INFORMATICA MUSICALE

UNIVERSITA' DEGLI STUDI DI CATANIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA
LAUREA TRIENNALE IN INFORMATICA
A.A. 2018/19
Prof. Filippo L.M. Milotta

ID PROGETTO: 14

TITOLO PROGETTO: Riconoscimento vocale

AUTORE: Parasiliti Palumbo Maria

Indice

1. Obiettivi del progetto

1.1 Introduzione al riconoscimento vocale

Il riconoscimento vocale è un sottocampo interdisciplinare della linguistica computazionale che sviluppa metodologie e tecnologie che consentono il riconoscimento della voce umana e l'elaborazione in testo attraverso un computer o un apposito sistema di riconoscimento vocale. È anche noto come riconoscimento vocale automatico (**A**utomatic **S**peech **R**ecognition), riconoscimento vocale del computer o sintesi vocale (**S**peech **T**o **T**ext).

L'uso del riconoscimento vocale ricopre svariati **campi applicativi:** dall'assistenza sanitaria ai sistemi in-car che consentono di avviare chiamate telefoniche o riprodurre musica senza distogliere lo sguardo dalla guida, dai programmi di dettatura vocale ai sofisticati sistemi di riconoscimento vocale degli aerei da caccia che permettono di aumentare le capacità del pilota di interagire con il velivolo.

Si rivela essere di estrema importanza **per le persone affette da disabilità**. Grazie a software di riconoscimento vocale che generano automaticamente i sottotitoli durante una sala conferenza, una lezione o una qualsiasi altra attività, i sordi non rimangono estraniati dalla società. I programmi di trascrizione automatica si rivelano utili per le persone affette da dislessia, RSI (**R**epetitive **S**train **I**njury / Disturbo degli arti superiori da lavoro) o altri disturbi alle mani.

I sistemi di riconoscimento vocale **possono essere classificati:**

SPEAKER DIPENDENTI

- ➔ Il modello viene adattato alla voce dell'utente
- ➔ Prevede una fase di addestramento in cui gli viene chiesto di leggere un testo a voce alta e con velocità naturale
- ➔ Permette la correzione di errori di interpretazione attraverso l'uso di funzioni vocali
- ➔ Sono dotati di algoritmi che tengono traccia delle correzioni per imparare dagli errori

SPEAKER INDIPENDENTI

- ➔ Il modello permette il riconoscimento di un parlato generico
- ➔ In termini di performance, è decisamente inferiore al primo
- ➔ Viene impiegato molto nei servizi automatici di informazione ed è efficiente in situazioni in cui l'utente è tenuto a dare risposte brevi
- ➔ Richiedono una elevata potenza di calcolo per essere quanto più performanti possibili

1.2 La storia in breve: dagli anni '50 a oggi

Nonostante le maggiori innovazioni in questo campo siano state realizzate in questi ultimi decenni, le sue radici affondano in anni più lontani. I primi tentativi di riconoscimento vocale vennero effettuati negli anni '50 negli Stati Uniti.

ANNO 1952 → Alcuni ricercatori dei Bell Laboratories costruirono *Audrey* (Automatic Digit Recognizer), **un sistema in grado di riconoscere i numeri da 0 a 9.**

ANNI 70 → La Carneige Mellor University perfezionò *Harpy*, un prototipo che consentiva il **riconoscimento di frasi complete**, ma con un dizionario limitato di 1011 vocaboli. Richiedeva un lungo periodo di apprendimento.

ANNI 80 → Venne applicato il metodo matematico sviluppato negli anni '60, chiamato **Hidden Markov Modeling**, per calcolare le probabilità che un suono corrispondesse a una certa parola. Nascono *3 grandi società:* la Covox, la Dragon Systems e la Kurzweil.

ANNI 90 → Il riconoscimento vocale venne perfezionato a tal punto che poteva essere impiegato per **automatizzare le chiamate ai servizi clienti.** Nel 1999 *Microsoft* acquistò la Entropic, che dichiarava di avere il più avanzato sistema di riconoscimento vocale.

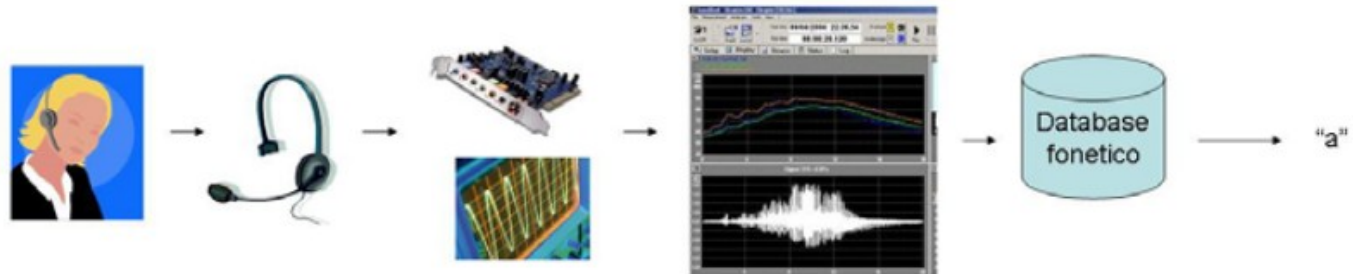
ANNI 2000 fino ai giorni nostri → I software di riconoscimento vocale erano diventati abbastanza accurati (80%) ed economici da essere inclusi gratuitamente in *Windows*. Nel 2003 la *Scansoft* acquistò la tecnologia di riconoscimento vocale della L&H e ViaVoice della IBM, per poi cambiare nome nel 2005 in **Nuance, tecnologia alla base di Siri**.

Oggi attraverso l'introduzione del **machine learning** e intelligenza artificiale, i sistemi di riconoscimento vocale sono diventati molto accurati. Siri, Assistente Google, Alexa e Cortana sono ormai entrati a far parte della nostra quotidianità.

1.3 Come funziona il riconoscimento vocale

Un sistema di riconoscimento automatico avviene secondo **le seguenti fasi:**

1. Trasformazione dei dati audio dal dominio del tempo al dominio della frequenza tramite **FFT** (**F**ast **F**ourier **T**ransform)
2. **Analisi del parlato:** viene analizzato il segnale vocale e vengono estratti i parametri spettrali
3. **Realizzazione del modello acustico:** tecnicamente conosciuta come fase di 'addestramento' o training. I segmenti di voce vengono classificati e identificati con simboli fonetici. L'inizio di un nuovo fonema viene identificato grazie alle pause.
4. **Realizzazione del modello linguistico:** i fonemi vengono composti per formare parole mediante l'applicazione del modello linguistico della lingua in uso.



È necessario convertire il segnale audio dal dominio del tempo a quello della frequenza, altrimenti l'individuazione del pattern non sarebbe possibile. La conversione viene realizzata attraverso un'analisi di spettro del segnale che consente l'identificazione delle frequenze che compongono i suoni. Applicando **la FFT** in un segmento di audio, si ottiene un pattern che identifica le ampiezze delle frequenze che compongono il suono, esso verrà confrontato con tutti i pattern già conosciuti all'interno di un database finché il sistema non individua quello più simile. Scendendo più nello specifico, il sistema dalla FFT ricava dei valori, in base ai quali viene calcolato un **feature number** per ogni centesimo di secondo in esame. Anche il database contiene i pattern di riferimento sotto forma di numeri.
Idealmente, ci sarebbe una corrispondenza diretta tra un feature number e un fonema, nella realtà una parola può essere pronunciata diversamente dalla stessa persona, pertanto **ogni fonema produce più feature number.**

I problemi legati al riconoscimento vocale sono:

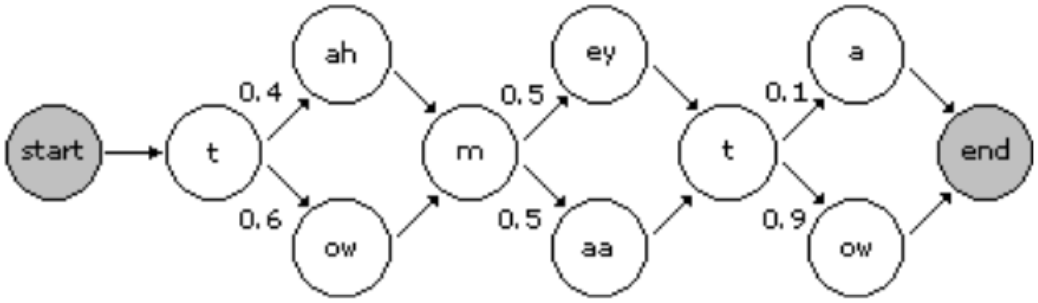
- persone diverse, pronunce diverse
- i rumori di sottofondo e altri elementi di disturbo
- il suono di ogni fonema varia in base al fonema precedente

Durante la fase di training, il software apprende i pattern relativi per ogni fonema e **una serie di dati statistici**, il più importante dei quali costituito da quante probabilità ci sono che un determinato fonema generi una certa sequenza di features.

Tra i modelli statistici matematici più utilizzati dai computer per il calcolo della probabilità, ritroviamo l'**Hidden Markov Models** (modello di Markov nascosto). Questo modello si basa sull'idea che i fonemi evolvano attraverso stati discreti.

Il risultato finale sarà dato da una catena costituita da un certo numero di stati, ciascuno dei quali genera un evento con una certa distribuzione di probabilità che dipende solo dallo stato che non può essere osservato.

Di fatto il modello è applicabile soltanto se il sistema possiede **la proprietà di Markov** e cioè quando gli stati che può assumere in futuro dipendono solo dallo stato presente e sono indipendenti dagli stati passati.



Esempio di modellazione della parola "Tomato".

Prevede la pronuncia in 3 diversi modi: inglese britannico, americano e possibile pronuncia quando si parla lentamente.

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.

Un altro metodo ampiamente utilizzato è quello basato sull'algoritmo **Dynamic Time Warping** che permette *l'allineamento ottimale non lineare* tra due sequenze temporali che possono variare in velocità. Supponiamo di avere due registrazioni vocali della stessa frase, DTW ci consentirà di trovare l'allineamento ottimale fra le due sequenze restituendo un valore di similarità, anche se le parole nella frase vengono pronunciate a velocità diverse.

DTW ha un costo computazionale quadratico rispetto alla dimensione dell'input.

Esistono anche algoritmi di approssimazione come *FastDTW* che restituisce un risultato diverso rispetto al DTW classico. Talvolta può essere impreciso, ma ha il vantaggio di avere un costo computazionale lineare e inoltre risulta essere di facile implementazione e adatto quando si lavora con pochi dati.

I modelli Markoviani presentano dei limiti: sono **classificanti**. Per ogni fonema viene salvata una unione di gaussiane che rappresentano la probabilità che il fonema si trovi in quel determinato stato. Si otterranno modelli separati per ogni fonema, che faticheranno a funzionare correttamente nel caso di suoni confondibili.

Questo problema viene superato con le **reti neurali**. Una rete neurale è **discriminante**: viene addestrata con esempi di tutti i possibili fonemi e riesce a ottenere risultati più soddisfacenti dei modelli Markoviani, su i fonemi confondibili.

Le **reti neurali artificiali** si ispirano al funzionamento delle reti neurali biologiche. Il nome "rete neurale" è dovuto alle similarità con il nostro cervello, in cui i neuroni sono in grado di funzionare singolarmente e sono collegati tramite gli assoni e le sinapsi.

Una **rete neurale** è composta da unità elaborative omogenee interconnesse **parallele**. Ciò consente ai singoli neuroni di lavorare indipendentemente l'uno dall'altro. È caratterizzata da una precisa architettura, elabora l'input seguendo un modello ben preciso e addestra la rete applicando un algoritmo di apprendimento. La scelta dell'architettura è legata all'algoritmo di apprendimento scelto. Esistono **3 differenti classi di architettura**:

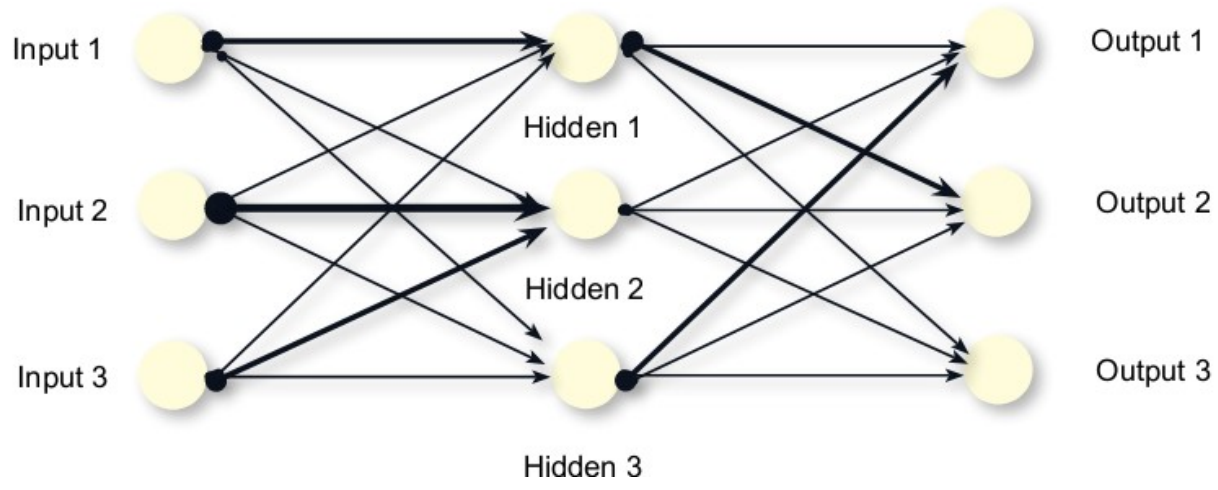
- **Single-layer feed-forward:** i dati si propagano in un unico senso, ha un livello di ingresso e uno di uscita
- **Multi-layer feed-forward:** come l'architettura precedente ma con più livelli hidden
- **Ricorrenti:** l'informazione torna indietro al neurone di input

Il numero canonico di strati in cui è suddivisa una rete neurale è tre:

• **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**. Strati e collegamenti hanno pesi e importanze diversa dagli altri. Ciò consente alla rete di fornire risultati diversi con gli stessi input. Quando arriva un segnale da analizzare, viene caricato nell'input del primo livello di neuroni, i quali generano un output che rappresenterà un nuovo input per le unità di livello superiore. Si procede per iterazione fino all'ultimo livello.

Addestrare una rete neurale consiste proprio nel **trovare i valori ottimali dei pesi** finché si raggiungerà lo scopo prefissato. Conclusa la fase di addestramento, una rete neurale, sarà in grado di fornire un output coerente, anche se riceve un input che non era stato presentato durante il training. Nel caso del riconoscimento vocale e dei caratteri, la rete viene addestrata finché non distinguerà perfettamente tutte le lettere dell'alfabeto.

1.4 L'importanza del riconoscimento vocale nel campo della disabilità

Come accennato nell'introduzione, fra i vasti campi applicativi, il riconoscimento vocale trova la sua massima espressione nel campo della disabilità.

La tecnologia mobile si evolve continuamente e progressivamente, tuttavia sono ancora poche le tecnologie sperimentate riguardanti il campo della sordità. In particolare, non esistono ancora applicazioni che interpretano il parlato dei sordi rendendolo comprensibile agli udenti. È stato questo che ha indotto alcuni ricercatori dell'università di Lahore nel 2018 a sviluppare un prototipo di applicazione che sfrutta la metodologia di riconoscimento vocale automatico per riconoscere il parlato della persona sorda e convertirlo in una forma riconoscibile a una persona udente.

Il nome dell'applicazione in questione è "Vocalizer to mute" (V2M) e dai risultati sperimentali è emersa una percentuale di accuratezza pari al 97,9%.

È convinzione comune che una persona sorda sia conseguentemente muta, ma non è così. In realtà mediante un percorso riabilitativo adeguato di logopedia e grazie a sofisticati dispositivi acustici, **i sordi possono imparare a parlare**. La capacità di recupero dipende dal grado di sordità della persona. Se la perdita di udito è profonda, difficilmente il sordo sarà in grado di farsi capire, se non attraverso l'uso della lingua dei segni. Questo inevitabilmente inciderà sullo sviluppo psicologico e sociale del soggetto, specialmente se è particolarmente giovane e ha una personalità fragile.

Per consentire ai sordi di interagire con gli altri, esistono tecnologie in via di sviluppo come i **guanti dotati di sensore** per l'interpretazione della LIS o il **Microsoft Kinect** che cattura i gesti della lingua dei segni e li traduce in linguaggio parlato e viceversa, **applicazioni per smartphone** come "Ear Hear" o "MonoVoix" che funge da interprete della lingua dei segni catturando i gesti tramite fotocamera del telefono e traducendoli nella lingua comprensibile agli udenti.

Il *prototipo di applicazione* realizzato dagli studiosi dell'università di Lahore rappresenta una novità, poiché **nessuno prima d'ora aveva tentato di rilevare il parlato di persone affette da sordità** considerata l'alta percentuale di errore. L'applicazione in via sperimentale è stata testata sottoponendo a esame 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. Per la fase di training e riconoscimento vocale, è stato impiegato il riconoscitore vocale di **HTK** (Hidden Markov Model Toolkit) , mentre è stato impiegato l'HMM in back-end. V2M ha registrato e processato i campioni vocali e dopo un'attenta analisi, ha mostrato sullo schermo il testo tradotto per le persone udenti. L'esperimento è stato un successo, ha infatti restituito una percentuale di accuratezza pari al 97%.

1.5 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 tasso di errore delle parole (**WER**), mentre la velocità con il **fattore tempo reale.** Fra le altre misure di accuratezza ritroviamo il tasso di errore a parola singola (**S**ingle **W**orld **E**rror **R**ate) e il tasso di successo dei comandi (**C**ommand **S**uccess **R**ate). Il riconoscimento vocale da parte della macchina è tuttavia un problema molto complesso, poiché le vocalizzazioni variano in termini di accento, pronuncia, articolazione, rugosità, nasalità, tono, volume e velocità, oppure in presenza di rumori e altri elementi di disturbo.

Le stime più utilizzate sono:

- la Word Error Rate (**WER**) che si basa sull'analisi di una trascrizione manuale del segnale e la trascrizione relativa ottenuta dal sistema. **Si confrontano** le due trascrizioni e si individuano le parole sbagliate (Substitutions), quelle mancanti (Deletetions) e quelle inserite erroneamente (Insertions). N corrisponde al numero di parole totali e C al numero di parole corrette.

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

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

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

dove H è uguale a N – (S+D), numero delle parole riconosciute correttamente

2. Riferimenti Bibliografici

- https://en.wikipedia.org/wiki/Speech_recognition

  Offre un quadro generale dell'argomento molto utile che mi ha permesso di focalizzare l'attenzione sui concetti chiave.

- http://www.crit.rai.it/eletel/2007-2/72-6.pdf

  Cenni storici importanti e analisi del funzionamento dei sistemi di riconoscimento vocale con modello di Markov nascosto e reti neurali

- "Audio e multimedia" di Vincenzo Lombardo e Andrea Valle

  Argomenti teorici trattati

- https://www.hindawi.com/journals/wcmc/2018/1013234/

  "A Novel Technique for Speech Recognition and Visualization Based Mobile Application to Support Two-Way Communication between Deaf-Mute and Normal Peoples" di Kanwal Yousaf, Zahid Mehmood, Tanzila Saba, Amjad Rehman, Muhammad Rashid, Muhammad Altaf e Zhang Shuguang.

  Si tratta di un articolo di ricerca riguardante una tecnica innovativa di riconoscimento vocale che favorisce la comunicazione fra i sordi e le persone udenti. Si è rivelato molto utile nella stesura dell'approfondimento sull'importanza del riconoscimento vocale nel campo della disabilità.

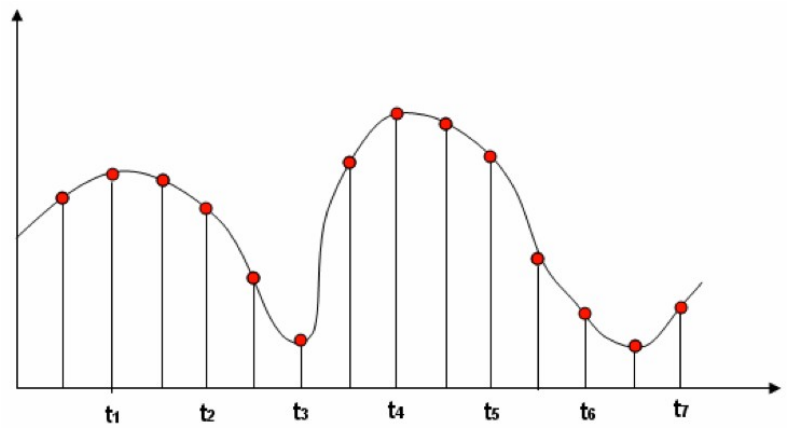
3. Argomenti Teorici Trattati

3.1 Suono analogico e digitale

La tecnologia analogica pervade la maggior parte dell'audio che ascoltiamo. La comunicazione del segnale in radio e in televisione e i dispositivi più diffusi sono di tipo analogico. Un **segnale analogico** può assumere infiniti valori, mentre quello **digitale** può assumere solo valori appartenenti a un insieme discreto. Un esempio di dispositivo di tipo analogico è il **microfono** che trasforma il *segnale sonore in segnale elettrico*. La tensione prodotta da un microfono è un segnale analogico. Per convertirlo in segnale digitale è necessario l'utilizzo di un convertitore analogico-digitale.

3.2 Campionamento del segnale

È la **discretizzazione del segnale analogico** nel tempo. Per ottenere un campionamento efficace, bisogna prelevare i campioni a una velocità adeguata in modo da tener conto di tutte le variazioni del segnale. Per un campionamento ottimale senza perdita di informazioni, bisogna utilizzare una frequenza di campionamento pari almeno al doppio della **frequenza di Nyquist.**



3.3 Trasformata di Fourier Veloce (FFT)

La **FFT** (**F**ast **F**ourier **T**ransform) è un algoritmo ottimizzato per calcolare la trasformata discreta di Fourier e viceversa.

Alla base della FFT c'è il **teorema di Fourier** che afferma che un segnale può essere scomposto in una serie di sinusoidi, ciascuna delle quali con una ben determinata frequenza, ampiezza e fase. Con la FFT è possibile analizzare lo spettro di un suono e vedere le sue componenti, passando dalla visione della forma d'onda con il tempo sull'asse X, alla visione in frequenza (con le frequenze sull'asse X), cioè passando **dal dominio del tempo a quello della frequenza**.