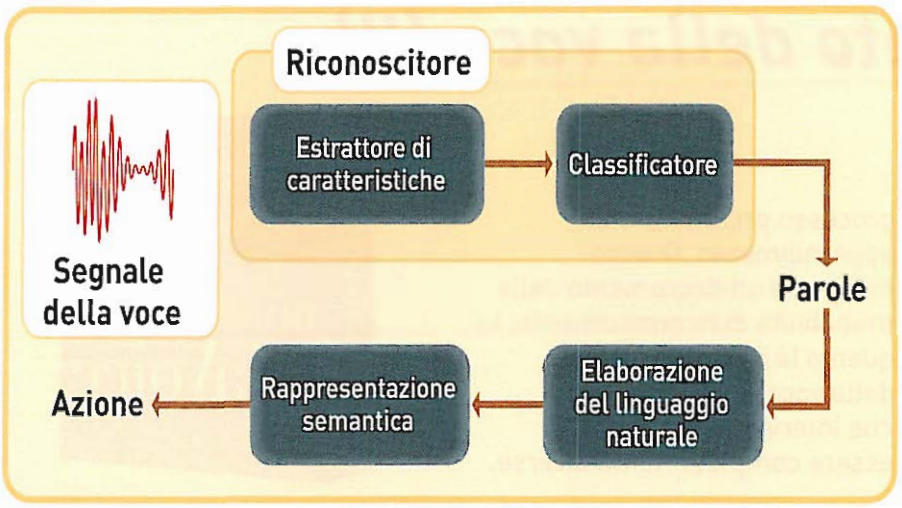


Attualmente esistono vocabolari con di più di 100.000 parole, come quello che incorpora il Via Voice di IBM.

Dimensione del vocabolario

Un altro fattore di grande importanza è la dimensione del vocabolario, dato che man mano

I sistemi dipendenti dalla persona che parla si utilizzano molto nei giochi per bambini.



Architettura di un sistema di riconoscimento della voce.

che cresce, aumenterà anche il tempo impiegato per realizzare un riconoscimento, in quanto il processo di ricerca sarà più lungo e sarà più alta anche la probabilità di errore causata dalla confusione di una parola con un'altra simile.

Variabilità e rumore

È stato verificato che per ottenere una buona percentuale di riconoscimento, è necessario tenere conto dei fattori a cui è esposto il dispositivo di acquisizione della voce, dato che il rumore prodotto nell'ambiente può far degradare in modo significativo la qualità del sistema.

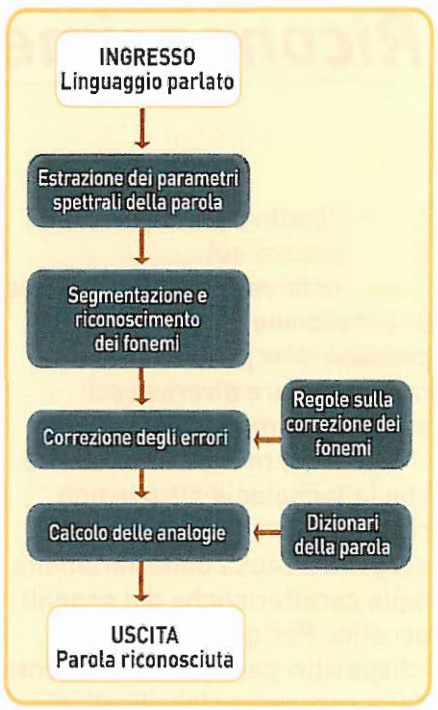
Architettura di un sistema di riconoscimento della voce

Per capire come funziona un'applicazione di riconoscimento della voce è necessario riconoscere quali sono i suoi due principali componenti: l'estrattore di caratteristiche e il classificatore. Il processo di riconoscimento della voce di una persona è composto da due fasi. La prima è quella dell'estrazione delle caratteristiche acustiche di una voce, il segnale viene diviso in un insieme di segmenti per poterne distinguere le caratteristiche significative. In seguito, nella seconda fase, che è la classificazione probabilistica, si crea un modello utilizzando alcune delle diverse tecniche esistenti.

Dopo di che si realizza una ricerca per trovare la sequenza dei segmenti con maggior probabilità di essere riconosciuta.

Classificazione dei sistemi di riconoscimento secondo la loro architettura

Esistono due principali tipi di architettura all'interno dell'applicazione di riconoscimento della voce. Il primo, chiamato sistema ad architettura integrata, incorpora tutte le sorgenti di riconoscimento che un'applicazione di questo genere può avere:



Schema del funzionamento di un sistema di riconoscimento della voce.

grammatica, dizionari, modulo acustico, ecc. Questo metodo utilizza un grande quantitativo di memoria, però è in grado di generare il modello acustico durante il processo di riconoscimento aumentando le possibilità di arrivare a un risultato soddisfacente. L'altra alternativa esistente è il sistema ad architettura non integrata, che divide il processo in due stadi, generando nel primo una mappa con le differenti pronunce di un fonema, in base al contesto in cui è pronunciato (mappa degli allofoni). Nella seconda fase un algoritmo di programmazione dinamica, capace di utilizzare informazioni riguardanti il lessico e la grammatica, genererà la frase o le frasi associate a questa mappa.

Questo secondo sistema utilizza molte meno risorse del computer, però genera un numero di errori più alto rispetto a quello ad architettura integrata.