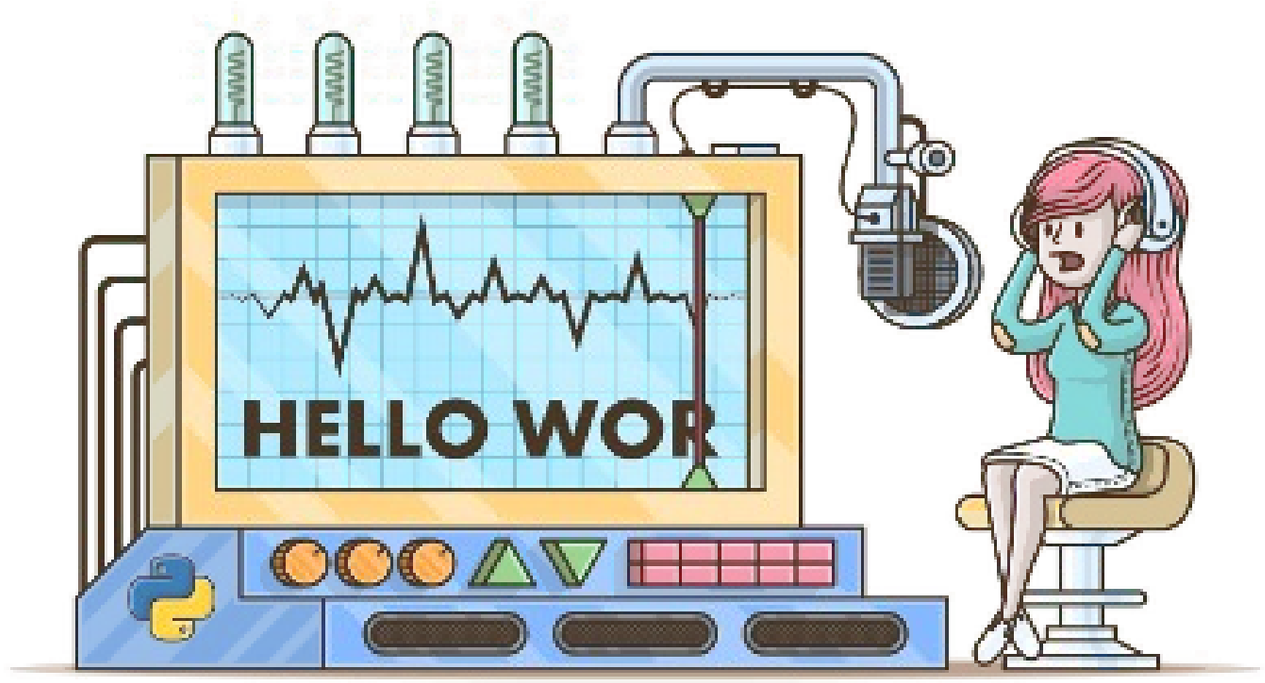
Riconoscimento vocale



Parasiliti Palumbo Maria (O46000888)

Progetti
Informatica Musicale 2018/19

Indice

- Introduzione al riconoscimento vocale
- La storia in breve
- Come funziona il riconoscimento vocale?
- Hidden Markov Models
- Dynamic Time Warping
- Reti neurali
- Importanza nel campo della disabilità
- "Vocalizer to mute", l'app sperimentale
- Analisi delle performance
- WER e WAcc

Introduzione al riconoscimento vocale

Il **riconoscimento vocale** è un sottocampo interdisciplinare della linguistica computazionale

CONSENTE il riconoscimento della voce umana e l'elaborazione in testo

L’uso del riconoscimento vocale ricopre **svariati campi applicativi**, come:

- ✤ SISTEMI IN-CAR
- ✤ ASSISTENZA SANITARIA
- ✤ DISABILITÀ
- ✤ MILITARE
- ✤ EDUCATIVO
- ✤ TRASCRIZIONE AUTOMATICA

Possiamo classificare i sistemi di riconoscimento in:

SPEAKER DIPENDENTI **SPEAKER INDIPENDENTI**

Storia in breve

ANNO 1952 → **Audrey**: sistema in grado di riconoscere i numeri da 0 a 9

ANNO 70 → **Harpy**: sistema in grado di riconoscere frasi complete con un dizionario di 1011 vocaboli

ANNO 80 → Applicazione del modello **Hidden Markov Modeling**
Nascono 3 grandi società: Covox, la Dragon Systems e la Kurzwei

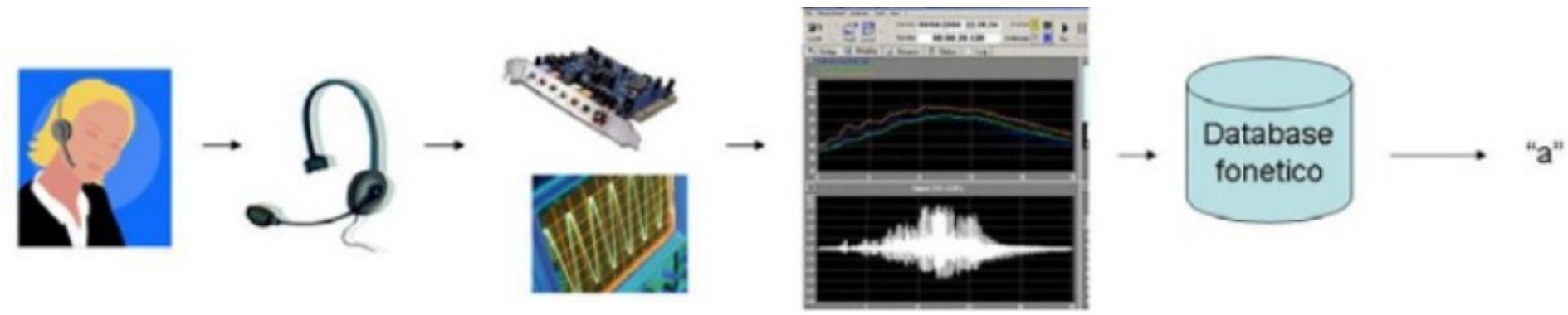
ANNO 90 → **Automatizzazione delle chiamate ai servizi clienti**

ANNO 2000 fino ai giorni nostri → L'accuratezza dei sistemi di riconoscimento vocale è pari all'80%. Nel 2003 la **Scansoft** acquista L&H e ViaVoice della IBM. Oggi attraverso l'introduzione del **machine learning** e intelligenza artificiale, i sistemi di riconoscimento vocale sono diventati molto accurati.

Come funziona il riconoscimento vocale?

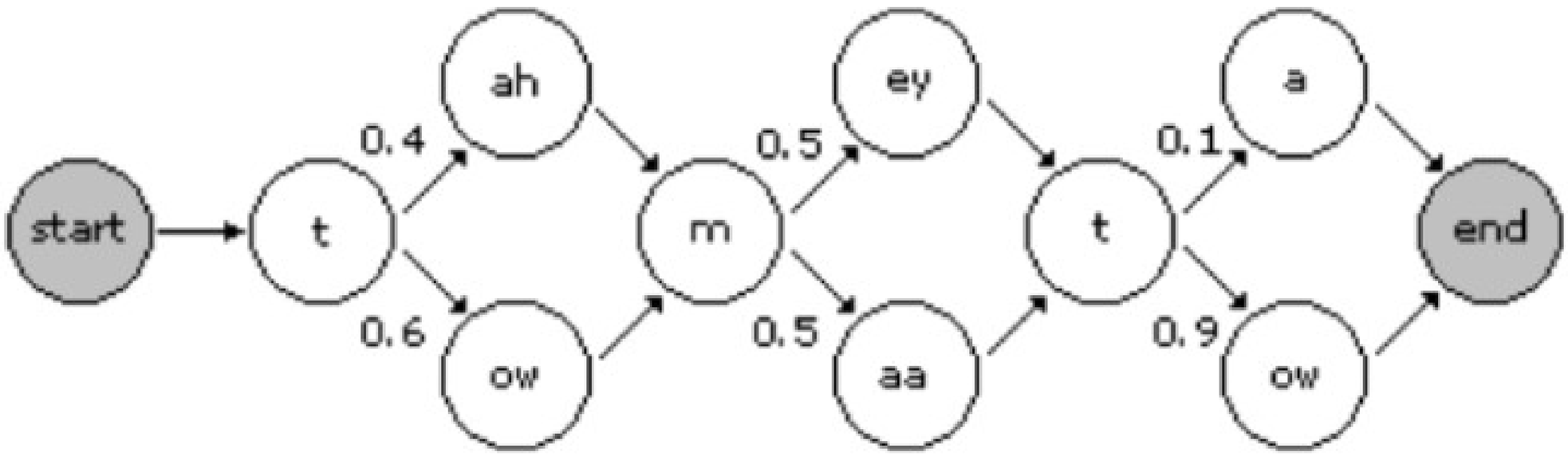
1) Trasformazione del segnale in **dominio della frequenza** (tramite FFT)
2) Si applica FFT in un segmento audio, il **pattern**, che identifica le ampiezze delle frequenze che compongono il suono, viene confrontato con tutti i pattern già conosciuti all'interno del **database** finché non viene individuato il più simile
3) Il sistema dalla FFT ricava dei valori, in base ai quali viene calcolato un **feature number** per ogni centesimo di secondo in esame.

   Ogni fonema produce più feature number
4) Durante la fase di training, il software apprende i pattern relativi per ogni fonema e una **serie di dati statistici**



Hidden Markov Models

- È un **modello probabilistico** che si basa sull'idea che i fonemi evolvano attraverso stati discreti.
- Applicabile se il sistema possiede la *proprietà di Markov*
- Il risultato finale è dato da una matrice di fonemi **collegati fra di loro in base alla probabilità** che un fonema sia legato all'altro.



Dynamic Time Warping

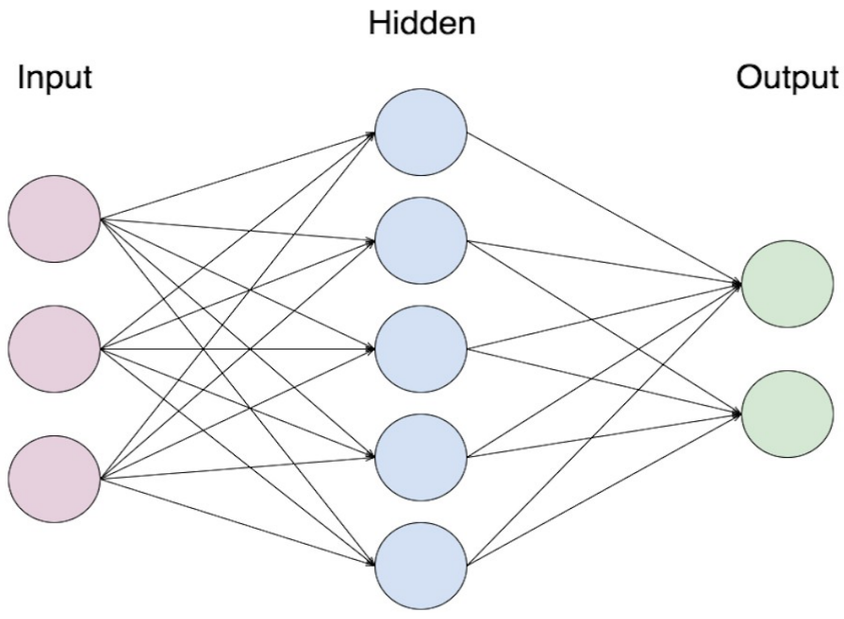
Algoritmo che permette l'allineamento ottimale non lineare tra due sequenze temporali che possono variare in velocità

Esistono anche algoritmi di approssimazione come **FastDTW** che ha il vantaggio di avere un costo computazionale lineare e inoltre risulta essere di facile implementazione e adatto quando si lavora con pochi dati

Reti neurali

Una rete neurale è composta da unità elaborative omogenee interconnesse **parallele** ed è caratterizzata da una **precisa architettura.**

Il numero canonico di strati di una rete neurale è 3:



- **INPUT**, converte i segnali d’input in segnali compatibili con gli strati successivi

- **HIDDEN**, lo strato nascosto si occupa dell’elaborazione dei segnali

- **OUTPUT**, raccoglie i risultati adattandoli alle richieste del blocco successivo

I neuroni sono collegati mediante parametri chiamati **pesi**.
Addestrare una rete neurale consiste proprio nel trovare i **valori ottimali dei pesi** finché si raggiungerà lo scopo prefissato.

Importanza nel campo della disabilità

Fra i vasti campi applicativi, il riconoscimento vocale trova la sua massima espressione nel campo della disabilità, in particolare la sordità.

Per consentire ai sordi di interagire con gli altri, esistono tecnologie in via di sviluppo come i guanti dotati di sensore o il Microsoft Kinect, o applicazioni per smartphone come “Ear Hear” o “MonoVoix”

“Vocalizer to mute”, l'app sperimentale

È stata realizzata da alcuni ricercatori dell’università di Lahore.

Rappresenta un’**innovazione** perché nessuno prima d’ora aveva tentato di rilevare il parlato di persone affette da sordità.

A 15 bambini di età compresa fra i 7 e i 13 anni è stato chiesto loro di pronunciare più volte le lettere dell’alfabeto, le cifre numeriche e una decina di brevi frasi.

- Mediante il riconoscitore vocale di **HTK** e l’**HMM** in back-end
- Percentuale di accuratezza pari al 97%

Analisi delle performance

Le prestazioni dei sistemi di riconoscimento vocale sono generalmente valutate in termini di **accuratezza** e **velocità**.

- La precisione viene valutata con il WER
- La velocità con il fattore tempo reale
- Altre misure: Single World Error Rate e il Command Success Rate

Il riconoscimento vocale da parte della macchina è tuttavia un **problema molto complesso**

WER e WAcc

STIME PIÙ UTILIZZATE

- la Word Error Rate (**WER**) si basa sull'analisi e confronto di una trascrizione manuale del segnale e la trascrizione relativa ottenuta dal sistema.

$$WER = \frac{S + D + I}{N} = \frac{S + D + I}{S + D + C}$$

- la WAcc (Word Accuracy), talvolta usata al posto di WER

$$WAcc = 1 - WER = \frac{N - S - D - I}{N} = \frac{H - I}{N}$$

Conclusioni

“C'è vero progresso solo quando i vantaggi di una nuova tecnologia diventano per tutti.”

- Henry Ford

@maryparp

Maria Parasiliti

maria.parasiliti.96@gmail.com

GRAZIE PER L'ATTENZIONE