Speech To Text

Carovano Natalino

Introduzione(1/2)

- La comunicazione tramite il linguaggio parlato può essere vista come un'interazione tra un **trasmettitore** (il parlante) ed un **ricevente** (l'ascoltatore) che avviene mediante un **mezzo** (l'aria)
- Questa definizione è però incompleta, Jakobson rileva la presenza di altri tre fattori costitutivi della comunicazione verbale:
  - **Messaggio**
  - **Codice**
  - **Contesto**

Introduzione(2/2)

- Si possono dare le seguenti definizioni:
  - **Messaggio**: informazione inviata dall'emittente al destinatario
  - **Codice**: sistema di segni interamente (o parzialmente) condiviso da emittente e destinatario, che stabilisce il significato dei segni
  - **Contesto**: si tratta della situazione nella quale di fatto si situa la comunicazione
- Questi tre fattori permettono di formulare ed interpretare il messaggio
- Da un punto di vista tecnico si possono trovare tante definizioni, che però non riescono a sintetizzare la complessità racchiusa nel termine **Parlato**

Parlato(1/2)

- Il Parlato ha origine dall'intenzione di trasmettere un'idea ed il sistema nervoso centrale, tramite i muscoli, la trasforma in atto ed è capace di trasferirla ad un ascoltatore
- Quest'ultimo recepisce il segnale, sottoforma di variazioni di pressione dell'aria, nel sistema uditivo, lo elabora e lo converte in stimoli neurologici che il sistema nervoso centrale riesce a comprendere

Parlato(2/2)

- Il parlante controlla costantemente gli organi di produzione basandosi sull'ascolto del segnale che sta producendo
- Quando le parole si combinano in sequenze, la pronuncia dei loro singoli segmenti può subire diverse alterazioni

Alterazioni del Parlato

- La velocità ed il ritmo possono fare adottare per alcuni segmenti un'articolazione più debole, far sì che altri siano omessi oppure che ne siano inseriti di nuovi, o che qualche segmento modifichi del tutto le sue caratteristiche
- Alcuni esempi di trasformazioni sono :
  - Elisione
  - Assimilazione

Trasformazioni nel parlato(1/2)

- Nel parlato (soprattutto in quello rapido) si possono tralasciare (o elidere) dei suoni
- Inoltre, suoni adiacenti si influenzano in maniera frequente gli uni con gli altri così da diventare più simili, ovvero si **assimilano**

- **Esempio :** parlando rapidamente le parole
  “con prudenza” sono pronunciate “comprudenza”

Trasformazioni nel parlato(2/2)

- Una nota va infine dedicata alla cosiddetta "coarticolazione" dei foni. Essa coinvolge i movimenti che l'apparato fonatorio compie nel passaggio dalla produzione di 2 foni ed è un fondamentale concetto nell'ambito del riconoscimento automatico

- **Esempio:**

  la "a" di "tap" è diversa da quella in "iam"

Disfluenze

- Altre volte tra due parole può essere introdotto un suono ad esempio nel caso delle "disfluenze" o delle "false partenze", che sono spesso presenti nelle conversazioni
- Esempi di questo tipo sono nei balbettii presenti nei momenti di incertezza o di imbarazzo :
  “ ti-ti posso invitare a cena?"

Rumore e Riverbero

- Il parlato è anche influenzato da fattori come il :
  - **Rumore**: è un segnale di disturbo rispetto all'informazione trasmessa in un sistema
  - **Riverbero**: è un fenomeno acustico legato alla riflessione del suono da parte di un ostacolo posto davanti alla fonte sonora
- Ad esempio: Se in una stanza una sorgente sonora cessa di irradiare, il livello sonoro diminuisce in dipendenza dell'assorbimento acustico delle pareti
- Il riverbero ha aspetti negativi, come il rischio di mascheramento delle sillabe del parlato, e positivi, come il rinforzo dell'intensità della sorgente

Parlato Umano

- Negli esseri umani il riconoscimento del Linguaggio Parlato, quello con cui ci esprimiamo ogni giorno, è un meccanismo naturale
- Oltre ad essere naturale, è anche “robusto” ed “efficiente”, nel senso che riesce a funzionare correttamente anche in situazioni sfavorevoli (presenza di rumori e di riverbero) oltre a gestire varie alterazioni del parlato (come quelle illustrate in precedenza)

Modello uditivo umano

- Il Parlato, dunque, presenta situazioni difficili da affrontare, se si ha intenzione di costruire un modello del sistema uditivo umano
- Bisogna tener conto di tutte le variazioni illustrate, che influiscono sui modelli di pronuncia delle parole che si vuole introdurre nel sistema artificiale, se si vuole costruire un sistema affidabile
- Inoltre, queste variazioni dipendono dalla lingua scelta quindi variano passando da una lingua all’altra

Parlato connesso(1/2)

- Un problema forte è inoltre costituito dal cosiddetto "parlato connesso"
- Nei dialoghi informali tra interlocutori vengono commessi errori di pronuncia, inoltre c'è la presenza di pause rumorose e di alterazioni imprevedibili dell'ambiente circostante
- Queste alterazioni che per un umano sono trascurabili, diventano fonti di grosso disturbo per una macchina

Parlato connesso(2/2)

- Comprendere le separazioni tra parole, distinguere il parlato dal rumore di sottofondo o dalla musica, non è un compito facile per un sistema artificiale, ma è un problema da affrontare se si vuole costruire un modello percettivo che emuli il comportamento dei sistemi biologici

Riconoscimento del parlato umano(1/2)

- Il sistema di riconoscimento umano del parlato compie computazioni, filtraggi e adattamenti ai diversi parlanti con cui ha a che fare, riesce dunque a trasformare un segnale vocale in una successione di vocaboli, alla quale poi da un'interpretazione
- Allo stato attuale della tecnologia non è ancora chiaro come questo avvenga

Riconoscimento del parlato umano (1/2)

- Dal punto di vista artificiale, si pensa che l'essere umano riconosca il parlato mediante una interazione complessa tra molti livelli di elaborazione, usando informazione sintattica e semantica assieme a potenti processi di elaborazione e classificazione
- I sofisticati algoritmi sviluppati finora, non sono sufficienti a reggere il confronto con ciò che avviene nel sistema nervoso centrale

Speech To Text

- **Speech To Text** (STT): è il processo che analizza il parlato e ne produce una forma testuale equivalente
- Un esempio tipico di applicazione di STT è un sistema di dettatura (ma non è l'unico!)
- Quindi possiamo dividere il processo in 2 sottoprocessi :
  - Riconoscimento del parlato
  - Trascrizione del parlato

ASR

- **Riconoscimento automatico del parlato**: è il processo grazie al quale un computer (o un altro tipo di macchina) identifica le parole pronunciate
- In inglese questo processo ha il nome di **Automatic Speech Recognition (ASR)**
- Sempre più spesso si parla di ASR, non solo per indicare il processo di riconoscimento del parlato, ma anche per intendere **Sistemi di riconoscimento automatico del parlato**

Cosa non è ASR ?

- In molti contesti è erroneamente conosciuta come voice recognition
- **Voice Recognition** (o **Speaker Recognition**) è un processo (in relazione con ASR) che cerca di identificare la persona che sta parlando
- Inoltre, capita spesso che si faccia coincidere il concetto di ASR (inteso come sistema) con sistemi di dettatura, ma tale definizione non è corretta

Perché ASR ?

- Il riconoscimento automatico del linguaggio parlato è molto interessante sotto diversi punti di vista:
  - **commerciale**: vede fiorire un insieme di applicazioni utili in diversi campi
  - **scientifico**: la sfida di riuscire a creare un modello del sistema di riconoscimento umano del linguaggio naturale ha la conseguenza di farci capire meglio come funziona l'organismo umano

Costruire un ASR

Il processo di costruzione di un ASR è molto laborioso perchè coinvolge tipicamente diverse fasi:

- stabilire quale sarà l'**Unità Base** sulla quale opererà il riconoscimento
- la raccolta del **Corpus**
- stabilire l'**architettura** del riconoscitore

Unità di base

- L'**Unità Base**, nel riconoscimento umano del parlato, può essere definita come la "forma minima" di informazione acustica attorno alla quale è organizzata la maggior parte dell'elaborazione del segnale per gli esseri umani
- Quindi una porzione significativa del processo di riconoscimento, umano o artificiale, del parlato dovrebbe lavorare su questa entità
- Non è ancora chiaro se questa unità esista (e sia unica) nell'ambito della produzione umana, è comunque essenziale nella costruzione di un ASR

Unità di Base ideale

A prescindere dalla sua esistenza nell'ambito della produzione umana del parlato, un'unità base, per essere utile ad un riconoscitore, dovrebbe essere:

- sufficientemente "accurata" da poter fornire informazione utile ad un riconoscimento in tutti i contesti possibili
- **addestrabile**: bisogna avere abbastanza dati per ogni evento possibile
- **generalizzabile**: ogni parola dovrebbe poter essere costruita a partire dall'entità in questione

Scelta dell'unità di base

Negli ultimi 50 anni, i ricercatori hanno proposto molti tipi diversi di unità base.
Alcune di queste hanno interessato unità sottofonetiche, foni, bifoni, trifoni, mezze sillabe, sillabe, parole, frasi intere e numerose altre entità .
Generalmente la scelta si riconduce alle seguenti unità:

- **Foni**: addestrabili e generalizzabili
- **Parole**: accurate e poco addestrabili
- **Sillabe**: un buon compromesso tra foni e parole

Foni

- Intuitivamente, si può pensare che le lettere dell'alfabeto di una lingua siano i mattoni con i quali vengono costruite anche le frasi parlate, e non solo quelle scritte
- I sistemi alfabetici, in uso in molte lingue nel mondo, nascono nel tentativo di rendere graficamente i suoni pronunciati nel linguaggio parlato
- Varie ragioni di ordine storico hanno portato ad una situazione in cui lingue che usano uno stesso alfabeto (ad esempio quello latino) associano agli stessi simboli pronunce diverse

Alfabeto fonetico

- Per eliminare questi problemi, ai fini dell'apprendimento di una lingua, o di una sua descrizione formale, i linguisti hanno ideato un sistema di trascrizione dei suoni pronunciati nel parlato di tutte le lingue
- Un alfabeto di questo tipo è detto **fonetico** ed è indipendente dalla lingua considerata (considera solo i suoni)
- L'alfabeto fonetico più diffuso è quello Internazionale (API o IPA)

Fonemi

- L'idea che ha generato la nascita dell'API è che le parole siano successioni di foni, suoni distinguibili che sono realizzazioni o istanze di fonemi, categorie astratte di suoni
- Ad esempio col simbolo /t/ si indica il fonema delle "t" ossia la classe di tutte le t pronunciabili
- Quindi possiamo definire un **fonema** come il nome di una classe astratta
- Le realizzazioni effettive della classe sono i **foni**

Trascrizione fonetica

- Una frase registrata da una conversazione sarà dunque trascritta come una successione di foni corrispondenti a fonemi
- Da un punto di vista più formale, si definisce trascrizione fonetica un'operazione consistente nel rappresentare per iscritto la forma fonica di una parola (o di un testo) facendo uso dell'alfabeto fonetico
- Alcune volte per la trascrizione sono scelti degli alfabeti diversi da quello fonetico (ad esempio lo SNOR, che verrà illustrato in seguito)

Sillabe(1/2)

- Storicamente, invece, l'adattamento del simbolo al suono fu molto graduale e andò di pari passo con la semplificazione dei segni che rappresentarono prima sillabe (solo dopo le singole lettere)
- Inoltre le sillabe includono aspetti come la coarticolazione tra foni
- Attraverso le sillabe si riesce anche a non imporre una grossa complessità computazionale (circa 1500 sillabe per coprire l' 80% di un vocabolario di 26000 termini)

Sillabe(2/2)

- Nonostante alcuni vantaggi, ci sono molte questioni ancora insolute circa il ruolo delle sillabe nel linguaggio naturale e sulle molte difficoltà pratiche che si incontrano nell'usarle in un sistema artificiale
- Inoltre presentano ancora alcuni problemi che ne limitano l'uso:
  - mancanza di una definizione formale ed universale di sillaba e dei suoi confini
  - ci sono ancora dei problemi ad estrarre informazioni utile all'elaborazione

Parole(1/3)

- Modelli che hanno cercato di riconoscere parole intere sono stati usati spesso ed hanno i seguenti vantaggi:
  - riescono a catturare la coarticolazione fonetica senza necessitare di modelli ulteriori (nel caso di riconoscitori fonetici c'è bisogno di un ulteriore modello)
  - quando questi modelli sono ben addestrati, forniscono in genere le prestazioni migliori, anche perchè hanno bisogno di un addestramento molto inferiore a quello fonetico o sillabico

Parole(2/3)

- Nel caso di applicazioni a vocabolari ristretti, i riconoscitori di intera parola sono considerati i migliori, mentre per grandi dizionari sorgano problemi difficili da risolvere dovuti:
  - al numero delle parole contenute nell'insieme dei dati di addestramento
  - al riconoscimento di parole non contenute nell'insieme dei dati di addestramento (si costruiscono sistemi che compongono più modelli di parola, anche se non è sempre sufficiente)

Parole(3/3)

- Il problema maggiore rimane la difficoltà nel creare sistemi artificiali basati sulle parole, che siano indipendenti dal parlante e da fattori esterni (rumore)
- Nonostante questi problemi, modelli che hanno cercato di riconoscere intere parole sono stati largamente usati nell'ambito del riconoscimento automatico del parlato

Corpus

- I più diffusi sistemi di riconoscimento hanno bisogno di un insieme di registrazioni vocali sulle quali addestrarsi, che costituiscono il cosiddetto **corpus**
- Si può distinguere 2 tipi di ASR a seconda del corpus usato durante l'addestramento:
  - basato sulla voce dell'utente che fa uso del sistema: solo questi potrà usare il sistema con una alta "accuratezza"
  - costituito da un grosso insieme di registrazioni di voci appartenenti a diverse persone: non è richiesta nessuna sessione di addestramento iniziale ed è indipendente dall'utente

Tipologie di ASR

Le diverse teorie sulla natura del parlato, sulle quali vengono costruiti i modelli che cercano di emulare il sistema di elaborazione acustica umano, hanno portato ad approcci diversi nell'implementazione di un ASR.
Si distinguono generalmente due approcci:

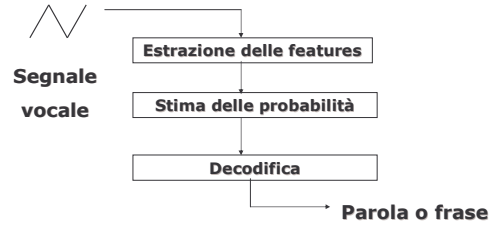
- sistemi esperti basati su conoscenze umane
- sistemi di tipo matematico (Reti Neurali, Modelli Markoviani)

Sistemi esperti

- I sistemi esperti basati su conoscenze umane sono dei programmi che incorporano un esplicito dominio di conoscenza (ad esempio: medicina, chimica …)
- Non essendo sistemi molto diffusi nell'ambito dello Speech To Text, ci concentreremo invece, sui sistemi di tipo matematico che, fino ad adesso, sono stati quelli che hanno dato i migliori risultati

Architettura generale

Una delle architetture più usate per un ASR di tipo matematico è la seguente :

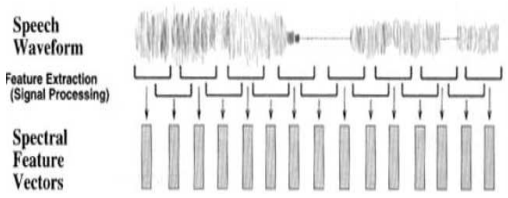


Estrazione di Features(1/2)

- In questo blocco vengono codificate delle porzioni del segnale, in genere proveniente da un microfono, e campionate mediante una tra un insieme di possibili tecniche di **Analisi del Segnale**
- Il problema della codifica è fondamentale per un ASR, in quanto da essa dipende la possibilità di discriminare un'entità linguistica da un'altra
- Una codifica breve e carica di informazione è la migliore possibile perchè consente un buon riconoscimento con il minimo addestramento

Estrazione di Features(2/2)

- Dalla fase di estrazione vengono derivati dei vettori numerici che esprimono le caratteristiche (features) del segnale



Stima delle probabilità(1/3)

- In questa parte i vettori numerici, associati dal blocco precedente a porzioni di segnale, vengono passati ad un sistema che associa loro una particolare classe di appartenenza (classificazione)
- Tale classe rappresenta una delle possibili realizzazioni di un' unità linguistica
- Ad esempio in un ASR basato sull'unità linguistica fonema, alla codifica sarà associato un particolare fonema

Stima delle probabilità(2/3)

In questa fase avviene una prima, fondamentale, fase di riconoscimento, ossia dalla codifica di una successione di segmenti del segnale di passa ad una successione di unità linguistiche che verranno analizzate dal blocco successivo
I sistemi di decodifica utilizzati più di frequente sono:

- Reti Neurali
- Modelli di Markov Nascosti(HMM)

Stima delle probabilità(3/3)

Il corretto funzionamento della sezione in questione, dipende fortemente dai dati di addestramento ed in genere, a causa della carenza di informazione della codifica, ha bisogno di numerosi campioni su cui basare la classificazione
Le esigenze del sistema sono ovviamente diverse a seconda del tipo di classificazione che si vuole realizzare

Decodifica

A causa della carenza di informazione contenuta nella codifica e della variabilità del parlato, la successione di stime proveniente dal blocco precedente ha una probabilità non sempre alta di essere corretta

La decodifica ha la funzione di rivolgersi ad una "grammatica" e ad un modello di pronuncia per le parole contenute in un "dizionario" di riferimento, allo scopo di ricostruire la frase corretta a partire dalle stime probabilistiche precedenti

Dizionario

Dizionario: è una lista di enunciati che possono essere riconosciuti dall'ASR

- Generalmente, vocabolari di dimensioni minori permettono un riconoscimento migliore e più rapido, mentre vocabolari più estesi creano maggiori difficoltà di riconoscimento
- A differenza dei normali dizionari, ciascun elemento presente nel dizionario di un ASR non deve necessariamente essere una singola parola
- Ad esempio: "Wake up" potrebbe essere inserito come un enunciato unico

Grammatica(1/2)

Nella fase di decodifica, viene utilizzata una grammatica utile a ricostruire la frase corretta

Questa grammatica viene generata :

- sulla base delle regole grammaticali della lingua a cui ci si riferisce (da qui il termine "grammatica di riferimento")
- da un'analisi statistica compiuta su un grosso insieme di dialoghi registrati e trascritti (corpus), dalla quale viene ricavata la probabilità che una parola succeda un'altra

Grammatica(2/2)

Ad esempio, se è già stata riconosciuta la parola "casa" e quella successiva è "mia", il sistema riconoscerà la frase "casa mia" come una plausibile trascrizione del segnale; se invece la parola successiva fosse "mela", allora esso darebbe poco credito alla trascrizione "casa mela", perché non avrebbe senso secondo la sua grammatica di riferimento

Analisi del segnale

- Nell'ambito della costruzione di sistemi di riconoscimento automatico del parlato (ASR) è richiesto l'estrazione di features
- **Feature**: caratteristica fondamentale di un segnale acustico che lo codifica perdendo la minor quantità possibile di informazione
- I metodi di analisi costruiti a questo scopo hanno cercato di individuare caratteristiche spettrali legate alle frequenze che formano i singoli foni del parlato o altre caratteristiche intrinseche del segnale analizzato (quasi sempre su intervalli di lunghezza compresa tra 20 e 250 ms)

Tecniche di Analisi

Sono disponibili diverse tecniche di analisi del segnale, tra le più utilizzate possiamo citare:

- **LPC**
- **MFCC**
- **Modulation Spectrogram**

Le prime sono tecniche standard che operano su intervalli dell'ordine dei 20 ms e cercano di codificare le caratteristiche dei foni componenti il segnale

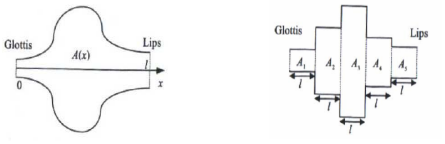
La Modulation Spectrogram invece cerca di elaborare segnali che operano su intervalli di circa 250 ms

Linear Predictive Coding(1/5)

- Il “Linear Predictive” Coding (LPC) è uno dei metodi classici e più datati per l'estrazione di features da un segnale
- Il suo scopo è quello di catturare le frequenze formanti dei foni che costituiscono un segnale vocale
- Non essendo basato su un modello molto accurato e robusto, la tecnica fornisce delle features che sono molto dipendenti dal rumore, dal riverbero e dalle variazioni delle modalità di registrazione

Linear Predictive Coding(2/5)

- Possiamo costruire un modello della produzione del suono in questo modo: immaginiamo di “srotolare” il cavo orale orizzontalmente mentre viene pronunciata una frase, quello che otterremo sarà un tubo di forma variabile nel tempo



Linear Predictive Coding(3/5)

- Nella fase successiva, discretizziamo il tubo suddividendolo in N parti di forma rettangolare, ognuno di dimensioni variabili nel tempo ma di area costante
- Il tubo, nell’istante in cui viene prodotto un fono, può essere visto come un filtro con una forma definita e costante che ha la funzione di cambiare (modulare) l’aspetto del treno d’impulsi o del rumore
- Il risultato finale sarà proprio il suono prodotto

Linear Predictive Coding(4/5)

- La tecnica LPC tenta di ottenere informazione sulla risposta in frequenza del filtro, perché è da quella che dipendono le frequenze di risonanza (frequenze fondamentali) del segnale che fuoriesce dalla bocca
- Si ipotizza allora che il cavo orale sia approssimabile con un filtro (una funzione) del tipo:

$$H(z) = \frac{1}{1 - \sum_{k=1}^{p} a_k z^{-k}}$$

Linear Predictive Coding(5/5)

- Quindi si risolve la seguente equazione:

$$S(z) = H(z) * E(z)$$

- Dove sia E(z) la trasformata Z dell’eccitazione glottidea, H(z) quella della risposta all’impulso del filtro e S(z) quella del segnale d’uscita
- E(z) e H(z) non sono noti a priori, quindi vanno approssimati
- Il Linear Predictive Coding prende il suo nome dal fatto che il campione all’istante n può essere “predetto” da una combinazione lineare di p campioni passati

Mel-Frequency Cepstrum Coefficients(1/4)

- I coefficienti tratti da questa tecnica di Analisi del Segnale costituiscono una rappresentazione del segnale definita a partire da un'analisi del segnale (in un breve intervallo di tempo) tramite trasformata di Fourier(DFT)
- Dato l'ingresso x[n], applichiamo prima una DFT:

$$X_a[k] = \sum_{n=0}^{N-1} x[n] e^{-j 2\pi n k / N} \qquad 0 \le k < N$$

Mel-Frequency Cepstrum Coefficients(2/4)

- E definiamo un banco di M filtri, ove il filtro m-esimo è della forma:

$$H_m[k] = \begin{cases} 0 & k < f[m-1] \\ \dfrac{(k - f[m-1])}{(f[m] - f[m-1])} & f[m-1] \le k \le f[m] \\ \dfrac{(f[m+1] - k)}{(f[m+1] - f[m])} & f[m] \le k \le f[m+1] \\ 0 & k > f[m+1] \end{cases}$$

- Tali filtri valutano lo spettro medio attorno ad un centro di frequenza, variabile a seconda di m

Mel-Frequency Cepstrum Coefficients(3/4)

- Il Mel-Frequency Cepstrum è la trasformata coseno delle uscite degli M filtri:

$$c[n] = \sum_{m=0}^{M-1} S[m] \cos(\pi n (m - 1/2)/M) \qquad 0 \le n < M$$

- Il numero di filtri, M, è legato all'intervallo di frequenze che vogliamo analizzare e alla frequenza di campionamento del segnale
- Fissare m ad un certo valore vuol dire limitare il domino delle frequenze analizzate

Mel-Frequency Cepstrum Coefficients(4/4)

- Il numero di filtri, M, è legato all'intervallo di frequenze che vogliamo analizzare e alla frequenza di campionamento del segnale
- Nell'ambito del parlato spontaneo, l'intervallo in questione arriva fino a circa 3500 Hz e, ad una frequenza di campionamento del segnale di 16000 Hz, M può essere posto uguale a 13
- La natura dei filtri rende i coefficienti MFCC relativamente resistenti al rumore, sebbene il loro legame con i segmenti fonetici ne limiti l'applicabilità solo all'ambito di ASR che sfruttino intervalli di tale lunghezza

Modulation Spectrogram(1/3)

- E' una tecnica che cerca di ottenere informazioni utili per un riconoscimento di un segnale, rivolgendosi a porzioni di questo che hanno lunghezza pari a 250 ms
- Lo scopo è quello di sfruttare l'informazione proveniente da lente variazioni nelle caratteristiche del segnale, al fine di ottenere delle features possibilmente indipendenti dal rumore, dal riverbero, dal parlante e da fattori che il sistema nervoso centrale non considera indispensabili quando opera un riconoscimento vocale

Modulation Spectrogram(2/3)

- Il segnale in ingresso viene passato in un banco di filtri. Il processo cerca di simulare la variazione di sensibilità del sistema uditivo alle diverse frequenze
- Ogni uscita del banco rappresenta il segnale filtrato secondo una banda acustica diversa
- Al segnale filtrato vengono tagliate le ampiezze negative (non utili) e poi passato in un filtro (inviluppo) per meglio evidenziare le separazioni tra unità grammaticali

Modulation Spectrogram(3/3)

- Il segnale inviluppato viene sottocampionato per ridurre la complessità di calcolo
- Viene calcolata la trasformata di Fourier in finestre di lunghezza 250 ms e si prendono in considerazione solo determinate componenti (a 4 Hz) per ogni segmento
- I segmenti di analisi lunghi possono essere molto resistenti ad interferenze o rumori
- Lo svantaggio che, essendo più grandi di quelli fonici, ci fanno rinunciare ad aspetti più sottili della struttura del parlato

ASR Multilivello(1/2)

- Tutti i metodi illustrati fino ad adesso presentano vari problemi di diverso tipo
- Rilasciando l'ipotesi di un'unica unità di base, quello che possiamo fare è cercare di "fondere" due sistemi per ottenerne uno migliore
- L'utilizzo del Modulation Spectrogram combinato ad altre features di tipo fonetico (ad esempio MFCC) ha mostrato un grosso incremento delle prestazioni degli ASR, soprattutto nel caso di parlato riverberante

ASR Multilivello(2/2)

- In questo approccio i risultati in uscita dal riconoscitore di ogni canale vengono "fusi" (fase di merging) per ottenere una valutazione finale
- Sempre più spesso gli ASR più recenti vengono costruiti seguendo questo principio, in questa maniera si è raggiunta una certa accuratezza nella fase di analisi del segnale audio
- Ulteriori ottimizzazioni riguarderanno anche altri aspetti (come la multigranularità o la generazione di particolari grammatiche di riferimento)

Approccio Multigranulare(1/2)

- Da quello che abbiamo visto, emerge un legame sulla scelta dell'unità base del parlato tra foni e sillabe
- Alcune recenti teorie, hanno ipotizzato che una tale entità non esista realmente, ma piuttosto che il parlato abbia una struttura a più "granuli", ossia basata su più livelli linguistici nel più basso dei quali si colloca una forma di informazione a metà tra fono e sillaba
- I riconoscitori implementati sulla base di questa ipotesi hanno fornito prestazioni più alte di quelli "classici" (anche in presenza di rumore)

Approccio Multigranulare(2/2)

- Il sistema comincia da un certo livello, ad esempio quello sillabico in cui procede da una sillaba all'altra finché non si trova in una situazione di incertezza, a quel punto passa al livello fonetico, oppure, se la grammatica lo permette, una volta riconosciuto un certo numero di sillabe o foni, salta direttamente alla parola successiva
- E' basato principalmente su una funzione euristica che deve scegliere quale livello utilizzare durante l'analisi

Problemi degli ASR(1/2)

I maggiori problemi degli ASR rimangono quelli di robustezza, vale a dire la conservazione delle prestazioni in presenza di :

- Rumore
- Riverbero
- Parlato non "chiaro" (pronuncia errata di parole, pause rumorose nel parlato)

Tecniche come il **Modulation Spectrogram** o l'**MFCC** sono nate proprio allo scopo di ottenere codifiche del segnale indipendenti almeno dal rumore e dal riverbero

Problemi degli ASR(2/2)

Esistono anche problemi dovuti ad alcune variazioni delle condizioni in cui il sistema è stato addestrato

Ci possono essere variazioni:

- della distanza dalla sorgente vocale
- del tipo di microfono usato per l'acquisizione del segnale
- del parlante

Valutazione di un ASR(1/2)

- Abbiamo parlato fino ad adesso di accuratezza nel riconoscimento, ma come si può valutare un ASR in maniera formale?
- Generalmente, le prestazioni di un ASR sono specificate in termini di accuratezza e velocità
- **Accuratezza**: quanto bene il riconoscitore è in grado di riconoscere enunciati. Questo include non solo la capacità di identificare un enunciato noto ma anche di determinare se un certo enunciato non è presente nel dizionario

Valutazione di un ASR(2/2)

- **Velocità**: è in genere espressa in relazione al tempo che occorre per elaborare un corpus di una certa lunghezza (espressa in ore)
- Le due misure che in genere vengono utilizzate per il calcolo di questi fattori sono :
  - **word error rate** per l'accuratezza
  - **real time factor** per la velocità
- Il livello di accuratezza minimo accettabile per un sistema dipende dal particolare tipo di applicazione in cui è utilizzato
- Alcuni sistemi possono presentare un'accuratezza del 98% (in condizioni ottimali)

Word Error Rate

- Il WER è derivato dalla distanza di Levenshtein (lavorando a livello di parole) allineando la sequenza di stringhe riconosciuta con una sequenza di riferimento

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

- S: numero di sostituzioni
- D: numero di cancellazioni
- I: numero delle inserzioni
- N: numero delle parole di riferimento
- Il Wer indica la percentuale di errori nella trascrizione

Real Time Factor

- Se per processare un input di durata **I**, occorre un tempo **P**, il real time factor è definito come:

$$\mathrm{RTF} = \frac{P}{I}$$

- Esempio : Se per processare una registrazione di durata 2 ore ne occorrono 8, il real time factor è uguale a 4
- Quando il real time factor è uguale a 1, il processo è elaborato in real time
- L'RTF è un parametro dipendente dall'hardware

Applicazioni

- Verrà illustrata una serie di possibili campi di applicazione nel campo dello speech to text:
  - Traduzioni automatica
  - Riconoscimento dei comandi(interfaccia vocale con il computer) (*)
  - Dettato
  - Home automation
  - Trascrizioni Mediche
  - Telefonia mobile
  - Valutazione della pronuncia in applicazioni di aiuto per l'apprendimento del linguaggio
  - Interazione uomo-robot
  - Verifica/Identificazione della voce

Sistemi opensource/commerciali

Esistono una vasta serie di sistemi opensource e commerciali disponibili per utenti di varie categorie, ad esempio:

- **XVoice** : è sistema di dettatura
- **CVoiceControl/kVoiceControl**: permette ad un utente di eseguire applicazioni usando comandi vocali
- **ISIP**: è un motore di riconoscimento del parlato reso disponibile dall'Institute for Signal and Information Processing della Mississippi State University
- **CMU Sphinx**: ancora in fase di sviluppo ma già opensource
- **IBM ViaVoice**: offre funzioni di dettato e di comando vocale (commerciale)
- **Abbot**: applicazione sviluppata all'Università di Cambridge (resa commerciale in seguito)

RT(1/2)

La “Rich Transcription” è una serie di valutazioni di ASR creata dal “National Institute of Standards and Technology” (NIST) per promuovere e stimare lo stato dell’arte in alcune tecnologie dello Speech To Text.
Lo scopo di questa serie di valutazioni è creare tecnologie di riconoscimento che producano trascrizioni più leggibili da umani e più utili per le macchine

Verranno illustrate queste valutazioni per osservare l’attuale stato dell’arte nel settore

RT(2/2)

La serie di valutazioni è iniziata nel 2002 e continua fino ad oggi
L’ultima valutazione, la "Rich Transcription Spring 2007 Meeting Recognition", è stata condotta nei mesi di Febbraio-Luglio del 2007
La comunità che si occupa del riconoscimento del parlato, sta espandendo gli scopi della valutazione del RT per includere ricerche che includano l’audio ed il video

Verrà illustrata la RT 2006 (quella del 2007 è attualmente in corso)

Task

La valutazione è stata effettuata su diversi task:

- **Speech To Text**: convertire parlato in uno stream di testo
- **Speaker Diarization**: processo di annotazione di un input audio con informazioni che attribuiscono ad un certo segnale audio la sua sorgente. Queste sorgenti possono includere speaker, musica, rumore di sottofondo
- **Speech Activity Detection**: individua quando almeno una persona nel meeting sta parlando (e quando nessuno parla)

Ogni partecipante è libero di partecipare ad ognuno (o almeno uno) di questi task

Input(1/3)

L’input dei sistemi è dato da diverse registrazioni di meeting avvenuti in differenti siti

L’input è stato suddiviso in 2 domini principali:

- **conference room**: i dati consistono in 180 minuti di registrazioni di 10 meeting collezionati in 6 siti differenti. Sono stati usati passaggi di 12 minuti selezionati da ogni meeting
- **lecture room**: 180 minuti di dati registrati durante alcuni meeting

Da ogni meeting delle lecture room, sono stati selezionati momenti di esposizione e momenti di domanda/risposta

Input(2/3)

Ci sono diverse condizioni su cui valutare l’input dei sistemi (condizioni in cui è stato registrato l’audio)

Queste condizioni includono l’audio da :

- **Microfoni Distanti Multipli (MDM)**: 3 o più microfoni posizionati centralmente su un tavolo
- **Array di microfoni (MSLA)**: sono stati usati Multiple Mark III digital microphone Arrays, microfoni digitali a 64 canali
- **Microfoni Distanti (ADM)**: microfoni posizionati in punti di interesse (a secondo del luogo in cui si è svolto il meeting)

Input(3/3)

Condizioni di contrasto (da contrappore alle valutazioni precedenti per ottenere una valutazione migliore):

- **Microfoni Singoli distanti (SDM)**
- **Microfono Individuale (IHM)**
  - performance sul parlato “pulito”
  - simile al parlato delle conversazioni telefoniche (un parlante, parlato spontaneo)

Negli ultimi anni l’unica lingua consentita è l’inglese

Output(1/2)

L'output del sistema è un file di tipo .CTM.
Un file di tipo CTM è un tipo di file "token-based" ed include le seguenti informazioni per ogni token riconosciuto:

- il nome del file sorgente
- l'inizio , la fine e la durata del riconoscimento di un token
- la rappresentazione del token riconosciuto (in Ascii o Unicode)
- una probabilità di confidenza

Output(2/2)

Inoltre va inserito il tipo del token:

- lessicale
- non lessicale (rumore)
- lessicale straniero (in un'altra lingua differente da quella inglese)
- pause

- Possono essere generate informazioni sullo speaker (identificatore dello speaker) altrimenti può essere usato un valore "unknow"
- La descrizione della stringa è definita secondo il formato SNOR per effettuare una valutazione uniforme per tutti i sistemi

SNOR

Tutti i token devono essere generati secondo le seguenti regole, nel formato SNOR (Standard Normal Orthographic Representation):

- Separati da spazi
- Case insensitive
- Lettere singole sono rappresentate con la lettera seguita da un segno d'interpunzione (ad esempio: "a. b. c.")
- Non sono rappresentati caratteri non alfabetici
- Le parole con un trattino sono divise nelle loro parti costituenti
- Parti mancanti del parlato sono rappresentate con un trattino

Valutazione del processo

- Il calcolo dell'accuratezza è svolto secondo una versione del WER (orientato a segmenti di parlato) sfruttando la rappresentazione dello SNOR
- Non ci sono limitazioni al tempo di elaborazione del parlato
- Ogni partecipante deve partecipare almeno ad un task (parleremo solo di quelli che partecipano al primo task)

Partecipanti

Partecipano alla valutazione :

- Karlsruhe University (UKA)
- Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur (LIMSI)
- Augmented Multiparty Interaction Program (AMI)
- IBM
- International Computer Science Institute and SRI International (ICSI/RSI)

Verranno illustrati solo due sistemi (i primi due dell'elenco) per dare un'idea di quanto il modello visto fino ad adesso corrisponda alla realtà.

Sistema 1(1/3)

Sviluppato dalla Karlsruhe University (UKA) come un sistema multilivello
Il sistema ha due frontend :

- il primo basato sul Mel-frequency cepstral coefficients (MFCC)
- il secondo basato su MVDR (minimum variance distortionless response), una tecnica di analisi del segnale basata su Linear Predictive Coding

I risultati all'uscita dei 2 frontend vengono "fusi" e passati al livello successivo

Non vengono usati ulteriori filtri durante l'acquisizione perchè se ne usa 20 (invece che 13) nel MFCC

Sistema 1(2/3)

- La stima della probabilità avviene con una variante delle HMM (la variante serve per diminuire lo spazio delle probabilità)
- Il training è stato effettuato su un corpus di 10 ore in lingua inglese
- Nella fase di decodifica viene usata un grammatica di riferimento generata a partire da una serie di corpora scelti appositamente ed in parte selezionati da web (attraverso delle query)

Sistema 1(3/3)

- Per la generazione è stato usato "SRILM toolkit" un tool (attualmente in fase di sviluppo ma già disponibile e funzionante in modalità opensource) per la generazione di grammatiche
- I termini riconosciuti dal sistema sono dati dalla fusione di 2 dizionari (Callhome English e LIMSI SI-284)
- L'errore medio stimato su diverse misurazioni era del 32% (calcolato con il wer)

Sistema 2(1/2)

- Sviluppato dal "Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur " (LIMSI)
- Ha un frontend basato su Mel-frequency cepstral coefficients (MFCC)
- Nella fase di stima è stata usata una rete neurale: sono stati usati algoritmi efficienti per addestrare ed usare la rete neurale (altrimenti il suo utilizzo sarebbe inefficiente)
- L'addestramento è stato realizzato su 97h di registrazioni da 6 sorgenti

Sistema 2(2/2)

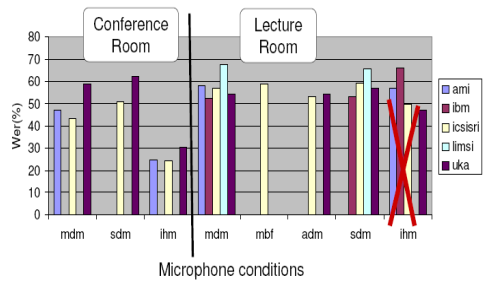
La fase di decodifica è stata particolarmente curata, oltre ad essere stata generata una grammatica (4-gram generata con un HMM):

- la decodifica viene verificata anche attraverso una probabilità di pronuncia

Una N-gram è una grammatica stocastica nella quale la probabilità di un occorrenza di un simbolo è condizionata alla probabilità di occorrenza degli altri N-1 simboli

L'addestramento è stato fatto su fatto su 20k articoli (circa 40M parole)

Risultati Finali



Commenti

- La percentuale di errore è risultata più alta degli altri anni perchè l' RT 2006 è stato più difficile in termine di acustica (ma non di linguaggio), è da notare infatti, che l'errore cresce sensibilmente con l'aumentare della distanza dei microfoni dalla sorgente (ad esempio ihm e mdm)
- In tutti i casi i risultati ottenuti dai sistemi sono stati inferiori alle aspettative (ed ai risultati ottenuti nei test in laboratorio dai vari gruppi), questo ad indicare la difficoltà del riconoscimento (e della valutazione)
- Le valutazioni dell'IHM nelle Lecture Room non sono state considerate valide

TCStar wokshop(1/2)

- Il TC-STAR è un progetto Europeo focalizzato sul Speech-to-Speech Translation (SST)
- Il progetto punta a domini di conversazioni, senza vincoli sull'argomento, in 3 linguaggi: Inglese, Spagnolo e Cinese
- Il worshop tratta i seguenti domini:
  - traduzioni automatica di testo/parlato
  - riconoscimento e sintesi del parlato
  - aspetti e valutazione di sistemi
- Lo scopo del workshop è di far lavorare insieme i ricercatori che si occupano di qualsiasi aspetto delle traduzioni da parlato a parlato e confrontare gli approcci innovativi del TCStar con quelli degli altri progetti di ricerca in corso

TCStar wokshop(2/2)

- Verrà illustrato uno screenshot di uno dei sistemi che hanno partecipato al TCStar utile a comprendere come lavorano questi sistemi
- A partire dall'alto, è possibile vedere:
  - il video della conferenza (di cui si effettua in real time la traduzione)
  - il testo riconosciuto
  - il segmento di testo corrente (quello con cui confrontare il testo riconosciuto)
  - il testo tradotto
- Non è stato analizzato il TCStar perchè simile sotto certi aspetti all'RT, anche se è stato introdotto per completezza e per far comprendere l'attuale interesse che sta suscitando il settore del Text To Speech

Conclusioni(1/2)

Abbiamo visto:

- problemi dell'analisi del parlato
- cosa è e come costruire un'asr
- alcuni sistemi in circolazione

- Allo stato attuale dell'arte si può notare un proliferare di sistemi che adottano diversi metodi e tecniche (alcune che tendono a sovrapporsi nelle varie fasi di elaborazione, nel tentativo di migliorare i risultati del riconoscimento) ma che tuttavia rimangono aderenti all'architettura presentata (se pur semplificata)

Conclusioni(2/2)

- Nonostante non sia stato analizzato il TCStar da un punto di vista tecnico, ci è comunque servito per capire l'interesse che il settore sta generando
- Il numero sempre maggiore di manifestazioni e progetti rivela quanto sia alto il grado di attenzione relativo a questo settore ed alle sue tecnologie

Sitografia

- http://www.ibiblio.org/pub/linux/docs/HOWTO/Speech-Recognition-HOWTO
- http://nist.gov/speech/tests/rt/index.htm
- http://en.wikipedia.org/wiki/Speech_recognition
- http://www.w3.org/TR/ngram-spec/
- http://www.hackerart.org/corsi/intromedia.htm#media

Bibliografia

- Riconoscimento Automatico del Parlato: un Approccio Sillabico (Gianpaolo Coro: Tesi di laurea)