COMUNICAZIONE VERBALE

Introduzione

La comunicazione verbale (basata sul parlato) occupa un ruolo importante nell'ambito della MMI (Man-Machine Interaction) avanzata.

La sostituzione di dispositivi convenzionali di input come pulsanti e tastiere con il controllo vocale può incrementare il comfort e la velocità di trasmissione degli input.

Inoltre, il parlato è una delle modalità preferite nell'interazione multimodale, ad esempio come modalità complementare ai gesti.

Infine, il parlato non contiene solo informazioni linguistiche ma anche informazioni sulle emozioni, le quali rappresentano un ulteriore aspetto importante della MMI avanzata.

COMUNICAZIONE VERBALE

Riconoscimento del parlato – Concetti generali

Per Automatic Speech Recognition (ASR) si intende la generazione di una stringa di parole mediante l'acquisizione di un input audio da un dispositivo (ad esempio uno smartphone). Il compito di un sistema ASR è quello di mappare una qualsiasi forma d'onda

nell'appropriata stringa di parole

It's time for lunch!

I sistemi Text-To-Speech (TTS) svolgono invece il task opposto rispetto ai sistemi ASR. I sistemi TTS mappano testo in una forma d'onda sonora.

COMUNICAZIONE VERBALE

Riconoscimento del parlato – Concetti generali

Prima di descrivere alcuni algoritmi utilizzati per il riconoscimento del parlato in sistemi ASR, risulta utile una descrizione delle caratteristiche del compito ASR.

Una prima caratteristica è rappresentata dalla dimensione del vocabolario. Alcuni compiti ASR possono essere risolti con un'accuratezza estremamente elevata, come ad esempio i compiti basati su un vocabolario di 2 parole (*yes*, *no*) o su un vocabolario di 10 parole (come ad esempio il compito di *digit recognition* che consiste nel riconoscimento di sequenze di numeri composti dalle cifre 0, …, 9). Altri compiti ASR riferiti ad un vocabolario di dimensione maggiore (ad esempio più di 60000 parole) sono ovviamente più difficili.

Una seconda caratteristica è rappresentata dall'interlocutore. È più semplice riconoscere il parlato degli umani alle macchine (dettando o parlando ad un *dialogue system*) rispetto al riconoscere il parlato degli umani verso altri umani. La lettura di un discorso (*read speech*), nel quale gli umani leggono ad alta voce, come ad esempio in audiolibri, è anch'essa relativamente semplice da riconoscere. Il riconoscimento del parlato di due umani coinvolti in una conversazione (ad esempio per la trascrizione di un meeting) è il compito più difficile da risolvere.

COMUNICAZIONE VERBALE

Riconoscimento del parlato – Concetti generali

Infatti sembra che quando gli esseri umani parlano con le macchine, o leggono senza la presenza di un pubblico, semplifichino il loro discorso, parlando più lentamente e in modo più chiaro.

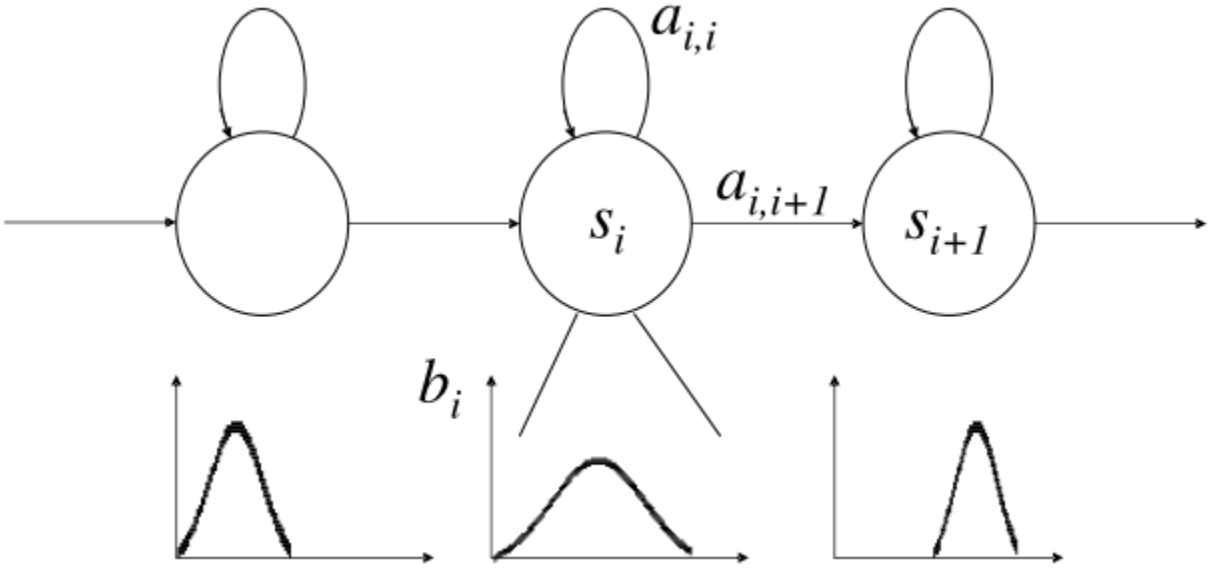
Una terza caratteristica è rappresentata dal canale e dal rumore. Il parlato è più facile da riconoscere se esso è registrato in una stanza silenziosa con microfoni montati sulla testa rispetto al caso in cui esso sia registrato da un microfono distante in una strada rumorosa o in una macchina con il finestrino aperto.

Una quarta caratteristica è legata all'accento e alla lingua dello speaker. Il parlato è più semplice da riconoscere se il sistema di riconoscimento viene utilizzato nelle stesse condizioni in cui è stato allenato.

COMUNICAZIONE VERBALE

Riconoscimento del parlato – Approccio basato su HMMs

Un approccio che può essere utilizzato per il riconoscimento del parlato è quello basato su *Hidden Markov Models* (*HMMs*). L'ipotesi fondamentale di tale approccio è il fatto che ogni suono vocale per un determinato linguaggio (più comunemente chiamato fonema) venga rappresentato con un *HMM*. Un *HMM* può essere definito come una macchina a stati finiti stocastica.



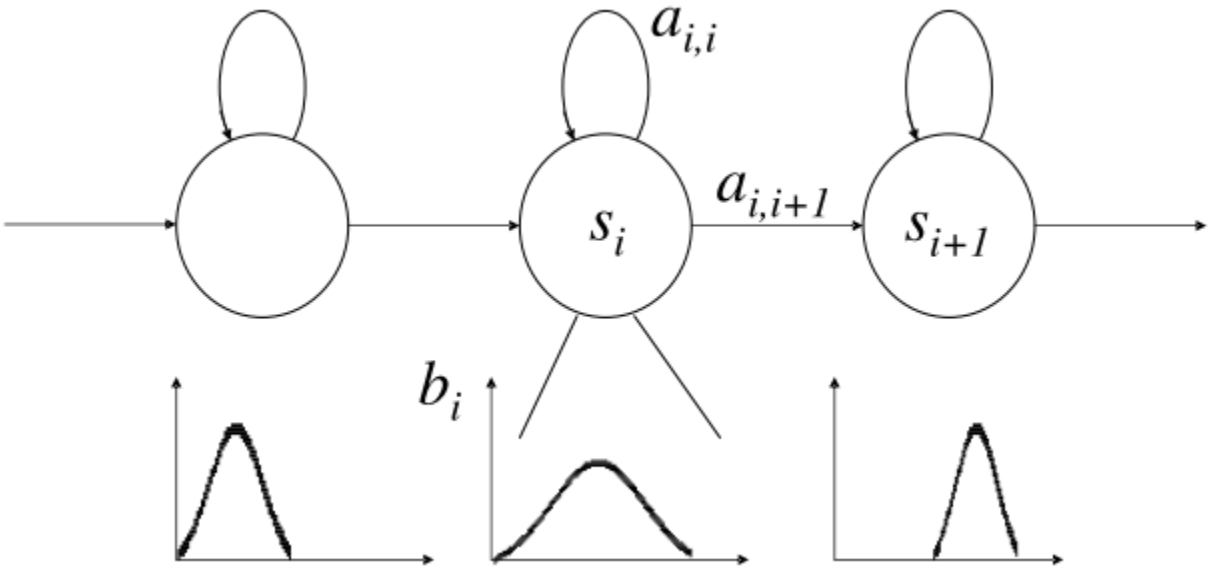
Esempio di Hidden Markov Model.

COMUNICAZIONE VERBALE

Riconoscimento del parlato – Approccio basato su HMMs

Un *HMM* è caratterizzato da un numero finito di stati e da possibili transizioni tra questi stati.

Il processo di produzione del parlato viene considerato come una sequenza di eventi acustici discreti, dove tipicamente ogni evento è caratterizzato dalla produzione di un *feature vector* che descrive il segnale vocale prodotto all'istante di tempo discreto equivalente del processo di produzione del parlato.

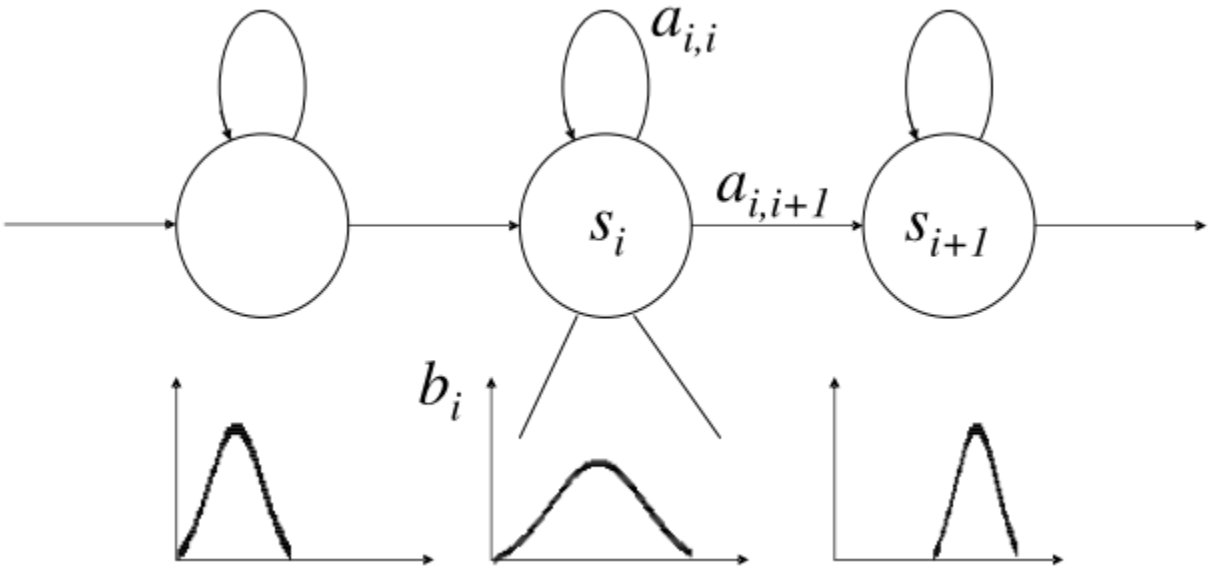


Esempio di Hidden Markov Model.

COMUNICAZIONE VERBALE

Riconoscimento del parlato – Approccio basato su HMMs

Si ipotizzi che il *HMM* sia in uno dei suoi stati ad ognuno dei passi di tempo discreto considerati; al passo successivo, il *HMM* eseguirà una transizione in uno stato che può essere raggiunto dal suo stato attuale in base alla topologia scelta (possono essere incluse anche transizioni verso lo stato attuale o verso gli stati precedentemente visitati).

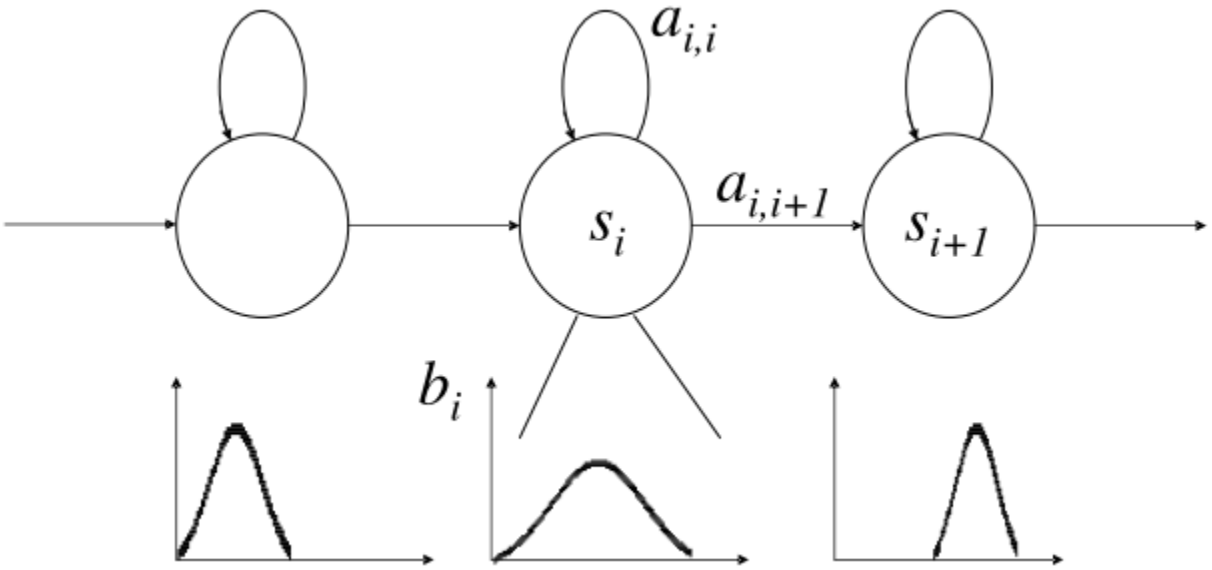


Esempio di Hidden Markov Model.

COMUNICAZIONE VERBALE

Riconoscimento del parlato – Approccio basato su HMMs

Nella figura è riportato un esempio di *HMM*. Esso è caratterizzato da due insiemi principali di parametri: il primo insieme è la matrice delle probabilità di transizione (transition probabilities), la quale descrive la probabilità $p(s(k-1) \rightarrow s(k))$, dove *k* è l'indice di tempo discreto e *s* è la notazione utilizzata per indicare uno stato. Tali probabilità sono rappresentate dai parametri *a* nella figura. Esse sono probabilità condizionate.

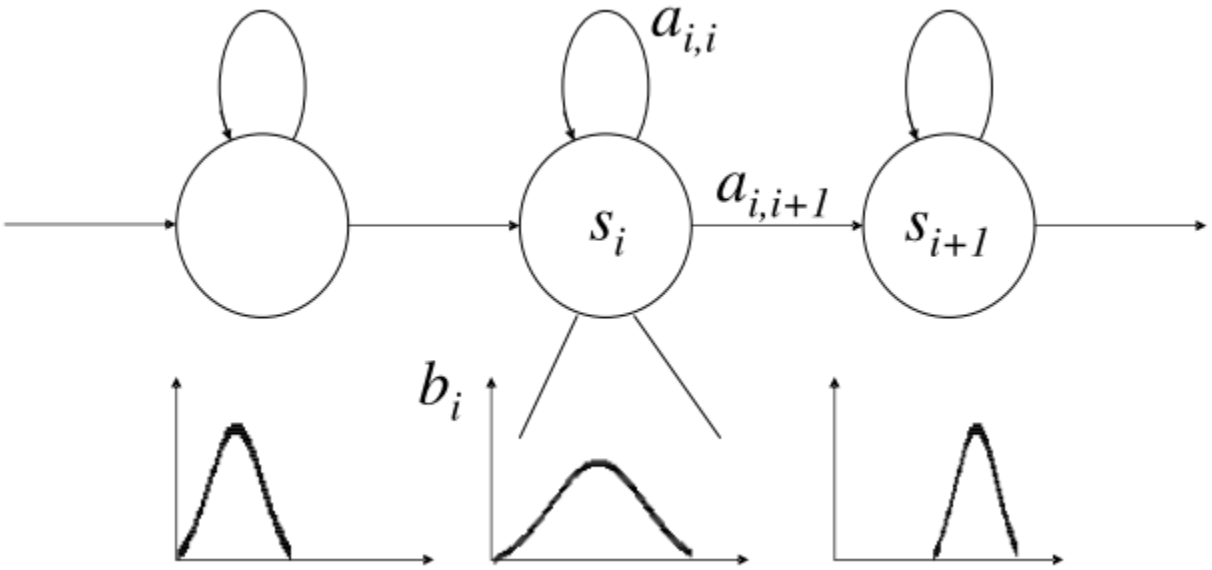


Esempio di Hidden Markov Model.

COMUNICAZIONE VERBALE

Riconoscimento del parlato – Approccio basato su HMMs

Il secondo insieme rappresenta le probabilità di emissione (emission probabilities) $p(x|s(k))$, le quali indicano la probabilità che un certo *feature vector* x si possa presentare mentre il *HMM* si trova nello stato s al tempo k. Tale probabilità è di solito espressa da una funzione di distribuzione continua (nella figura, b_i indica una funzione densità). Tale funzione di distribuzione è in molti casi una combinazione di distribuzioni Gaussiane.

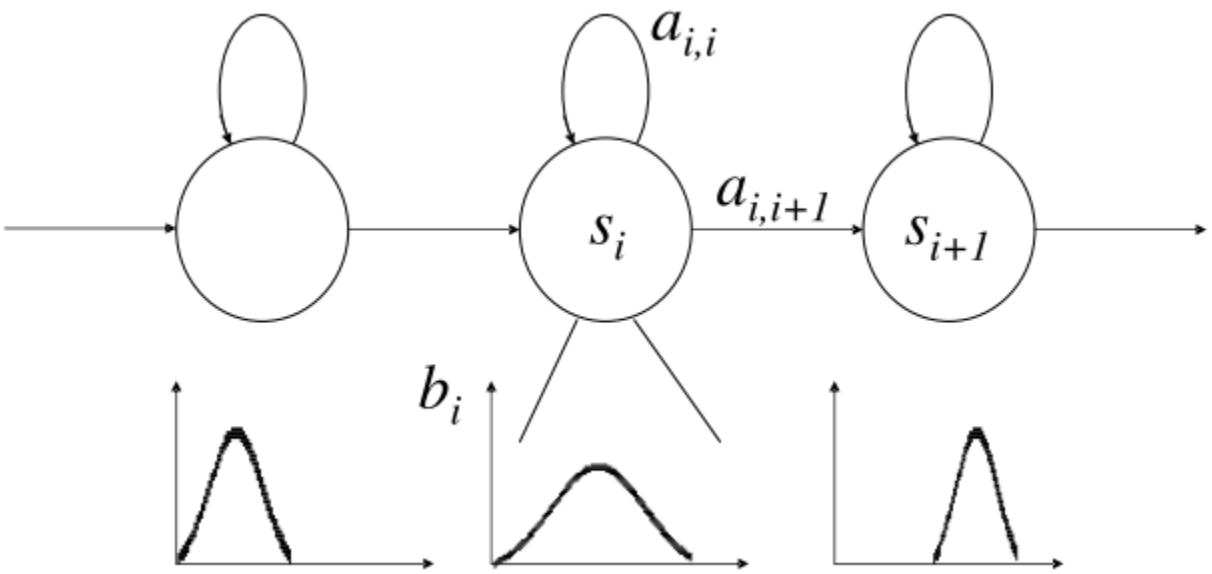


Esempio di Hidden Markov Model.

COMUNICAZIONE VERBALE

Riconoscimento del parlato – Approccio basato su HMMs

Se si ipotizza una transizione in un altro stato al passo *k* nel quale si presenta un *feature vector x(k)*, è molto probabile che questa transizione avvenga in uno stato con un'alta probabilità di emissione per *x(k)*, cioè in uno stato che rappresenti adeguatamente le caratteristiche del *feature vector* considerato.

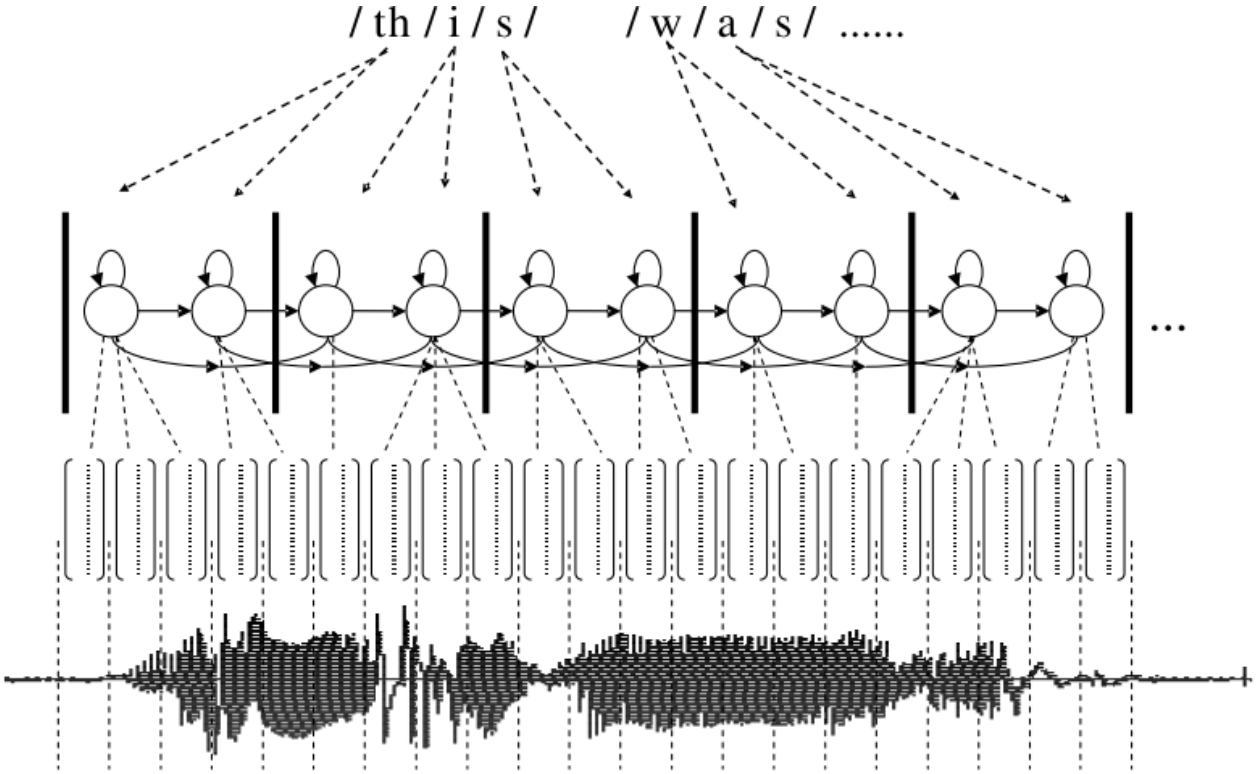


Esempio di Hidden Markov Model.

COMUNICAZIONE VERBALE

Riconoscimento del parlato – Approccio basato su HMMs

Nella parte inferiore della figura, è riportato il segnale vocale corrispondente alla sequenza di parole «this was». Tale sequenza è riportata nella parte superiore della figura. Nell'esempio riportato, il segnale è stato suddiviso in finestre di tempo di lunghezza costante (ad esempio 10 ms) e per ogni finestra è stato generato un *feature vector*.

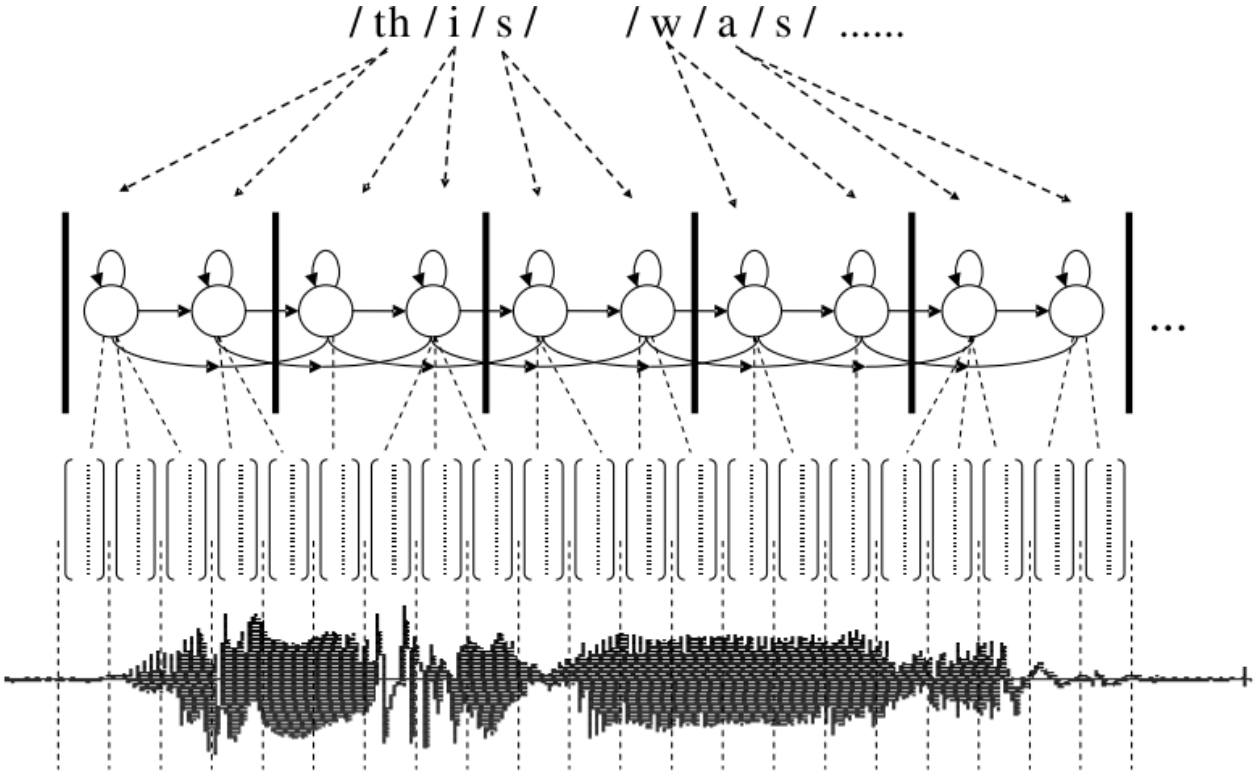


Esempio di riconoscimento del parlato basato su HMMs.

COMUNICAZIONE VERBALE

Riconoscimento del parlato – Approccio basato su HMMs

La generazione dei *feature vectors* può essere eseguita applicando una trasformazione in frequenza al segnale (esempio: trasformata di Fourier). Si ottiene quindi una sequenza di vettori (rappresentata sopra il segnale vocale). Tale sequenza viene quindi «osservata» dal *HMM* riportato nella figura. Tali vettori rappresentano le osservazioni.



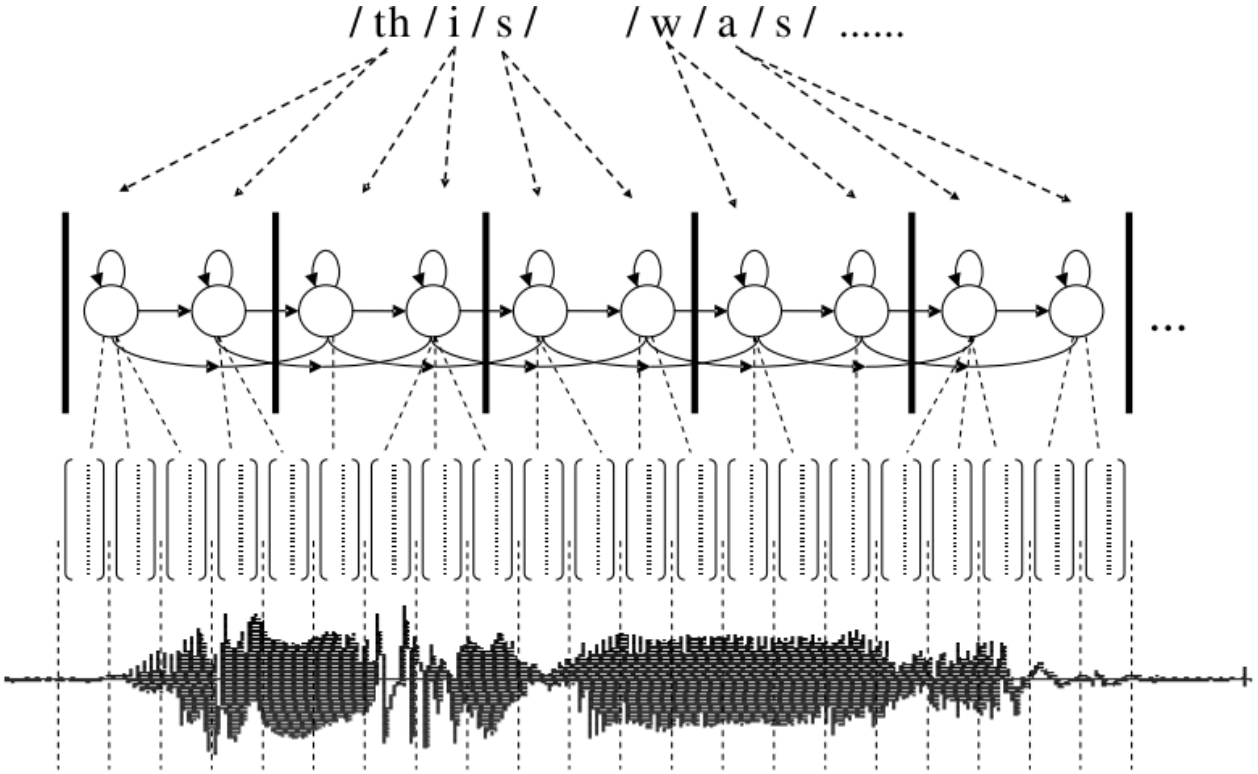
Esempio di riconoscimento del parlato basato su HMMs.

COMUNICAZIONE VERBALE

Riconoscimento del parlato – Approccio basato su HMMs

Il *HMM* complessivo è stato generato rappresentando ogni fonema della sequenza di parole mediante un *HMM* a 2 stati e concatenando gli *HMMs* ottenuti.

Un concetto importante è rappresentato dalle linee nere tratteggiate riportate nella figura. Esse indicano l'associazione di ogni *feature vector* ad uno stato specifico del *HMM*.



Esempio di riconoscimento del parlato basato su HMMs.

COMUNICAZIONE VERBALE

Riconoscimento del parlato – Approccio basato su HMMs

Si hanno quindi tante associazioni quanti sono i vettori nella sequenza di *feature vectors*

$$X = [x(1), x(2), ..., x(K)]$$

dove *K* è il numero di vettori. Tale concetto di associazione dei *feature vectors* agli stati (state-alignment) è una delle capacità principali degli *HMMs*; esistono diversi algoritmi per il calcolo di tali associazioni. Il principio di base di tale procedura di allineamento può essere spiegato considerando una singola transizione del *HMM* al passo *k* dallo stato *s(k-1)* allo stato *s(k)* e ipotizzando che allo stesso istante di tempo si presenti il *feature vector x(k)*.

Si può calcolare la probabilità congiunta di tale evento, cioè la probabilità che i due eventi descritti si verifichino congiuntamente:

$$\begin{aligned} p(x(k), s(k-1) \rightarrow s(k)) &= p(x(k)|s(k-1) \rightarrow s(k)) \cdot p(s(k-1) \rightarrow s(k)) \\ &= p(x(k)|s(k)) \cdot p(s(k-1) \rightarrow s(k)) \end{aligned}$$

COMUNICAZIONE VERBALE

Riconoscimento del parlato – Approccio basato su HMMs

Si hanno quindi tante associazioni quanti sono i vettori nella sequenza di *feature vectors*

$$X = [x(1), x(2), ..., x(K)]$$

dove *K* è il numero di vettori. Tale concetto di associazione dei *feature vectors* agli stati (state-alignment) è una delle capacità principali degli *HMMs*; esistono diversi algoritmi per il calcolo di tali associazioni. Il principio di base di tale procedura di allineamento può essere spiegato considerando una singola transizione del *HMM* al passo *k* dallo stato *s(k-1)* allo stato *s(k)* e ipotizzando che allo stesso istante di tempo si presenti il *feature vector x(k)*.

Si può calcolare la probabilità congiunta di tale evento, cioè la probabilità che i due eventi descritti si verifichino congiuntamente:

$$p(x(k), s(k-1) \rightarrow s(k)) = p(x(k)|s(k-1) \rightarrow s(k)) \cdot p(s(k-1) \rightarrow s(k))$$
$$= p(x(k)|s(k)) \cdot p(s(k-1) \rightarrow s(k))$$

si ricordi la definizione di probabilità condizionata $\mathrm{P}(B \mid A) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(A)}$

COMUNICAZIONE VERBALE

Riconoscimento del parlato – Approccio basato su HMMs

Si hanno quindi tante associazioni quanti sono i vettori nella sequenza di *feature vectors*

$$X = [x(1), x(2), ..., x(K)]$$

dove *K* è il numero di vettori. Tale concetto di associazione dei *feature vectors* agli stati (state-alignment) è una delle capacità principali degli *HMMs*; esistono diversi algoritmi per il calcolo di tali associazioni. Il principio di base di tale procedura di allineamento può essere spiegato considerando una singola transizione del *HMM* al passo *k* dallo stato *s(k-1)* allo stato *s(k)* e ipotizzando che allo stesso istante di tempo si presenti il *feature vector x(k)*.

Si può calcolare la probabilità congiunta di tale evento, cioè la probabilità che i due eventi descritti si verifichino congiuntamente:

$$p(x(k), s(k-1) \rightarrow s(k)) = p(x(k)|s(k-1) \rightarrow s(k)) \cdot p(s(k-1) \rightarrow s(k))$$
$$= p(x(k)|s(k)) \cdot p(s(k-1) \rightarrow s(k))$$

ipotesi semplificativa: la probabilità di un'osservazione dipende solo dallo stato che la ha generata

COMUNICAZIONE VERBALE

Riconoscimento del parlato – Approccio basato su HMMs

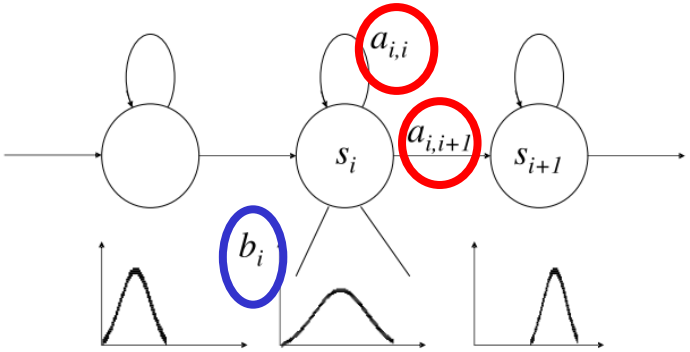
Si hanno quindi tante associazioni quanti sono i vettori nella sequenza di *feature vectors*

$$X = [x(1), x(2), ..., x(K)]$$

dove *K* è il numero di vettori. Tale concetto di associazione dei *feature vectors* agli stati (state-alignment) è una delle capacità principali degli *HMMs*; esistono diversi algoritmi per il calcolo di tali associazioni. Il principio di base di tale procedura di allineamento può essere spiegato considerando una singola transizione del *HMM* al passo *k* dallo stato *s(k-1)* allo stato *s(k)* e ipotizzando che allo stesso istante di tempo si presenti il *feature vector x(k)*.

Si può calcolare la probabilità congiunta di tale evento, cioè la probabilità che i due eventi descritti si verifichino congiuntamente:

$$p(x(k), s(k-1) \rightarrow s(k)) = p(x(k)|s(k-1) \rightarrow s(k)) \cdot p(s(k-1) \rightarrow s(k))$$
$$= p(x(k)|s(k)) \; p(s(k-1) \rightarrow s(k))$$



COMUNICAZIONE VERBALE

Riconoscimento del parlato – Approccio basato su HMMs

Si consideri l'equazione

$$p(x(k), s(k-1) \rightarrow s(k)) = p(x(k)|s(k-1) \rightarrow s(k)) \cdot p(s(k-1) \rightarrow s(k))$$
$$= p(x(k)|s(k)) \cdot p(s(k-1) \rightarrow s(k))$$

Come già affermato in precedenza, una transizione in uno degli stati successivi sarà molto probabile, il che si traduce in un'alta probabilità congiunta, espressa nella formula precedente.

Si può quindi pensare alla progettazione di un algoritmo per il calcolo della sequenza di stati più probabile considerando l'associazione al *feature vector* osservato. Tale algoritmo deve restituire la sequenza di stati che possa garantire la massimizzazione del prodotto di tutte le probabilità calcolate mediante l'equazione precedente per tutti i *K feature vectors*.

COMUNICAZIONE VERBALE

Riconoscimento del parlato – Approccio basato su HMMs

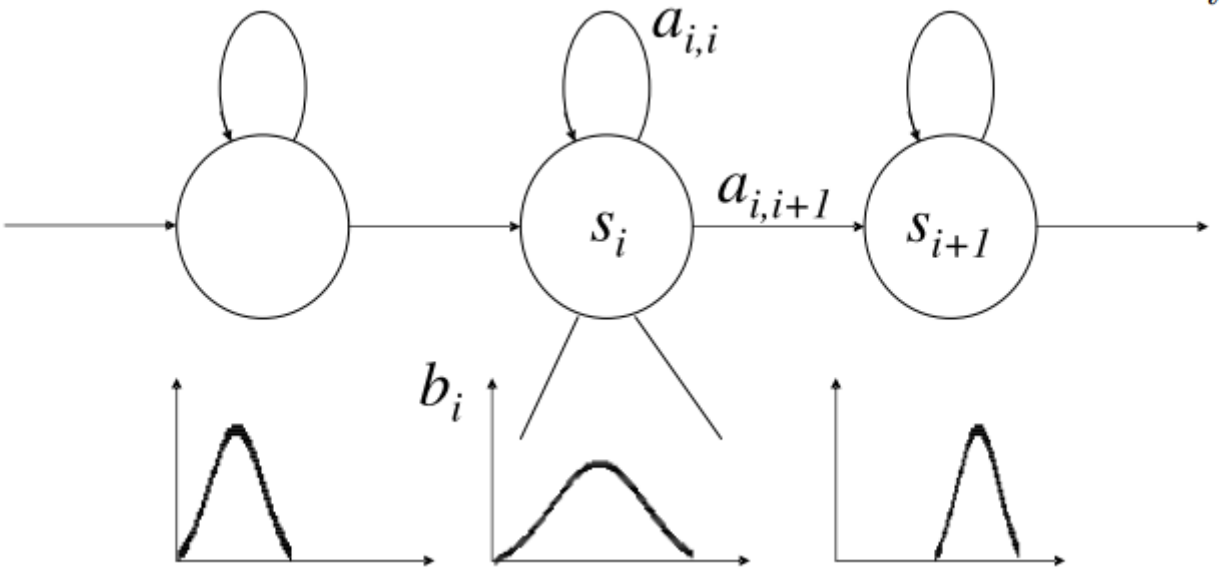
Un *HMM* λ composto da *N* stati è completamente specificato da $\lambda = (\boldsymbol{\pi}, \mathbf{A}, \mathbf{B})$ dove:

- $A = [a_{i,j}]_{N \times N}$ con $\sum_j a_{i,j} = 1$ (nel caso specifico riportato in figura si ha inoltre $a_{i,j} = 0$ per $j \notin \{i, i+1\}$)

matrice delle probabilità di transizione

$$a_{i,j} = P(q_t = s_i \,|\, q_{t-1} = s_j)$$

stato al tempo *t*



Esempio di Hidden Markov Model.

Proprietà di Markov

$$P(q_i|q_1, \ldots, q_{i-1}) = P(q_i|q_{i-1})$$

COMUNICAZIONE VERBALE

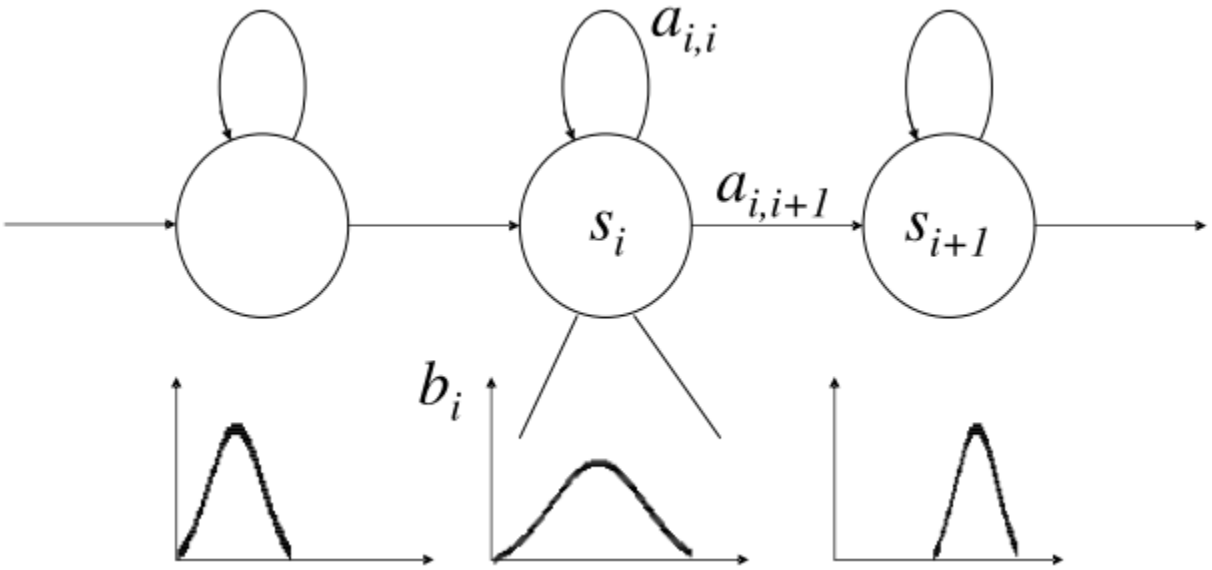
Riconoscimento del parlato – Approccio basato su HMMs

Un *HMM* λ composto da N stati è completamente specificato da $\lambda = (\boldsymbol{\pi}, \mathbf{A}, \mathbf{B})$ dove:

- $\mathbf{B} = b_i\big(x(o)\big) \quad \begin{matrix} i = 1 \ldots N \\ o = 1 \ldots K \end{matrix}$ sequenza delle probabilità di emissione

- $\boldsymbol{\pi} = (\pi_1 \; \pi_2 \; \ldots \; \pi_N)$ con $\pi_i = P(q_1 = s_i)$ e $\sum_i \pi_i = 1$

  vettore della probabilità di stato iniziale — stato iniziale

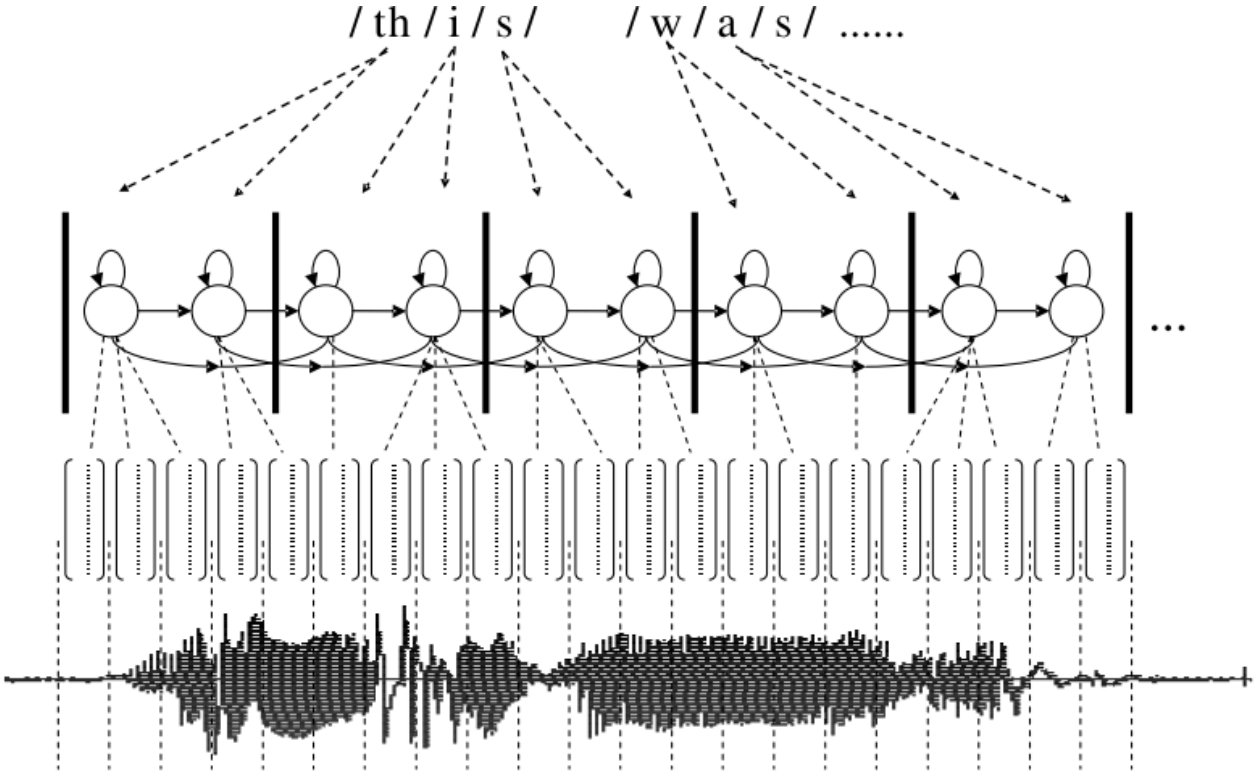


Esempio di Hidden Markov Model.

COMUNICAZIONE VERBALE

Riconoscimento del parlato – Approccio basato su HMMs

Si ipotizzi che i parametri dei *HMMs* siano noti e che sia stata determinata la sequenza ottima di stati (come mostrato dalle linee nere tratteggiate nella figura; tali linee rappresentano l'associazione di ogni *feature vector* allo stato più probabile). In questo caso tale approccio ha prodotto due risultati principali.



Esempio di riconoscimento del parlato basato su HMMs.

COMUNICAZIONE VERBALE

Riconoscimento del parlato – Approccio basato su HMMs

Il primo risultato è un risultato di segmentazione che assegna ogni vettore a uno stato. Con tale risultato è possibile determinare quali vettori (e quindi quale parte del segnale vocale) sono stati assegnati ad un fonema specifico, ad esempio al suono /i/ riportato nella figura della slide precedente (cioè i vettori nr. 6 – 9).
Il secondo risultato si riferisce alla probabilità menzionata in precedenza. La sequenza di *feature vectors* che rappresenta il segnale vocale è stata assegnata alla sequenza ottima di stati. Poiché tale probabilità è la massima possibile e nessun'altra sequenza di stati produrrà un valore più grande, tale probabilità può essere considerata come la probabilità che il segnale vocale sia stato prodotto dal *HMM* mostrato nella figura della slide precedente.

Quindi, un *HMM* è in grado di processare un segnale vocale e di produrre due importanti risultati: la segmentazione del segnale in unità (ad esempio fonemi) e la probabilità che tale segnale possa essere stato prodotto dal modello probabilistico considerato.

Riferimenti Bibliografici

[1] Kraiss, K. -F. (2006). Advanced Man-Machine Interaction: Fundamentals and Implementation. Springer-Verlag Berlin Heidelberg. ISBN-10: 3-540-30618-8

[2] Daniel Jurafsky, James H. Martin (2024). Speech and Language Processing – An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.