COMUNICAZIONE VERBALE

Riconoscimento del parlato – Approccio basato su HMMs

Nella fase di training, i parametri del *HMM* non sono noti, ma è nota la trascrizione della corrispondente sequenza di parole («this was» nell'esempio). Può essere utilizzata una procedura iterativa per il calcolo dei parametri del *HMM* scelto.

Considerando il *HMM* riportato nell'esempio, possono essere ipotizzati i parametri per iniziare la procedura iterativa. Si calcola quindi la sequenza ottima di stati corrispondente a tali parametri. Quasi sicuramente tale sequenza non sarà la sequenza globalmente ottima, poiché i parametri iniziali potrebbero non essere stati scelti correttamente.
Tuttavia, è possibile derivare una nuova stima dei parametri del *HMM* da tale sequenza di stati. Per quanto riguarda le probabilità di transizione, occorre contare le occorrenze delle transizioni tra gli stati della sequenza di stati calcolata e dividerle per il numero totale delle transizioni.

COMUNICAZIONE VERBALE

Riconoscimento del parlato – Approccio basato su HMMs

Per quanto riguarda le probabilità $p(x(k)|s(k))$, esse possono essere ricavate mediante il calcolo di proprietà (esempi: valor medio e varianza) dei *feature vectors* associati ai differenti stati.

Tale procedura può essere ripetuta: con i parametri aggiornati del *HMM*, può essere calcolata una associazione (state-alignment) aggiornata tra i *feature vectors* e gli stati; poi possono essere utilizzate nuovamente le proprietà della sequenza di stati al fine di aggiornare i parametri del *HMM*.

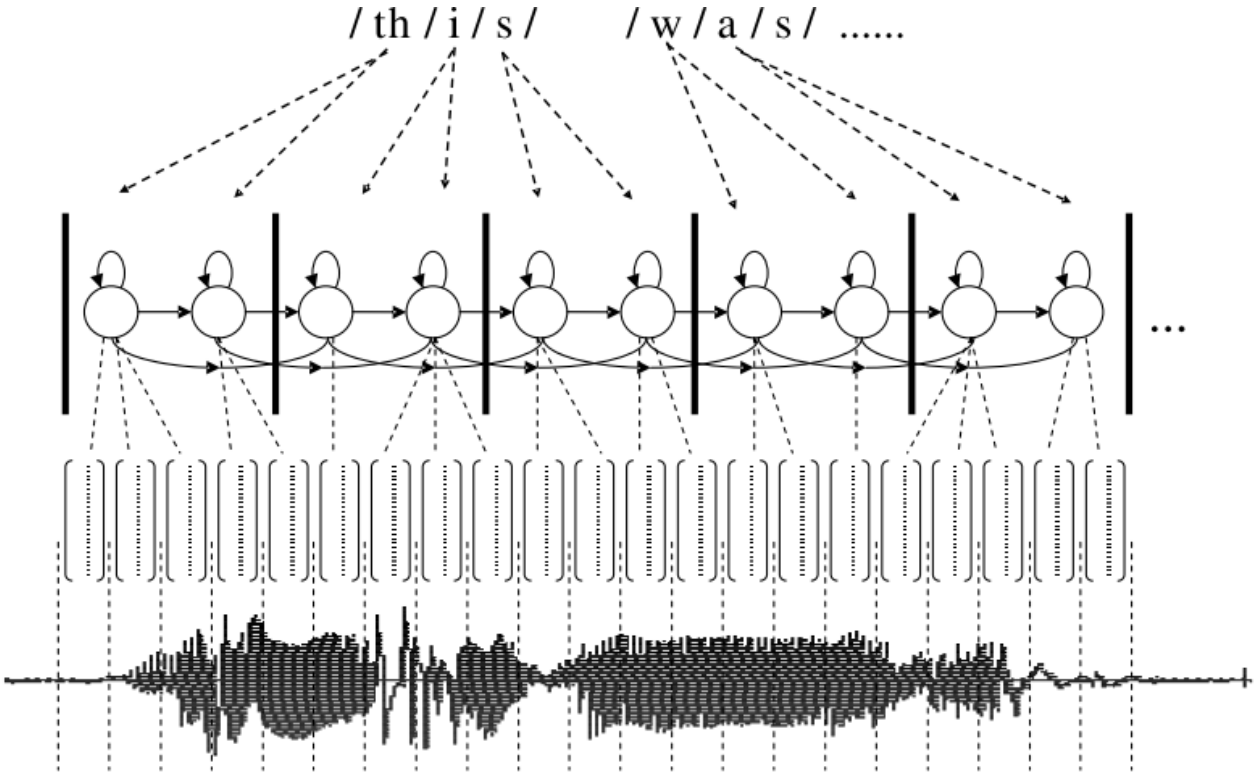
Tipicamente, tale procedura fornisce una stima soddisfacente dei parametri del *HMM* dopo alcune iterazioni. Inoltre, alla fine di tale procedura, si può «tagliare» il *HMM* concatenato, ottenendo un singolo *HMM* per ogni fonema. Ognuno di questi *HMMs* è caratterizzato da un insieme di parametri che approssimano le caratteristiche dei differenti suoni. Quindi, ad esempio, il singolo *HMM* per il fonema /i/ dell'esempio avrà parametri differenti rispetto al *HMM* per il fonema /a/.

La tecnologia basata su *HMMs* permette il training di modelli probabilistici per ogni unità del parlato (tipicamente l'unità «fonema») mediante l'elaborazione di lunghe frasi senza la necessità di etichettare i fonemi o di pre-segmentare le frasi manualmente. Ciò costituisce un enorme vantaggio.

COMUNICAZIONE VERBALE

Riconoscimento del parlato – Approccio basato su HMMs

Nella fase di riconoscimento, i parametri del *HMM* sono noti (dalla fase di training) e la stringa associata al segnale vocale è sconosciuta. Quindi la trascrizione riportata nella figura non è nota e deve essere ricostruita con la procedura di riconoscimento.

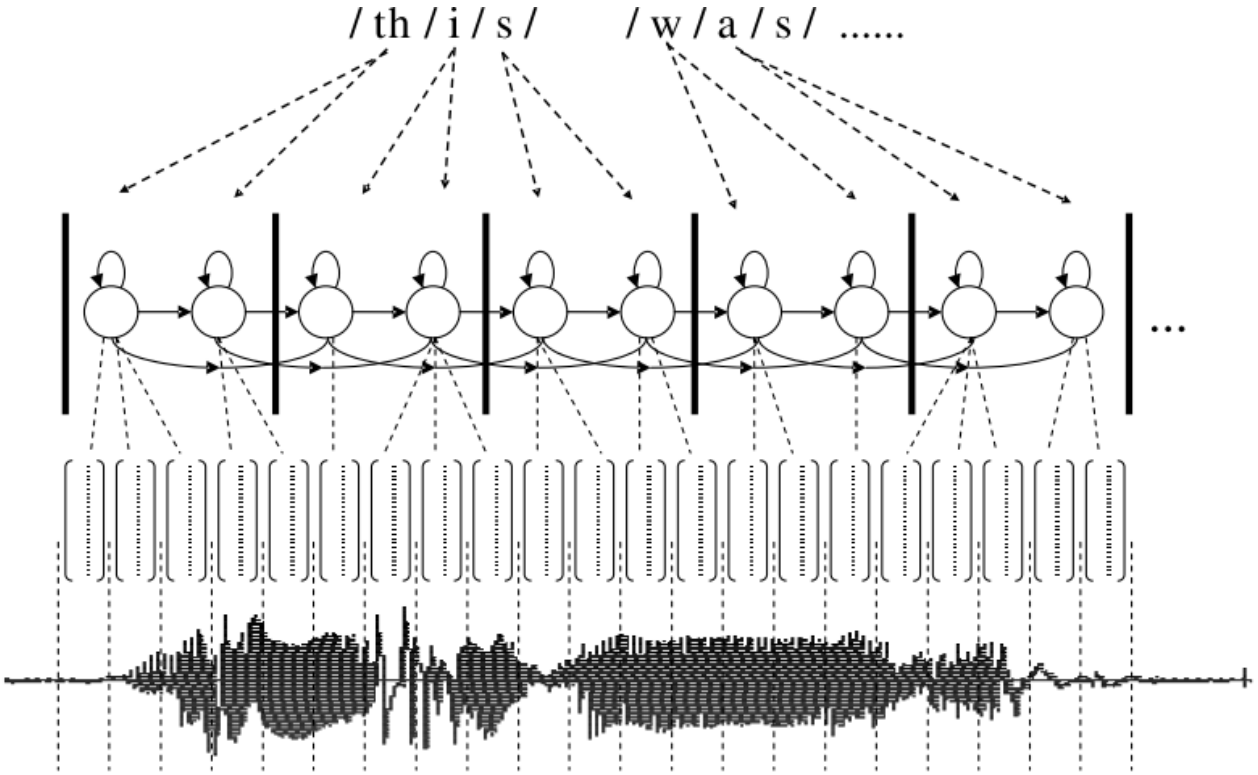


Esempio di riconoscimento del parlato basato su HMMs.

COMUNICAZIONE VERBALE

Riconoscimento del parlato – Approccio basato su HMMs

Esistono algoritmi efficienti per il calcolo dell'associazione (state-alignment) tra la sequenza dei *feature vectors* e gli stati del *HMM* (si osservino le linee nere tratteggiate nella figura). Uno di tali algoritmi è l'algoritmo di Viterbi.



Esempio di riconoscimento del parlato basato su HMMs.

COMUNICAZIONE VERBALE

Riconoscimento del parlato – Approccio basato su HMMs

Il diagramma riportato mostra il tempo con i *feature vectors x* sull'asse orizzontale (in istanti di tempo discreto) e gli stati di un modello *HMM* (a tre stati) sull'asse verticale. Ipotizzando che si parta dal primo stato, il primo *feature vector* viene assegnato a tale stato iniziale. Per quanto riguarda il secondo *feature vector*, in accordo con la topologia del *HMM*, è possibile che il modello rimanga nello stato 1 (eseguendo quindi una transizione dallo stato 1 allo stato 1) o che si muova verso lo stato 2 (eseguendo

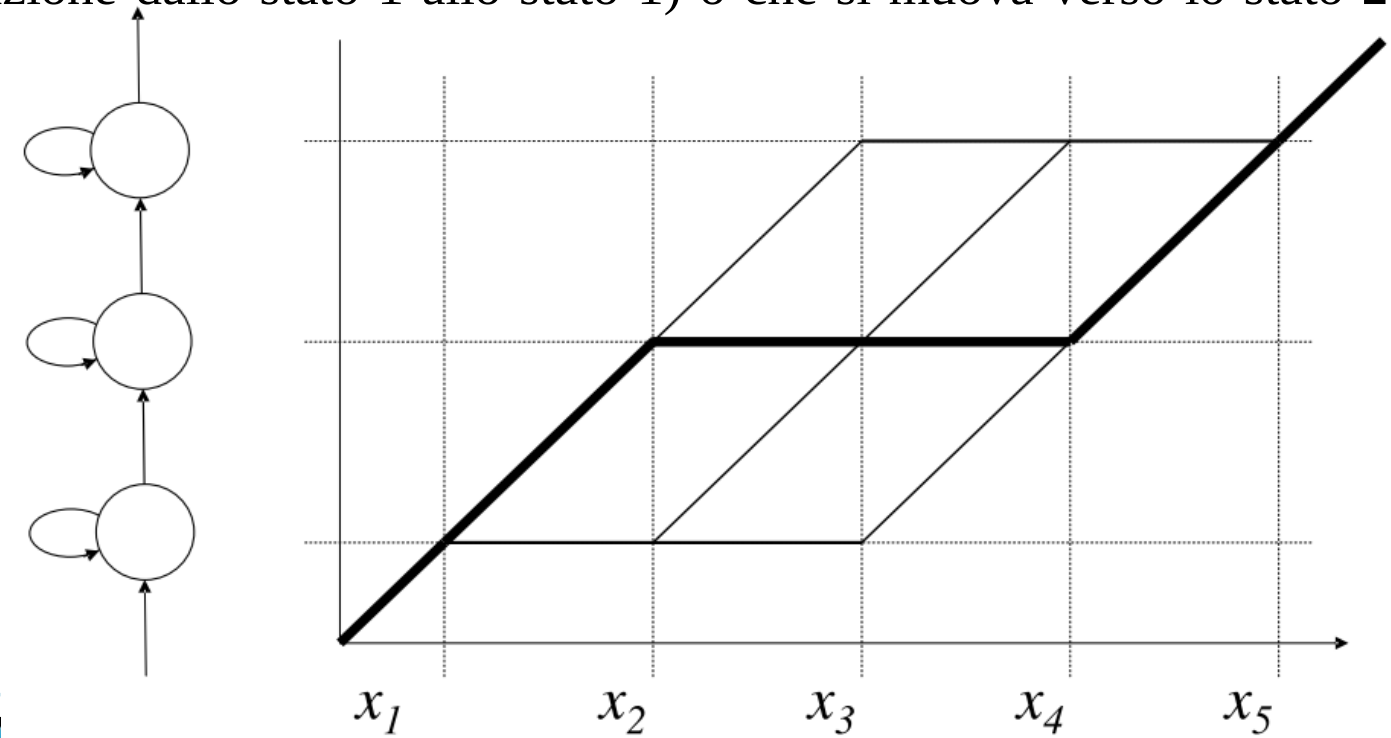


Diagramma a griglia (trellis diagram) per l'algoritmo di Viterbi.

quindi una transizione dallo stato 1 allo stato 2). Per entrambe le opzioni, la probabilità può essere calcolata utilizzando l'equazione

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

Entrambe le opzioni sono mostrate nella figura come possibili percorsi.

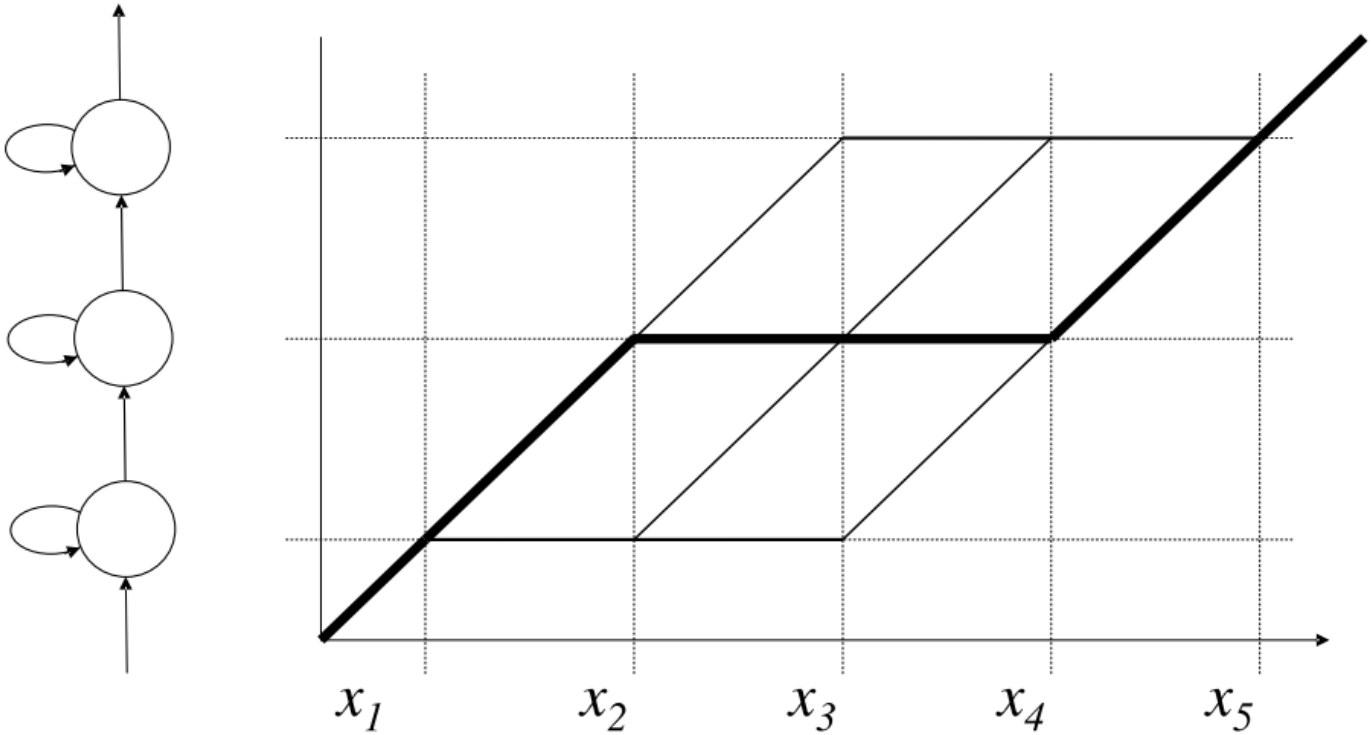


Diagramma a griglia (trellis diagram) per l'algoritmo di Viterbi.

COMUNICAZIONE VERBALE

Riconoscimento del parlato – Approccio basato su HMMs

Quindi a partire dai punti finali associati al secondo istante di tempo, il percorso può essere esteso in modo da calcolare la transizione dallo stato 1 (nello stato 1 o nello stato 2) o la transizione dallo stato 2 (nello stato 2 o nello stato 3). Quindi, per il terzo istante, il modello può trovarsi nello stato 1, nello stato 2 o nello stato 3.

A tutte queste opzioni può essere associata una probabilità, moltiplicando le probabilità ottenute per il secondo istante di tempo per la probabilità di transizione e per la

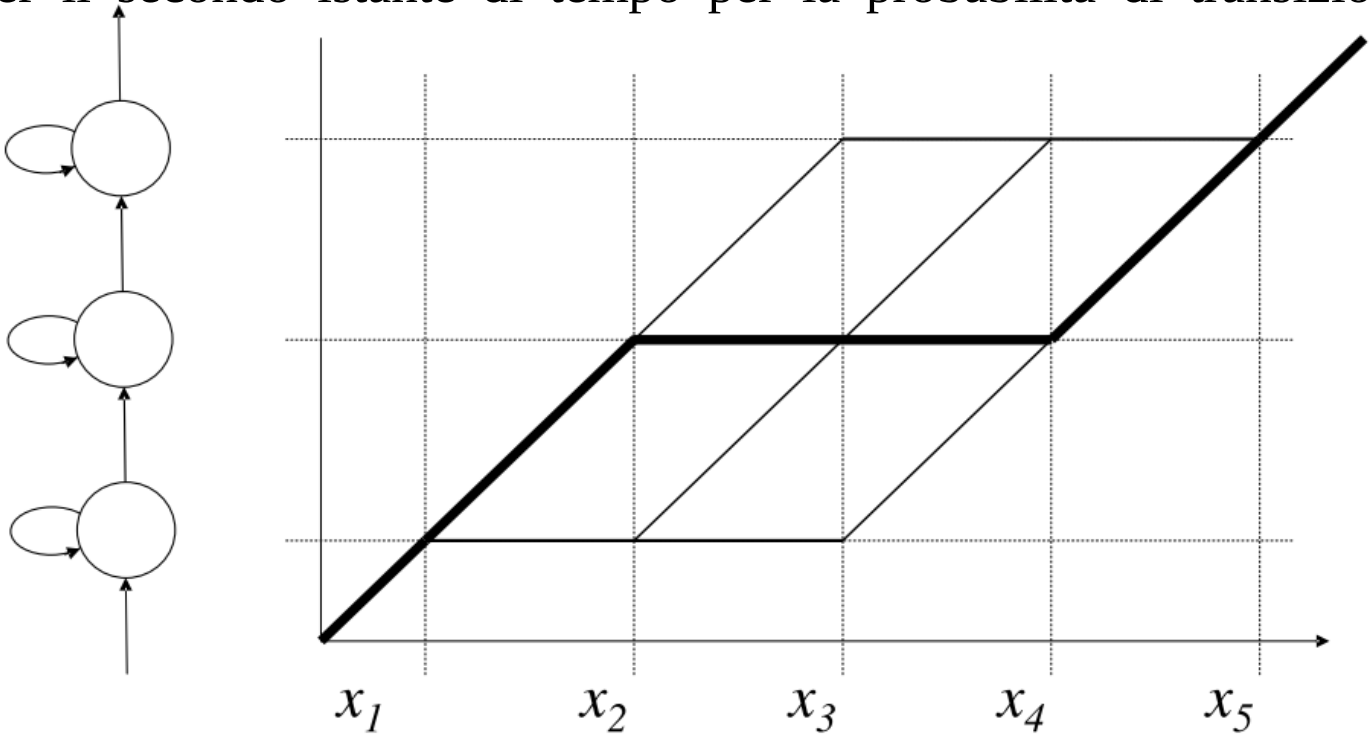


Diagramma a griglia (trellis diagram) per l'algoritmo di Viterbi.

probabilità di emissione associata al terzo *feature vector*.

Nella figura si può notare come ogni possibile sequenza di stati può essere rappresentata in una griglia (denominata trellis), la quale mostra tutti i possibili percorsi dallo stato 1 allo stato finale 3, tenendo conto delle osservazioni di cinque *feature vectors* differenti.

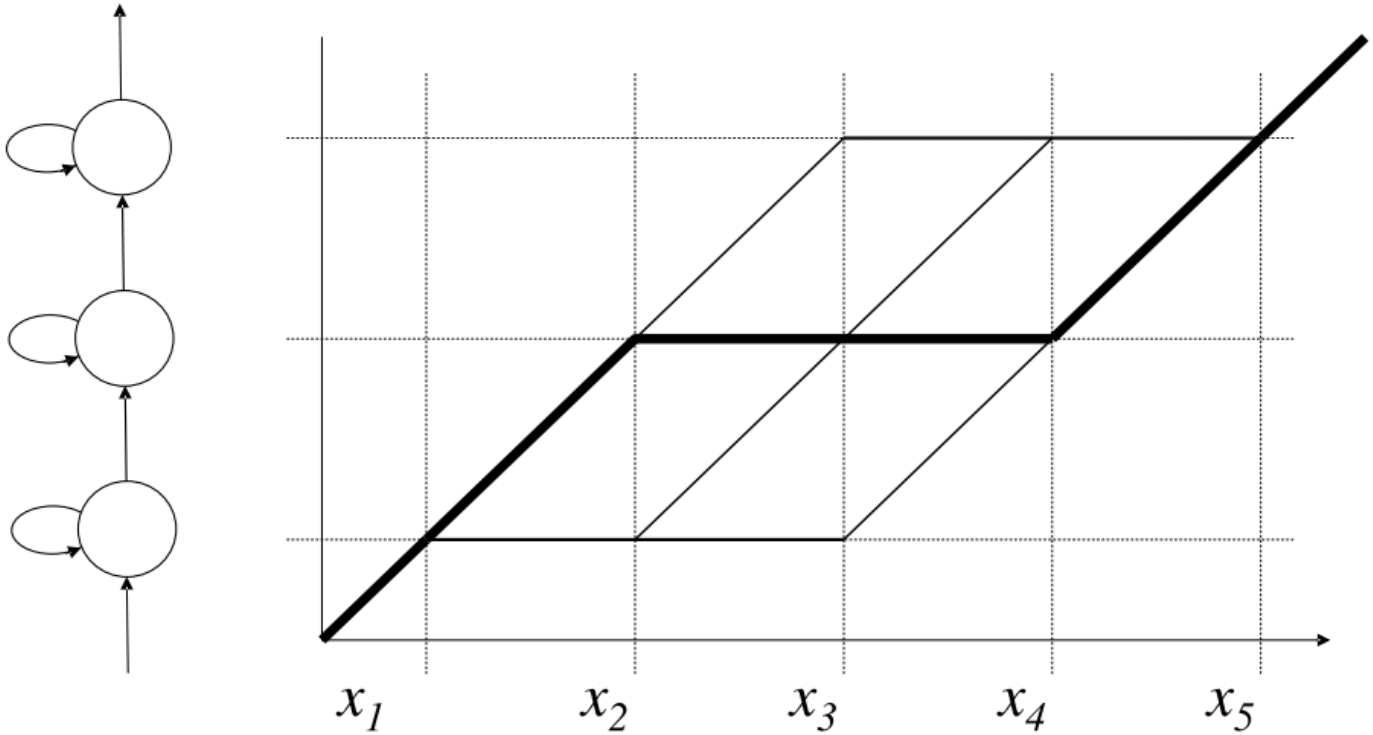


Diagramma a griglia (trellis diagram) per l'algoritmo di Viterbi.

COMUNICAZIONE VERBALE

Riconoscimento del parlato – Approccio basato su HMMs

Nella figura, la linea in grassetto mostra un possibile percorso nella griglia; per ogni percorso può essere calcolata una probabilità associata al fatto che il percorso considerato sia stato eseguito (utilizzando la procedura descritta, la quale indica di eseguire la moltiplicazione tra le probabilità associate ai singoli *feature vectors*).

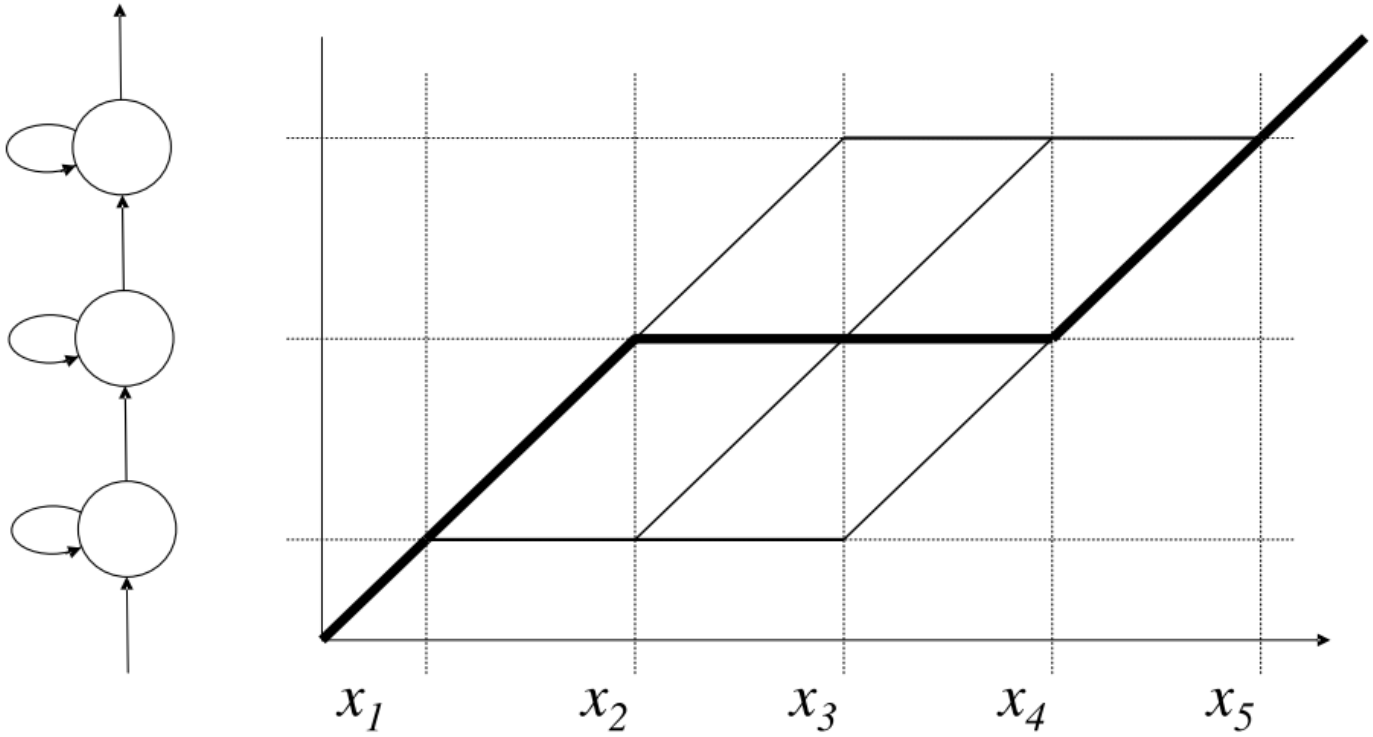


Diagramma a griglia (trellis diagram) per l'algoritmo di Viterbi.

COMUNICAZIONE VERBALE

Riconoscimento del parlato – Approccio basato su HMMs

Il percorso ottimo può essere calcolato utilizzando il principio della programmazione dinamica (appartenente alla teoria dell'ottimizzazione). La procedura appena descritta riassume i passi fondamentali dell'algoritmo di Viterbi. I principali risultati di tale algoritmo sono la sequenza ottima di stati e la probabilità associata a tale sequenza. Tale probabilità è la probabilità che il *HMM* dato abbia prodotto la sequenza di *feature vectors X*.

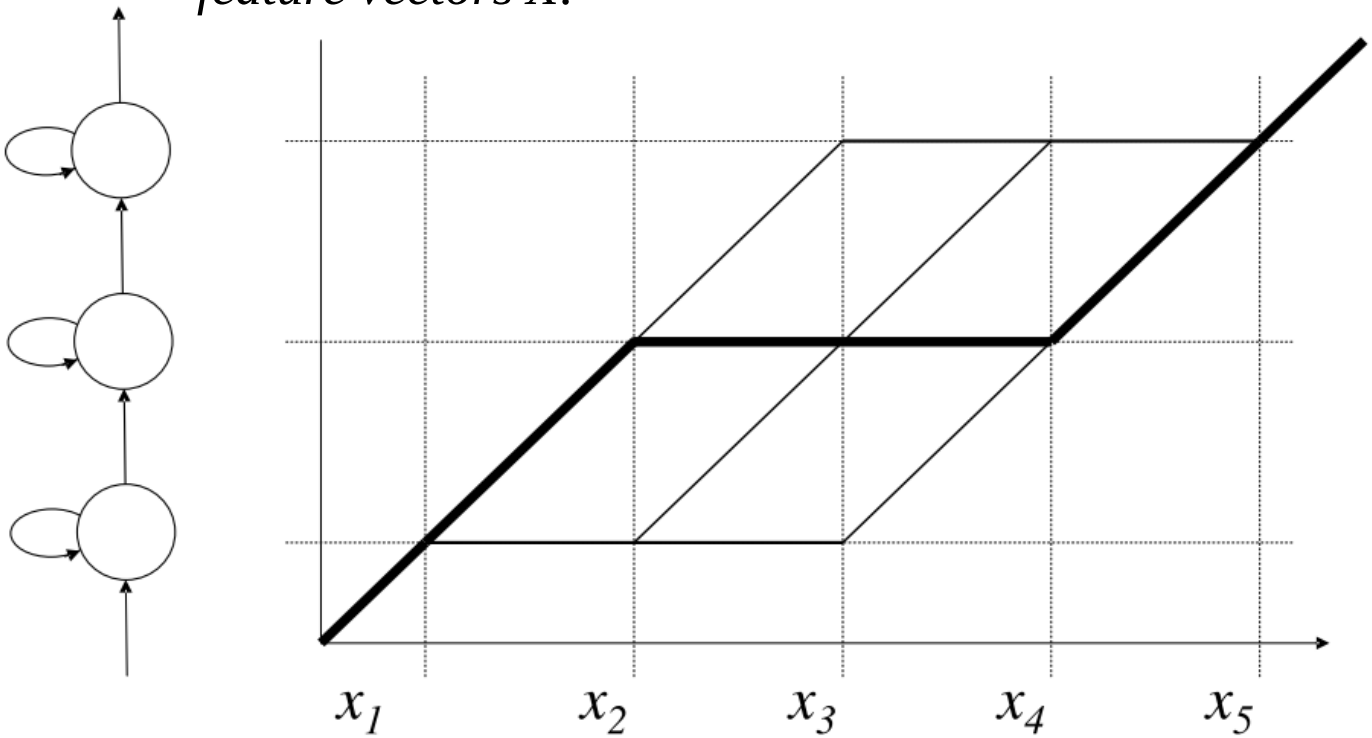


Diagramma a griglia (trellis diagram) per l'algoritmo di Viterbi.

COMUNICAZIONE VERBALE

Riconoscimento del parlato – Approccio basato su HMMs

Si ipotizzi che il *HMM* (caratterizzato da tre stati) mostrato nella figura della slide precedente rappresenti un fonema. Tipicamente, un segnale vocale sconosciuto rappresenta una parola o una frase del parlato. Come può una frase essere riconosciuta dall'algoritmo di Viterbi mediante la procedura di state-alignment di un *HMM* che rappresenta un singolo fonema?

Tale obiettivo può essere raggiunto estendendo l'algoritmo in modo da calcolare la sequenza di *HMMs* (ognuno dei quali è associato ad un singolo fonema) che massimizza la probabilità di emissione associata alla sequenza osservata di *feature vectors*. Ciò significa che, dopo che è stato raggiunto lo stato finale del *HMM* riportato nella figura precedente (ipotizzando che la sequenza di *feature vectors* non sia ancora terminata), sarà aggiunto un altro *HMM* (ed eventualmente altri). L'algoritmo continua elaborando tutti i *feature vectors*, ottenendo così uno stato finale (che rappresenta la fine di una parola).

Poiché non è noto quali *HMMs* devono essere aggiunti e non è noto quale sarà la sequenza ottima di *HMMs*, tale procedura rappresenta un *search problem*. Per tale motivo, tale processo viene chiamato anche *decoding*.

COMUNICAZIONE VERBALE

Riconoscimento del parlato – Approccio basato su HMMs

Tale procedura di *decoding* può essere supportata in differenti modi. Ad esempio si può considerare il fatto che l'ordine dei fonemi all'interno delle parole è piuttosto fisso e che la variazione può avvenire principalmente solo tra le parti iniziali e finali di una parola. Quindi la procedura di *search* sui fonemi rappresenta una procedura di *search* sulle parole; essa può essere supportata ulteriormente dal modello del linguaggio (tale concetto verrà spiegato più dettagliatamente nelle slide successive). Il modello del linguaggio, utilizzando la probabilità, permette di estendere il percorso di *search* con un modello di una nuova parola se la procedura di *search* ha raggiunto lo stato finale del modello della parola precedente. Quindi, l'algoritmo può calcolare la sequenza di parole più probabile e il risultato finale della procedura di riconoscimento è la frase riconosciuta.

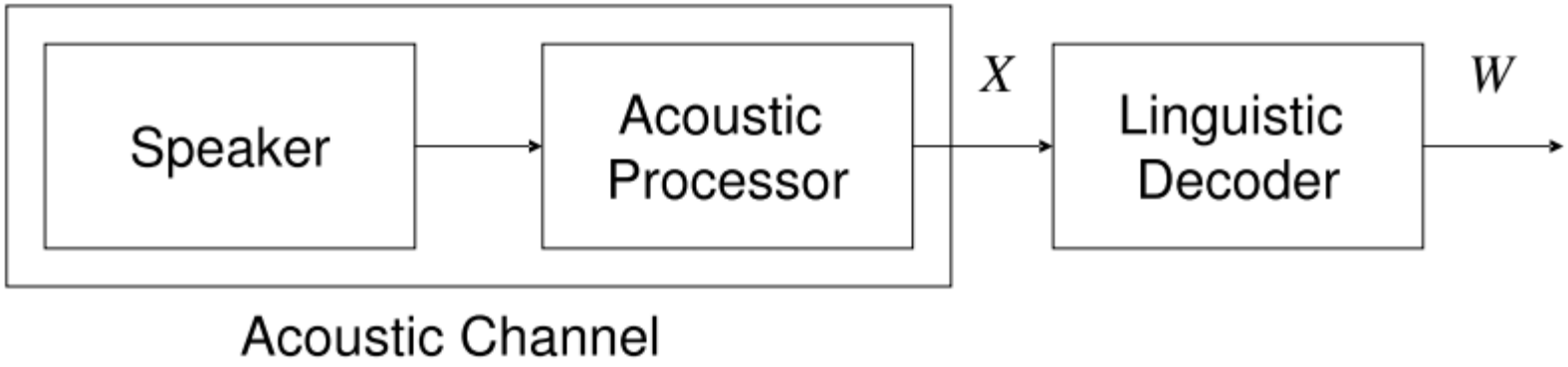
COMUNICAZIONE VERBALE

Riconoscimento del parlato – Approccio basato su HMMs

L'ASR con approccio basato su *HMMs* può essere interpretato dal punto di vista della teoria dell'informazione, come mostrato nella figura. Uno speaker formula una frase come una sequenza di parole indicata con

$$W = [w(1), w(2), ..., w(N)]$$

Lo speaker utilizza un microfono che cattura il segnale vocale. Il segnale vocale costituisce l'input del sistema ASR. Il sistema non vede la sequenza originale pronunciata dallo speaker; il sistema vede la versione codificata («encoded») che consiste in una sequenza di *feature vectors X*. Tale sequenza viene ricavata dalla forma d'onda sonora come output del canale acustico.



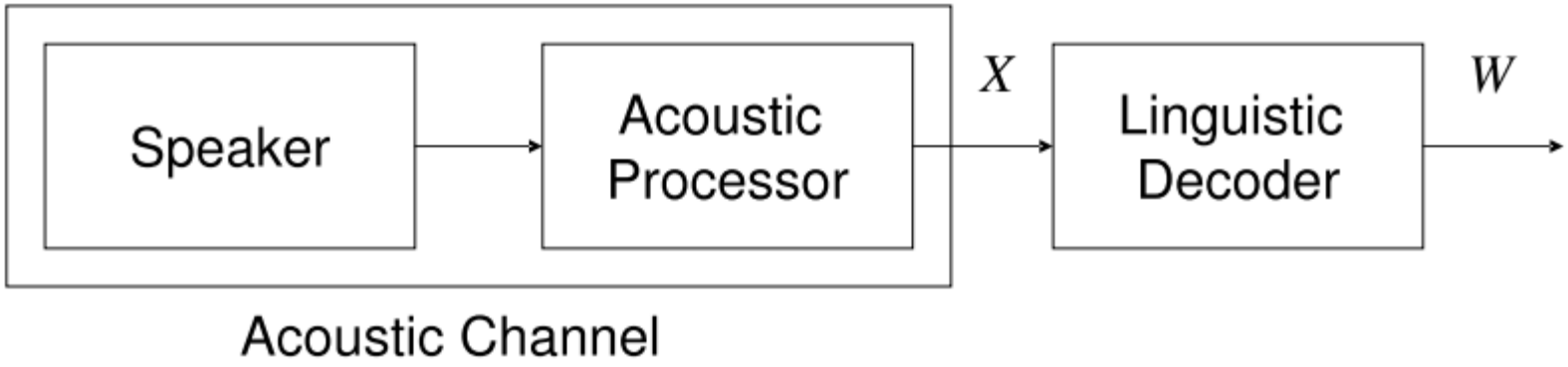
Interpretazione dell'ASR dal punto di vista della teoria dell'informazione.

COMUNICAZIONE VERBALE

Riconoscimento del parlato – Approccio basato su HMMs

I *HMMs* che rappresentano la sequenza di fonemi della sequenza di parole *W* permettono di modellizzare una relazione tra *W* e *X* basata sulla probabilità. Tale relazione restituisce la probabilità che la sequenza di parole *W*, rappresentata dall'opportuna sequenza di fonemi modellizzati tramite *HMMs*, abbia generato la sequenza osservata di *feature vectors*. Tale probabilità può essere indicata con $p(X|W)$.

La seconda parte della figura mostra il «Linguistic Decoder», cioè il modulo responsabile di decodificare l'informazione originale *W* elaborando la sequenza *X*. Per eseguire tale elaborazione, viene utilizzata la conoscenza relativa al modello (*HMMs*) espressa in $p(X|W)$.



Interpretazione dell'ASR dal punto di vista della teoria dell'informazione.

COMUNICAZIONE VERBALE

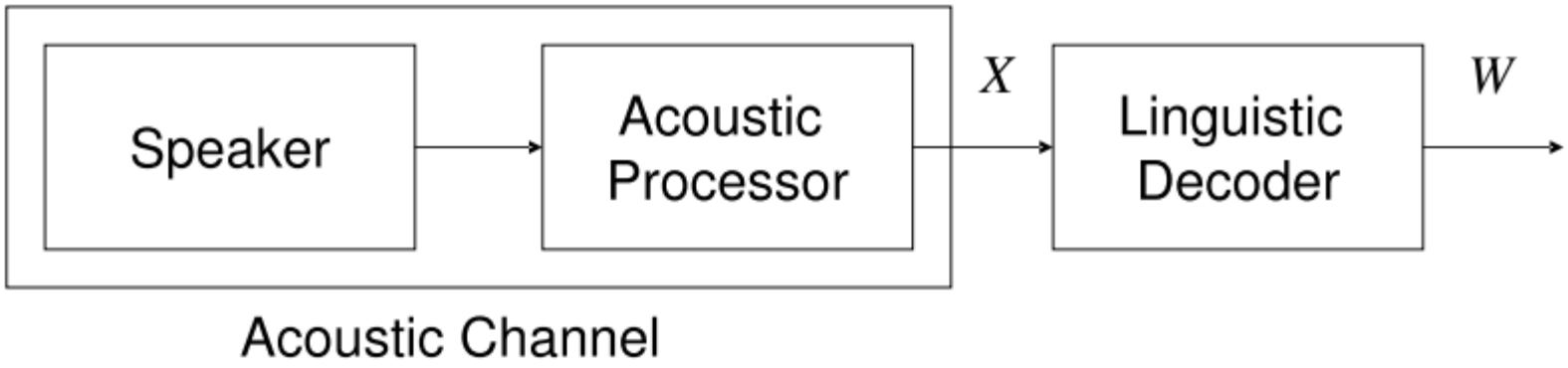
Riconoscimento del parlato – Approccio basato su HMMs

La strategia di *decoding* del modulo «Linguistic Decoder» si può formulare come

$$\max_W p(W|X) = \max[p(X|W) \cdot \frac{p(W)}{p(X)}]$$

Formula di Bayes

$$P(A_i \mid B) = \frac{P(A_i)\, P(B \mid A_i)}{P(B)}$$



Interpretazione dell'ASR dal punto di vista della teoria dell'informazione.

COMUNICAZIONE VERBALE

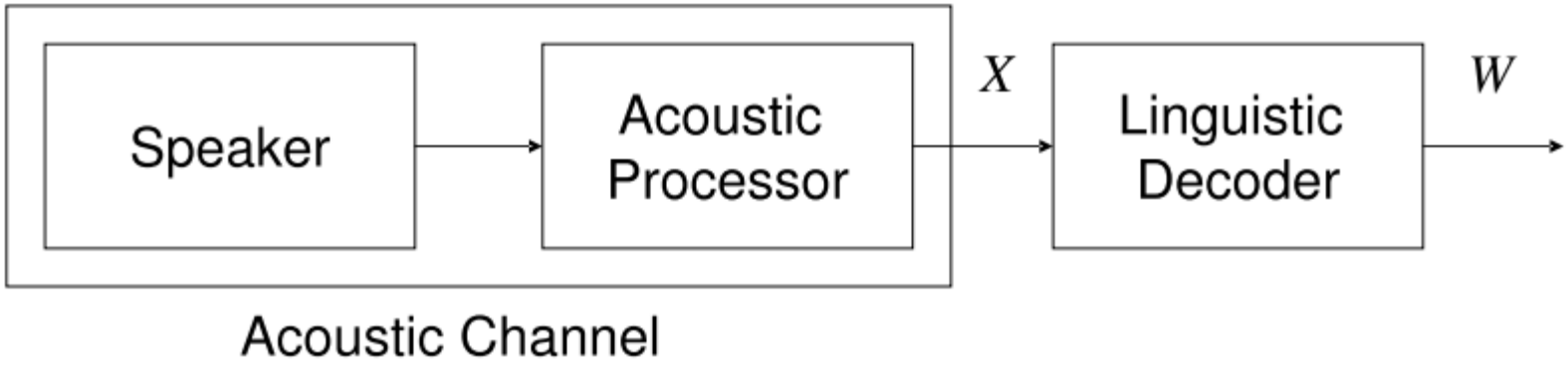
Riconoscimento del parlato – Approccio basato su HMMs

La strategia di *decoding* del modulo «Linguistic Decoder» si può formulare come

$$\max_W p(W|X) = \max[p(X|W) \cdot \frac{p(W)}{p(X)}]$$

Poiché la ricerca della sequenza di parole ottima *W* non dipende dalla probabilità $p(X)$, la formula precedente diventa

$$\max_W [p(X|W) \cdot p(W)]$$



Interpretazione dell'ASR dal punto di vista della teoria dell'informazione.

COMUNICAZIONE VERBALE

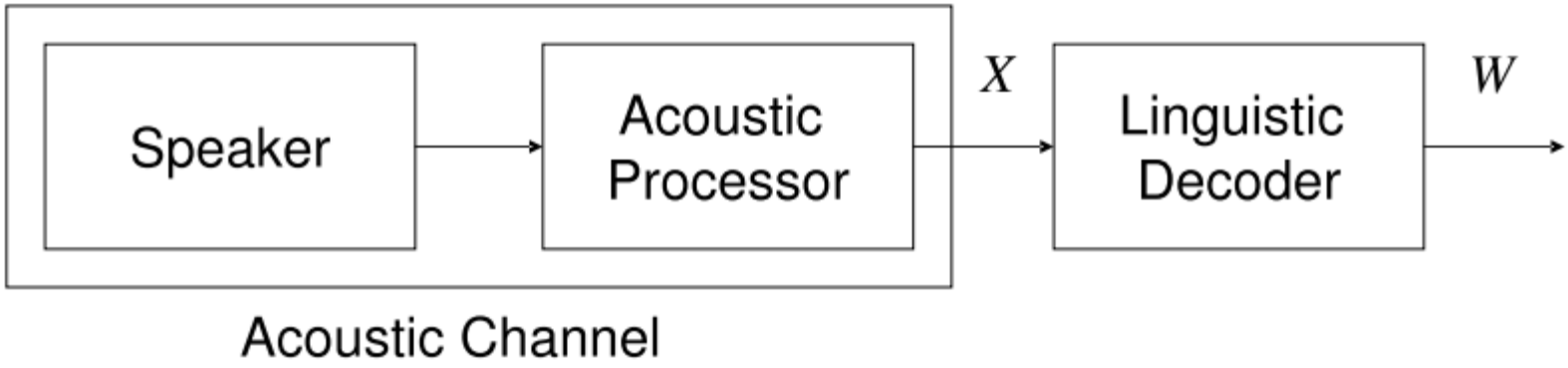
Riconoscimento del parlato – Approccio basato su HMMs

$$\max_{W}[p(X|W) \cdot p(W)]$$

Nella procedura di *search* dell'algoritmo di Viterbi deve essere quindi massimizzato il prodotto di probabilità riportato nella formula precedente.

Il termine $p(X|W)$ è la probabilità associata alla sequenza *X* nell'ipotesi che essa sia stata generata da *W*. Tale probabilità è data dalla concatenazione degli *HMMs* associati ai fonemi in un modello che rappresenta la stringa di parole risultante.

La formula precedente esprime la strategia di decodifica (*decoding*) menzionata precedentemente, cioè la ricerca della combinazione degli *HMMs* associati ai fonemi che massimizzano la probabilità di emissione.



Interpretazione dell'ASR dal punto di vista della teoria dell'informazione.

COMUNICAZIONE VERBALE

Riconoscimento del parlato – Approccio basato su HMMs

$$\max_W [p(X|W) \cdot p(W)]$$

Tuttavia, nella formula è presente il termine $p(W)$, il quale rappresenta la probabilità associata alla frase («sentence probability»), cioè la probabilità che la sequenza di parole *W* si presenti, indipendentemente dall'osservazione acustica *X*. Tale probabilità è descritta dal modello del linguaggio come la probabilità di una parola in una frase data la sequenza dei suoi predecessori:

$$p(w(n)|w(n-1), w(n-2), ..., w(n-m))$$

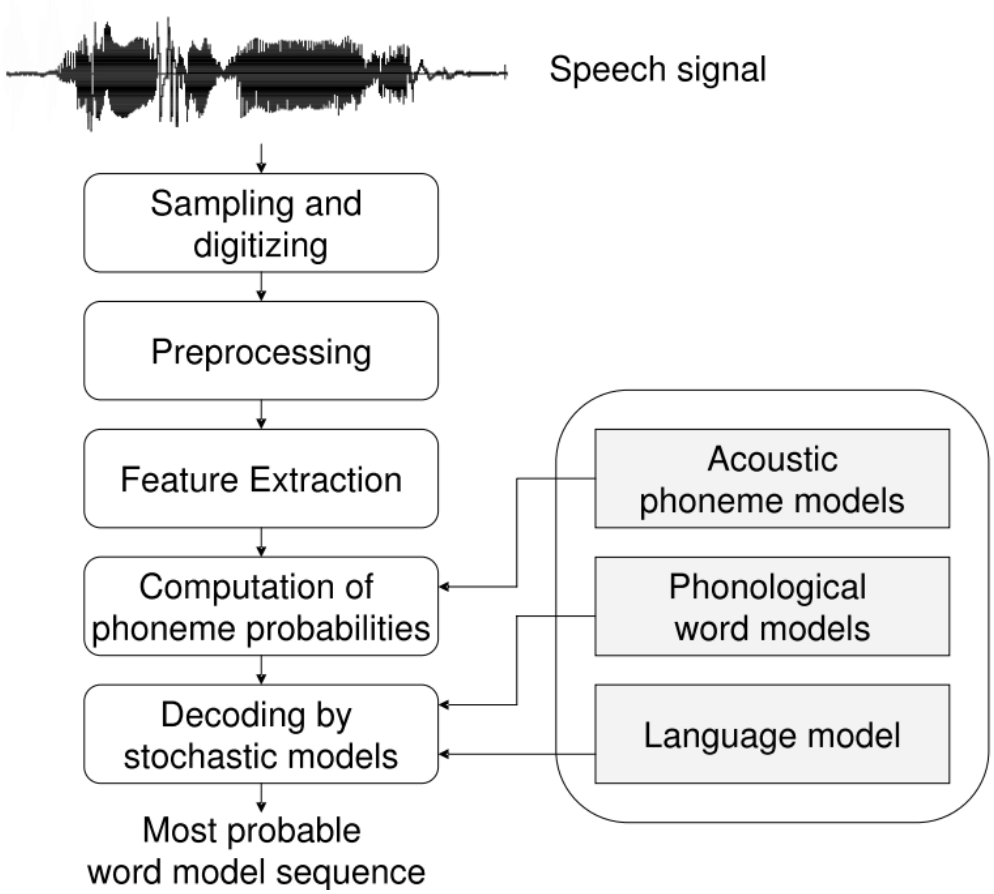
m indica la «word history», cioè il numero delle parole predecessore che sono rilevanti per il calcolo della probabilità di occorrenza della parola considerata. Quindi, la «sentence probability» può essere espressa come il prodotto delle probabilità associate alle singole parole:

$$p(W) = \prod_{n=1}^{N} p(w(n)|w(n-1), ..., w(n-m))$$

dove *N* è la lunghezza della frase e *m* è la «word history». Come menzionato, tali probabilità sono indipendenti da ogni osservazione acustica e possono essere calcolate ad esempio dall'analisi di testi opportuni.

COMUNICAZIONE VERBALE

Riconoscimento del parlato – Approccio basato su HMMs



Schema a blocchi per sistemi ASR basati su HMMs.

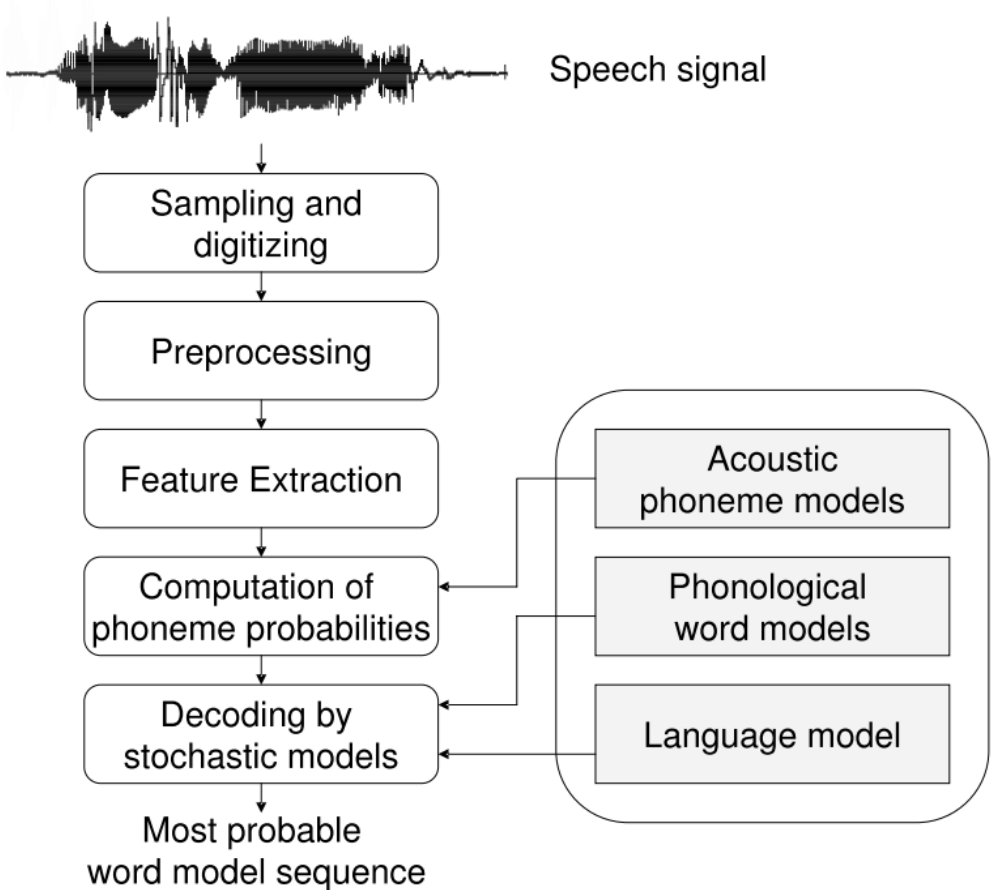
Il segnale vocale viene acquisito mediante un microfono, campionato e digitalizzato. La fase di preelaborazione include alcuni processi di filtraggio e di compensazione del rumore. Lo step successivo è l'estrazione delle features, nel quale il segnale viene diviso in finestre (ad esempio di 10 ms). Per ogni finestra viene calcolato un *feature vector* (tipicamente nel dominio della frequenza).

Successivamente, nella fase di riconoscimento, per ogni vettore della sequenza risultante di *feature vectors* possono essere calcolate le probabilità condizionate associate agli stati, inserendo, per ogni stato considerato nella procedura di *decoding*, il *feature vector x(k)* nella parte destra dell'equazione

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

COMUNICAZIONE VERBALE

Riconoscimento del parlato – Approccio basato su HMMs

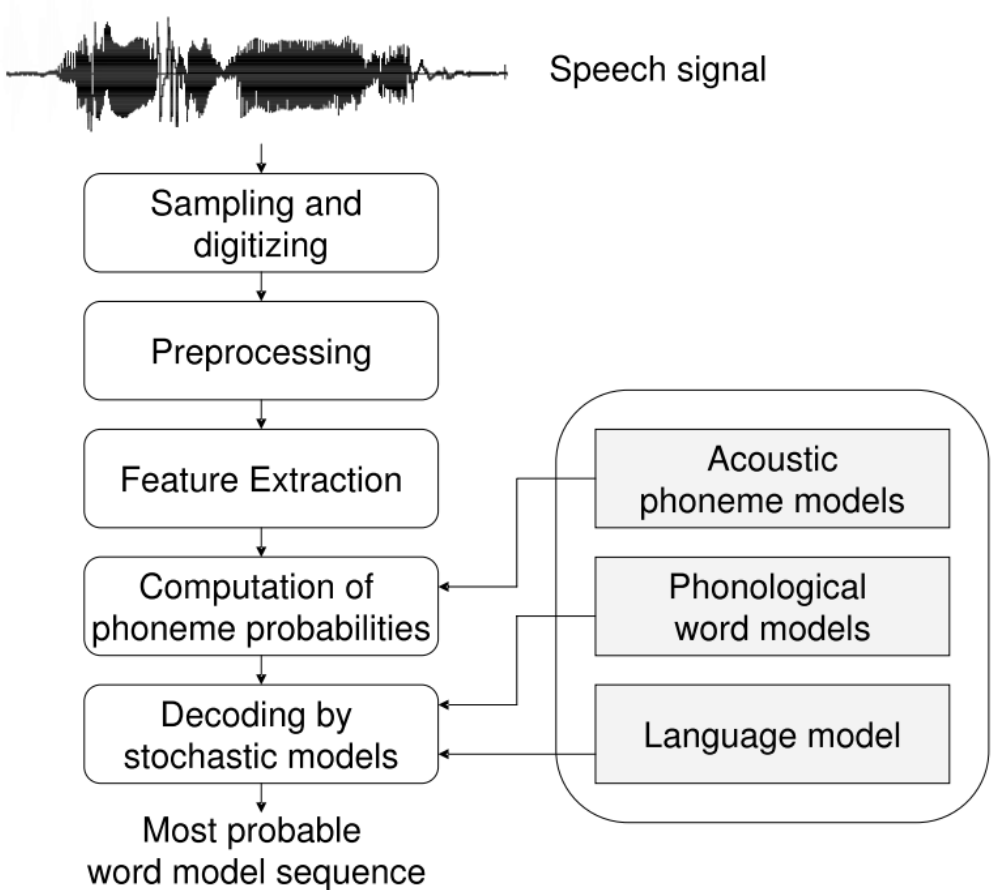


Schema a blocchi per sistemi ASR basati su HMMs.

Gli «acoustic phoneme models» contengono i parametri delle funzioni di distribuzione utilizzate per calcolare le probabilità condizionate associate agli stati. Tale calcolo viene integrato nella procedura di *search* precedentemente descritta, che cerca di trovare la sequenza di *HMMs* (dove ogni *HMM* è associato a un singolo fonema) che massimizza la probabilità di emissione della sequenza di *feature vectors*. Tale procedura di *search* viene controllata con i «phonological word models» (i quali descrivono come una parola possa essere suddivisa in fonemi) e con il modello del linguaggio. Il modello del linguaggio indica le probabilità per esaminare quale sarà la prossima parola della sequenza se la procedura di *search* ha raggiunto lo stato finale di un *HMM* associato al modello della parola precedente. Quindi il calcolo delle probabilità condizionate

COMUNICAZIONE VERBALE

Riconoscimento del parlato – Approccio basato su HMMs



Schema a blocchi per sistemi ASR basati su HMMs.

associate agli stati non deve essere eseguito per tutti i possibili stati, ma solo per gli stati che sono considerati probabili dalla procedura di *search*. Il risultato della procedura di *search* è la sequenza di modelli di parole più probabile. Tale sequenza viene mostrata all'utente del sistema ASR come una trascrizione che rappresenta la frase riconosciuta.

COMUNICAZIONE VERBALE

Riconoscimento del parlato – Word Error Rate

La metrica standard per la valutazione delle prestazioni di un sistema ASR è il *word error rate* (*WER*). Essa calcola quanto la stringa di parole restituita dal sistema ASR (cioè la stringa di parole stimata) differisce dalla trascrizione di riferimento.

Il calcolo del *WER* consiste nel ricavare la *minimum edit distance* in parole tra la stringa stimata e la stringa reale. Viene quindi calcolato il numero minimo di *word substitutions*, di *word insertions* e di *word deletions* necessarie per far corrispondere la stringa stimata alla stringa reale. Il *WER* viene quindi definito dalla formula seguente (si noti che siccome l'equazione include le *insertions*, il *WER* può essere maggiore del 100%):

$$\text{Word Error Rate} = 100 \times \frac{\text{Insertions} + \text{Substitutions} + \text{Deletions}}{\text{Total Words in Correct Transcript}}$$

COMUNICAZIONE VERBALE

Riconoscimento del parlato – Word Error Rate

REF:	i	***	**	UM	the	PHONE	IS		i	LEFT	THE	portable	****	PHONE	UPSTAIRS	last night
HYP:	i	GOT	IT	TO	the	*****	FULLEST		i	LOVE	TO	portable	FORM	OF	STORES	last night
Eval:		I	I	S		D	S			S	S		I	S	S	

Esempio di allineamento tra stringa reale e stringa stimata.

Nell'esempio riportato si notano 6 *substitutions*, 3 *insertions* e 1 *deletion*.

$$\text{Word Error Rate} = 100\frac{6+3+1}{13} = 76.9\%$$

Il metodo standard per calcolare il *WER* è un pacchetto chiamato ***sclite*** (National Institute of Standards and Technologies (NIST), 2005). Tale pacchetto, oltre all'esecuzione di allineamenti e al calcolo del *WER*, esegue altri task molto utili. Ad esempio, fornisce informazioni utili sull'analisi degli errori come le matrici di confusione (le quali mostrano quali parole vengono confuse con altre) e riassume alcune statistiche associate a parole che vengono spesso *inserted* o *deleted*. Inoltre, tale pacchetto calcola gli *error rates* per ogni speaker (se le frasi sono etichettate con un identificatore per ogni speaker) e alcune statistiche utili come ad esempio il *sentence error rate*, cioè la percentuale di frasi con almeno una parola errata.

COMUNICAZIONE VERBALE

Riconoscimento del parlato – Word Error Rate

English Tasks	**WER%**
LibriSpeech audiobooks 960hour clean	1.4
LibriSpeech audiobooks 960hour other	2.6
Switchboard telephone conversations between strangers	5.8
CALLHOME telephone conversations between family	11.0
Sociolinguistic interviews, CORAAL (AAL)	27.0
CHiMe5 dinner parties with body-worn microphones	47.9
CHiMe5 dinner parties with distant microphones	81.3
Chinese (Mandarin) Tasks	**CER%**
AISHELL-1 Mandarin read speech corpus	6.7
HKUST Mandarin Chinese telephone conversations	23.5

audiolibri

audiolibri

WER e CER (Character Error Rate) di alcuni sistemi ASR in compiti di riconoscimento del parlato (2020).

Si noti che il *WER* associato alla lettura di un discorso (*read speech*) nel caso di audiolibri è circa uguale al 2%. Tale compito di riconoscimento può considerarsi risolto (anche se tale risultato è associato a sistemi ASR che richiedono enormi risorse computazionali). Invece i risultati in caso di trascrizione di conversazioni telefoniche tra umani sono peggiori; ciò si verifica anche nel caso di interviste e nel caso di trascrizione di conversazioni svolte durante una festa.

COMUNICAZIONE VERBALE

Riconoscimento del parlato – Tecnologia dei sistemi ASR

Per quanto riguarda la tecnologia dei sistemi ASR basati su *HMMs*, si possono distinguere le seguenti categorie:

- Sistemi ASR *speaker-independent* con vocabolari di piccole dimensioni (esempio: 10-50 parole), utilizzati ad esempio per applicazioni di telefonia. In tali sistemi viene sfruttata la capacità degli *HMMs* di modellizzare lunghe frasi ottenute da molti speakers differenti.

- Sistemi ASR *speaker-independent* con vocabolari di media dimensione, utilizzati ad esempio nell'automotive e in ambienti multimediali. In tali ambiti vengono spesso combinati *HMMs* e tecniche di riduzione del rumore. L'efficienza dei *HMMs* nel *decoding* di intere sequenze di fonemi e parole viene sfruttata per il riconoscimento del parlato in ambienti sfavorevoli.

- Sistemi ASR *speaker-dependent* e/o *speaker-adaptive* con vocabolari molto ampi (esempio: 100000 parole), utilizzati per dictation systems.

COMUNICAZIONE VERBALE

Riconoscimento del parlato – Tecnologia dei sistemi ASR

I sistemi ASR possono diventare *sistemi embedded* tramite implementazione su smartphone e altri dispositivi elettronici. Ciò è possibile grazie all'esistenza di schede con capacità adeguate a contenere i parametri acustici e i modelli del linguaggio.

Al fine di mitigare i problemi di memoria, sono nati i sistemi Distributed Speech Recognition (DSR). Si parla di sistemi DSR quando solo l'estrazione delle features viene eseguita sul dispositivo considerato e le features vengono poi trasferite a un server più grande dove vengono calcolate le probabilità di emissione e dove viene effettuato il *decoding* nella sequenza di parole stimata. Può essere sfruttato un server arbitrariamente grande, con sufficiente potenza di calcolo e sufficiente memoria per modelli di linguaggio estesi e per una grande quantità di parametri (ad esempio Gaussiani) relativi alla modellizzazione acustica.

COMUNICAZIONE VERBALE

Riconoscimento del parlato – Alcune aree di applicazione dei sistemi ASR

Alcune aree di applicazione dei sistemi ASR sono:

- Settore delle telecomunicazioni: il riconoscimento del parlato rappresenta un ponte naturale tra il settore delle telecomunicazioni e l'Information Technology; esso fornisce un'interfaccia naturale la quale permette di inserire dati tramite un canale di comunicazione basato sul parlato.

- Office Automation: un esempio in tale ambito è il dettato utilizzato da un segretario per creare una lettera tramite input vocale.

- Settore sanitario: un esempio in tale settore è la creazione di un report medico (ad esempio relativo alla radiologia) mediante analisi di una radiografia tramite dettato acquisito da un microfono. Un altro esempio in tale ambito è rappresentato dall'utilizzo di sistemi ASR per comandare dispositivi e interfacce impiegati per compensare disabilità associate a malfunzionamenti delle mani o delle braccia.

COMUNICAZIONE VERBALE

Riconoscimento del parlato – Alcune aree di applicazione dei sistemi ASR

Alcune aree di applicazione dei sistemi ASR sono:

- Settore della produzione e settore manifatturiero: esempi in tali settori sono la programmazione di macchine mediante il parlato e il controllo di impianti tramite comandi vocali.

- Applicazioni multimediali: lo sviluppo della tecnologia multimediale ha portato ad un incremento della domanda per interfacce basate sul parlato, il quale rappresenta una delle principali modalità delle interfacce multimodali. Ad esempio, esistono applicazioni basate sull'utilizzo del parlato per gestire alcuni *smart environments* o per accedere a documenti web.

- Settore privato: settore automotive, settore dei dispositivi elettronici e settore dei *games*.

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.