AISV Scuola Estiva 2008 - *Archivi di Corpora Vocali*

Riconoscimento del parlato

Carlo Drioli
Dipartimento di Informatica dell'Università di Verona

Outline

- Riconoscimento del parlato: problematiche
- Il front-end acustico
- I diversi approcci al riconoscimento di sequenze
- Dynamic Time Warping
- Approccio statistico e Hidden Markov Models

Riconoscimento del parlato (ASR)

Problematiche del riconoscimento automatico del parlato

- Dal segnale verbale è possibile estrarre informazioni di natura molto diversa:
  - Informazioni linguistiche (fonemi, parole, frasi, prosodia)
  - Caratteristiche paralinguistiche (emozioni, stato d'animo)
  - Informazioni dialogiche (gesti di avvicendamento, natura del dialogo)
  - Caratteristiche del parlatore (genere, identità, stato di salute)

- Ai fini del riconoscimento, il segnale verbale va opportunamente elaborato per:
  - Eliminare componenti indesiderate o di disturbo (ad es. rumore)
  - Enfatizzare le componenti utili al riconoscimento
  - Trasformare il dominio di analisi per rendere l'individuazione di pattern più semplice ed efficace (estrazione di cues acustiche)

Riconoscimento del parlato

Variabilità intraparlatore

- Variabilità temporale: uguali fonemi, sillabe o parole possono essere pronunciati con modalità e velocità diverse.

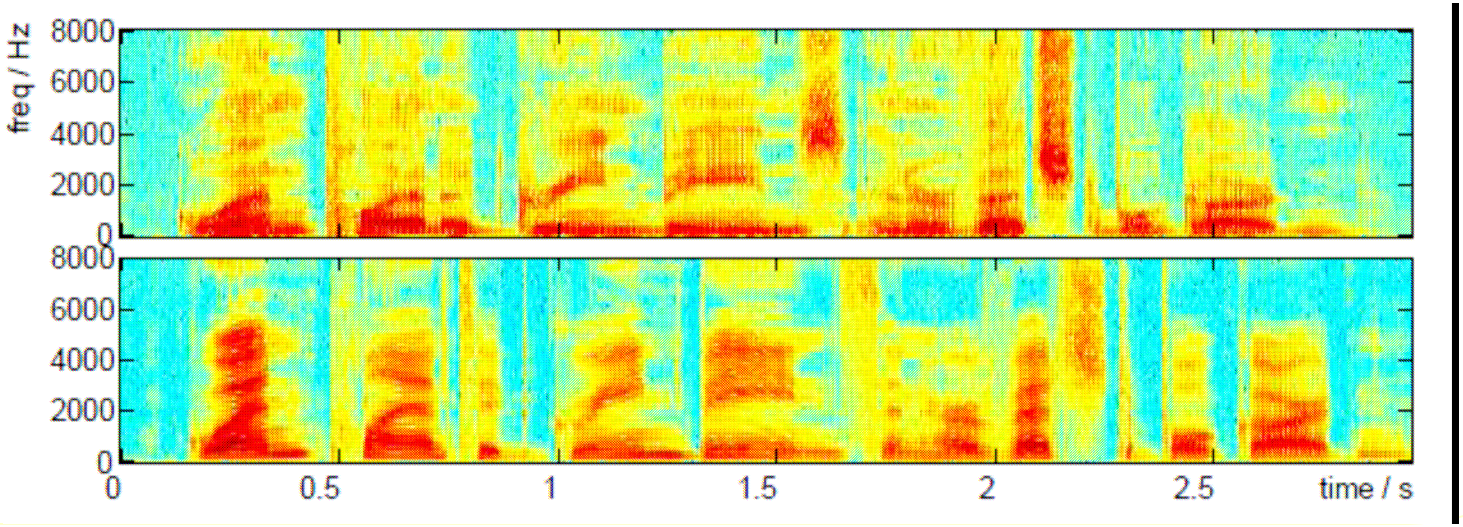
- Stile del parlato: può variare molto, da formale a informale, a seconda dell'intensità usata per la fonazione, o a seconda dello stato emotivo.

- Condizioni psico-fisiche del parlatore: uno stato di salute alterato può influire notevolmente sulle caratteristiche prosodiche o timbriche del parlato

Riconoscimento del parlato

Variabilità interparlatore

- Variabilità timbrica, a seconda del genere, dell'età, dello stato emotivo
- Caratteristiche individuali
- Caratteristiche dialettali



Riconoscimento del parlato

Variabilità dovuta alle condizioni ambientali

- Rumore di sottofondo (traffico, ventole, brusio)
- Riverbero ambientale (in ambienti chiusi)
- Tipo di microfono e sistema di registrazione usati

ASR: cues acustiche

Funzione e Scelta delle cues acustiche

- Il front-end acustico di un riconoscitore è preposto a fornire una rappresentazione del segnale vocale più efficace della forma d'onda ai fini di analisi dei pattern

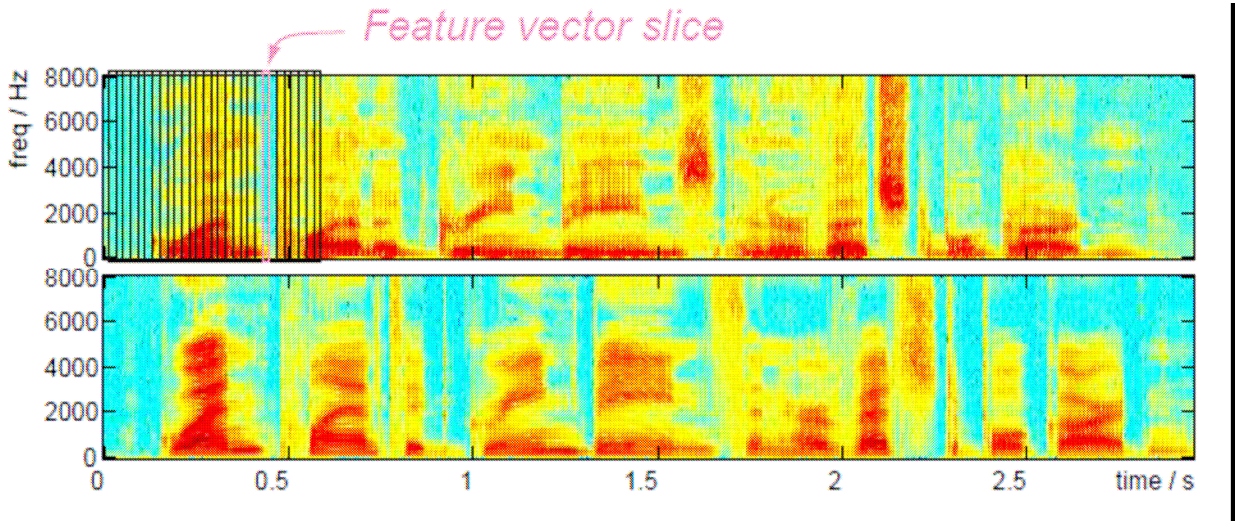
- La scelta delle caratteristiche acustiche ricade molto spesso su pochi parametri dalla robustezza consolidata (ad esempio, mfcc).

- Spesso le cues vengono selezionate anche in funzione del classificatori.

ASR: cues acustiche

Lo spettrogramma come parametro acustico del front-end

- E' possibile usare l'intero vettore che rappresenta l'analisi STFT di un frame, ad esempio:



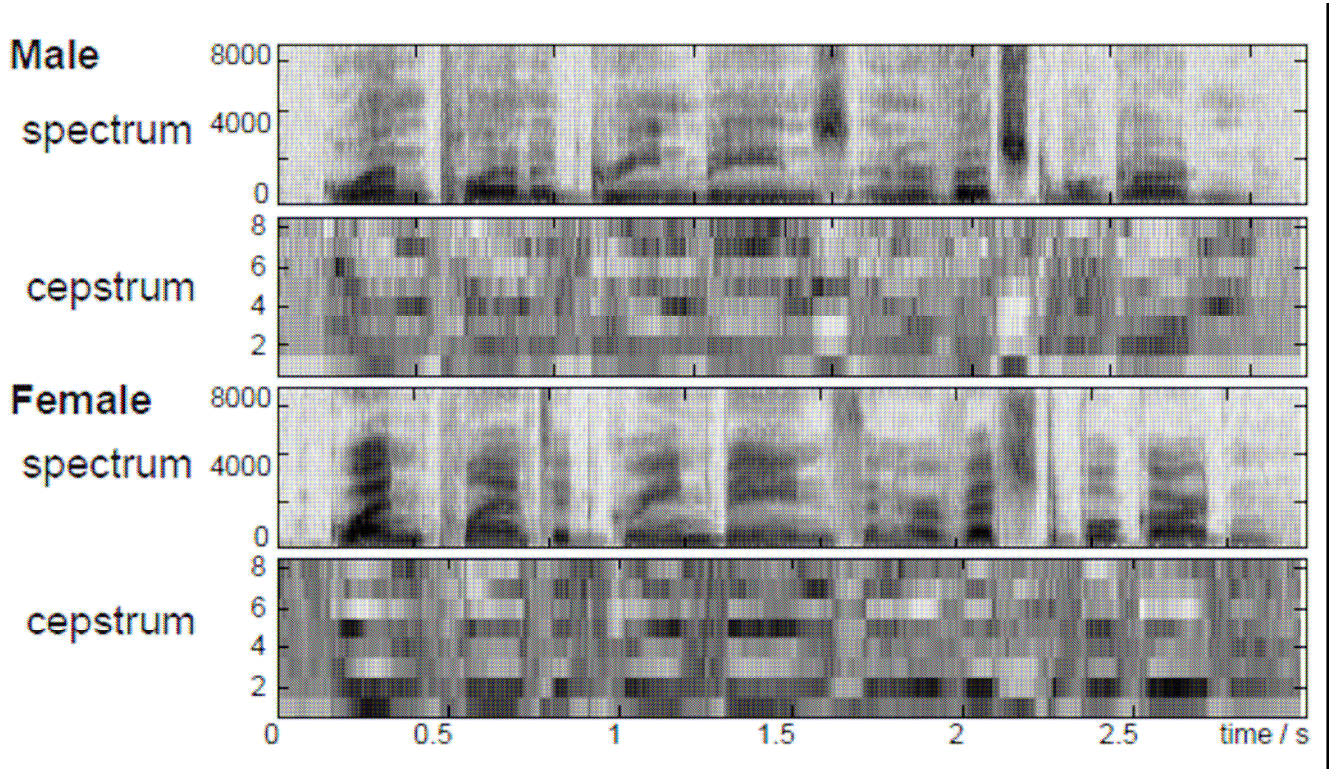
- Il confronto fra i due spettrogrammi evidenzia la necessità di ulteriori elborazioni per allineare e confrontare gli spettrogrammi efficacemente

ASR: cues acustiche

Il cepstrum come parametro acustico

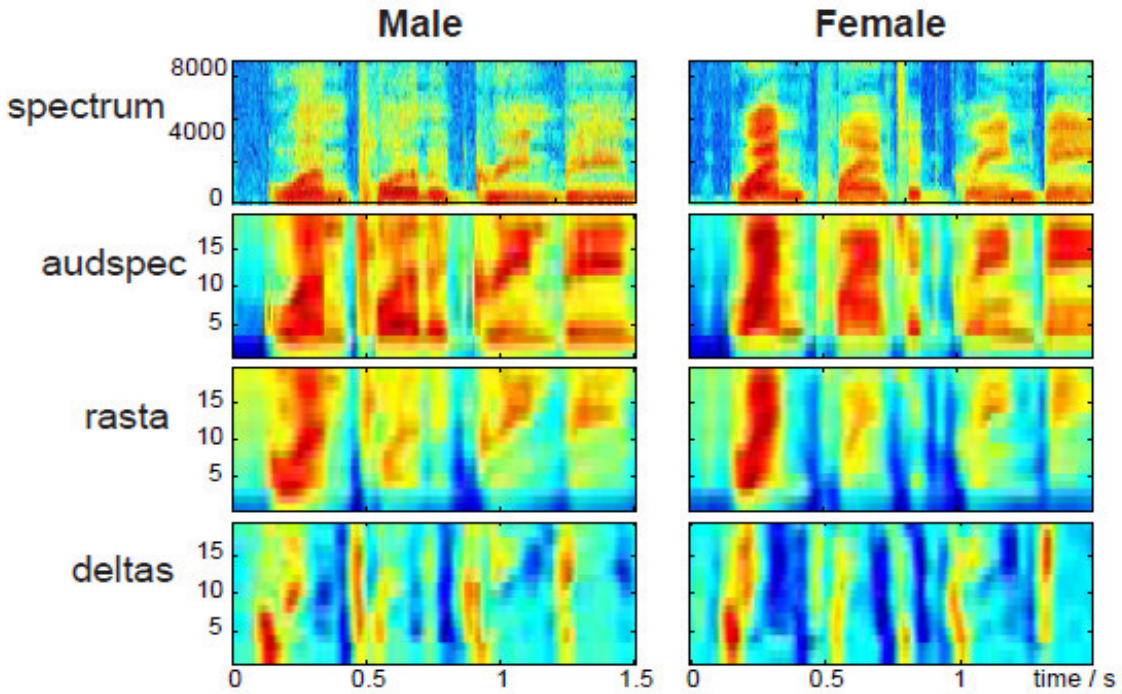
- Fornisce una rappresentazione compatta delle informazioni spettrali (dimensione ridotta dei vettori)

- Si presta ad essere modellato con misture gaussiane



ASR: cues acustiche

Confronto fra diverse caratteristiche



- Effetti positivi: fonemi simili sono normalizzati mentre il contrasto fra fonemi diversi risulta accentuato

Riconoscimento del parlato

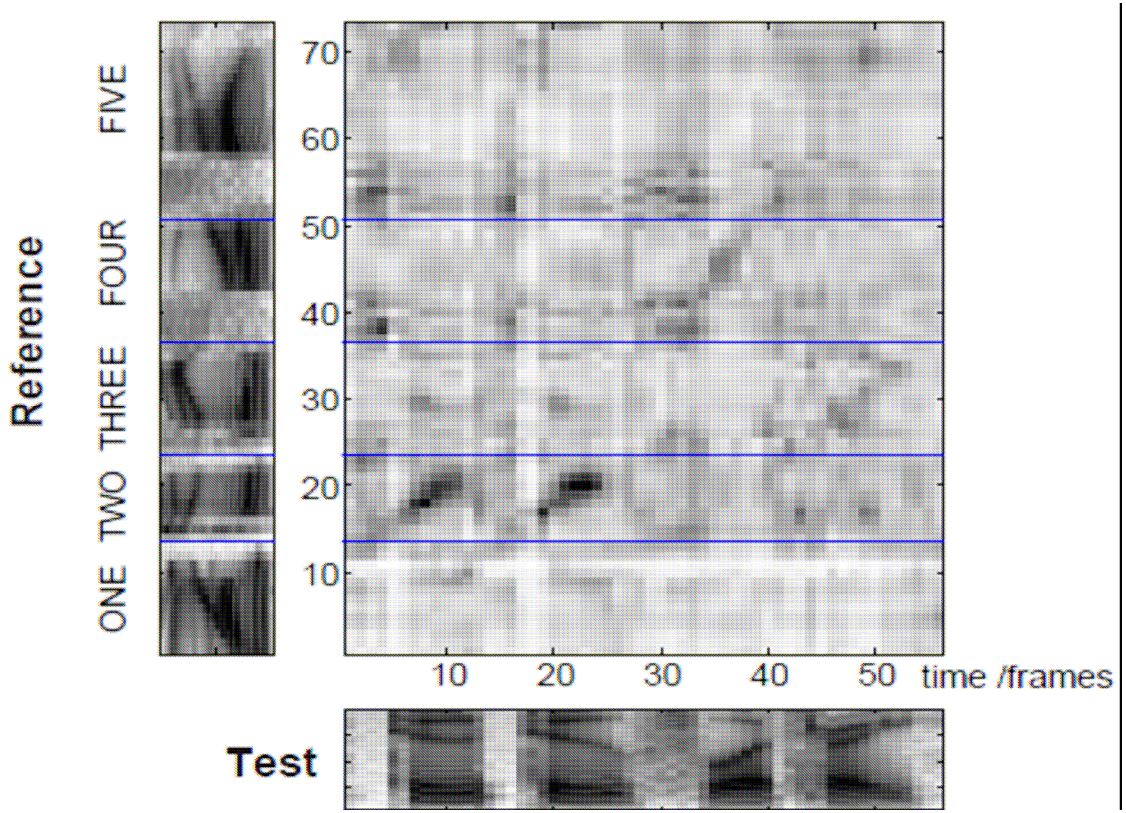
Confronto fra pattern: I diversi approcci possibili

- Confronto con pattern di riferimento.
- Approcci statistici: modelli acustici del parlatore e modelli di linguaggio
- Catene di Markov per modellare la tempo-varianza e gli aspetti di contesto

ASR: riconoscimento di pattern

Dynamic Time Warping (DTW) per il riconoscimento di sequenze

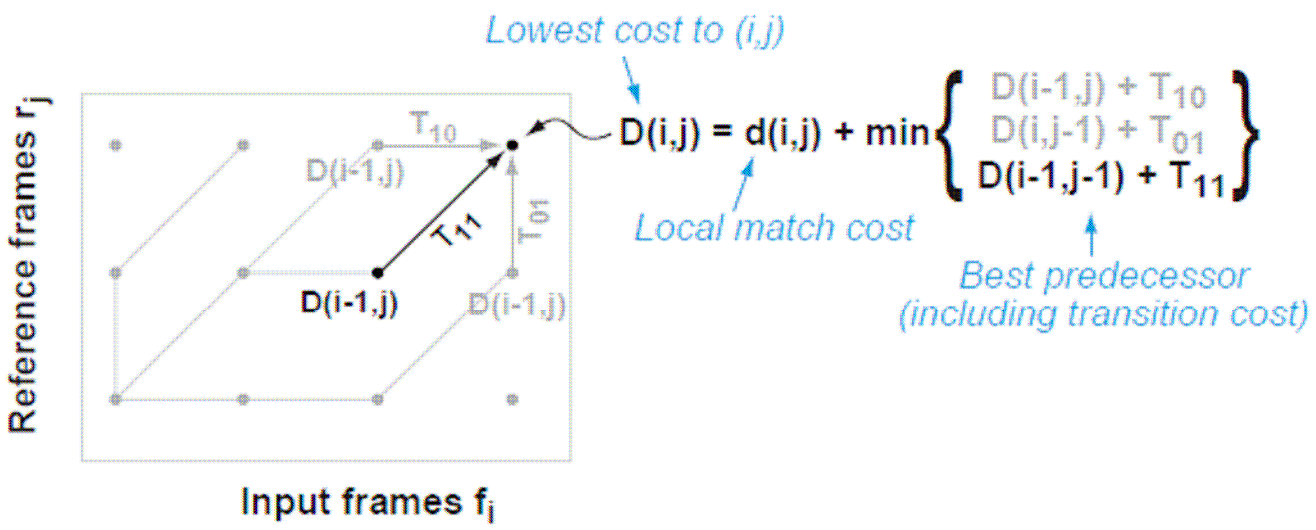
- Il confronto avviene frame per frame con dei pattern di riferimento preregistrati.



ASR: riconoscimento di pattern

Dynamic Time Warping (DTW)

- Viene cercato il percorso migliore sulla base di una funzione di costo
- Il costo è valutato sulla base della matrice delle distanze spettrali e dei costi di transizione
- Il percorso migliore è determinato alla fine sulla base del percorso migliore fra quelli tenuti in memoria



ASR: riconoscimento di pattern

Riconoscimento di sequenze: approccio statistico

- DTW ha dei limiti nella difficoltà di training e nella natura deterministica del procedimento
- L'approccio statistico si è rivelato ben presto una scelta più adeguata dal punto di vista teorico e pratico
- Il problema del riconoscimento viene riformulato secondo l'approccio Bayesiano:

$$M^* = \text{argmax} P(M|X)$$

  in cui M sono i possibili modelli rappresentanti parole o sequenze di parole e X sono le osservazioni

- $P(M|X)$ rappresenta la probabilità che la parola pronunciata sia M, data l'osservazione X

Riconoscimento statistico di sequenze

Riconoscimento statistico di sequenze:

- La difficoltà di calcolare $P(M|X)$ si può aggirare ricorrendo alla regola di Bayes, secondo cui:

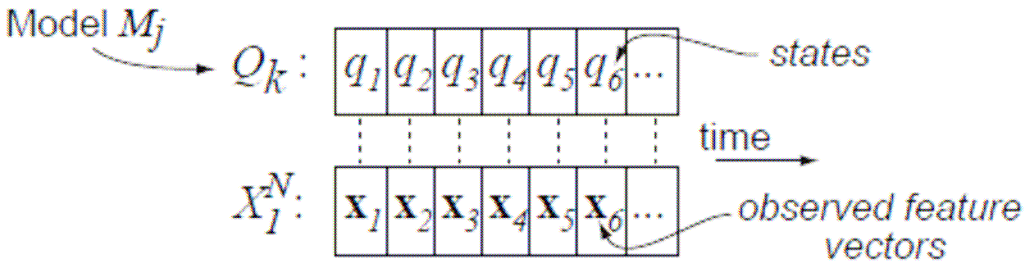
$$M^* = \text{argmax} P(M|X) = \text{argmax} P(X|M)P(M)$$

  in cui:
  - X sono le osservazioni acustiche
  - $P(X|M)$ è chiamato modello acustico
  - $P(M)$ è il modello del linguaggio
  - argmax è la ricerca attraverso tutte le sequenze

ASR: riconoscimento statistico di sequenze

Modelli a stati

- Si fa l'ipotesi che il segnale verbale evolva secondo un modello a stati
- Ogni modello evolve nello spazio degli stati, producendo vettori di caratteristiche acustiche osservabili



- La probabilità delle osservazioni è:

$$P(X|M) = \sum_Q P(X, Q|M) = \sum_Q P(X_1 \dots X_N|Q, M)P(Q|M)$$

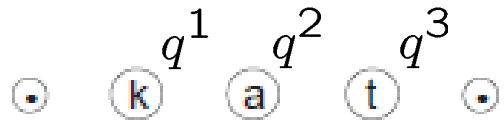
dove la somma è efettuata su tutte le sequenze possibili.

ASR: riconoscimento di pattern

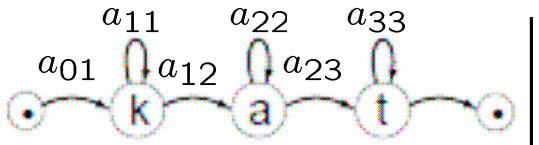
Hidden Markov Models (HMM)

I parametri che descrivono un modello HMM M_k sono:

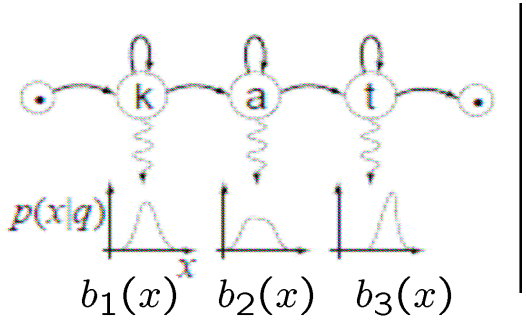
- Stati q^i



- Probabilità di transizione $a_{ij} = p(q_n^j | q_{n-1}^i)$



- Distribuzioni di emissione $b_i(x) = p(x|q^i)$



ASR: riconoscimento di pattern

Hidden Markov Models (HMM)

- La probabilità congiunta di una sequenza di osservazioni $X = x_1 x_2 \ldots$ e di una sequenza di stati $Q = q_1 q_2 \ldots$ è data da

$$P(X, Q|M) = a_{q_0 q_1} b_{q_1}(x_1) a_{q_1 q_2} b_{q_2}(x_2) a_{q_2 q_3} \ldots = a_{q_0 q_1} \prod_{t=1}^{T} b_{q_t}(x_t) a_{q_t q_{t+1}} \tag{1}$$

- Log-likelyhood: $\log P(X, Q|M) = \sum_{t=0}^{T} \log a_{q_t q_{t+1}} + \sum_{t=1}^{T} \log b_{q_t}(x_t)$

- Modello delle distribuzioni di emissione: Gaussian Mixture Model (GMM)

$$b_j(x_t) = \sum_{k=1}^{K} c_{jk} N(x_t; \mu_{jk}, \Sigma_{jk}) \tag{2}$$

Una GMM rappresenta la distribuzione statistica dei vettori mfcc osservati per un dato stato stato

ASR: riconoscimento di pattern

Hidden Markov Models (HMM)

A partire da $P(X, Q|M)$ è possibile infine calcolare $P(X|M)$ in vari modi:

- Sommando $P(X, Q|M)$ rispetto a tutte le possibili sequenze di stati:

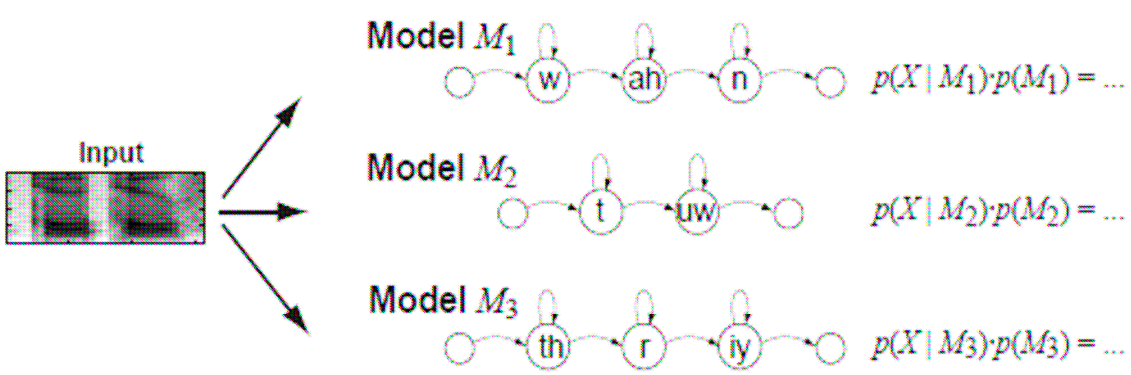
$$P(X|M) = \sum_{Q} P(X, Q|M)$$

- Con approssimazione di Viterbi, cercando la sequenza di stati che fornisce la massima probabilità: $P(X|M) = \max_Q P(X, Q|M)$

ASR: riconoscimento di pattern

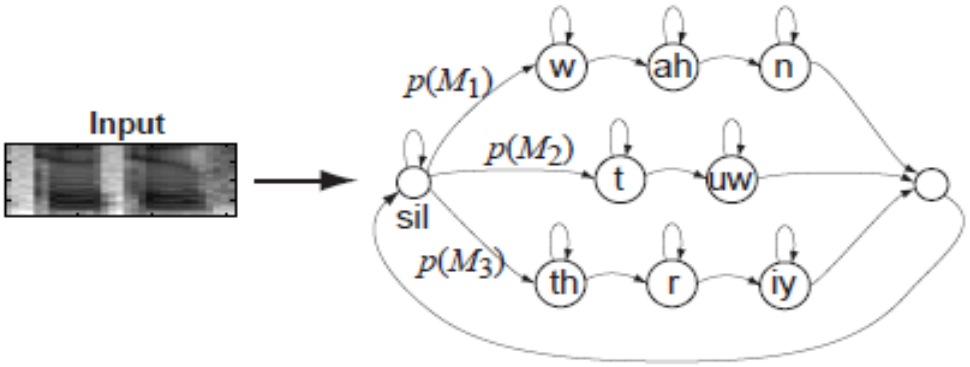
Riconoscimento con HMM

- Riconoscimento di parole isolate



Ricerca di
$\max\{P(M_i|X)\} =$
$\max\{P(X|M_i|)P(M_i)\}$

- Riconoscimento di parlato continuo

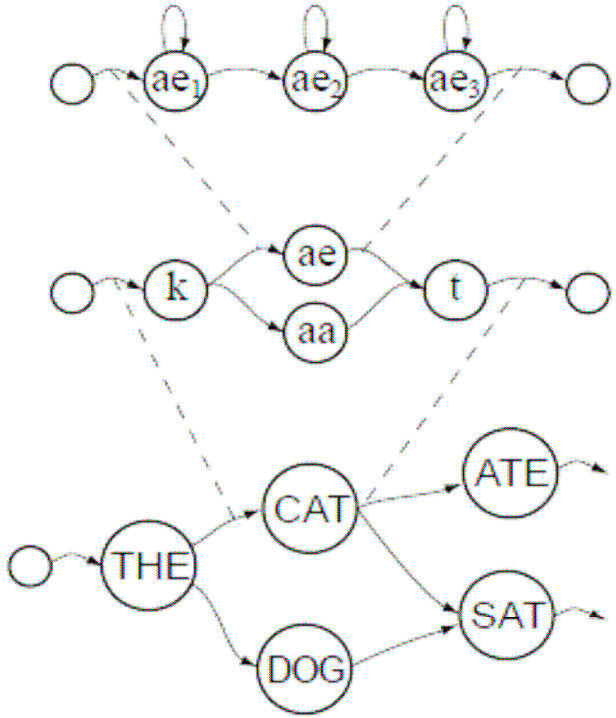


ASR: riconoscimento di pattern

HMM: strutture gerarchiche

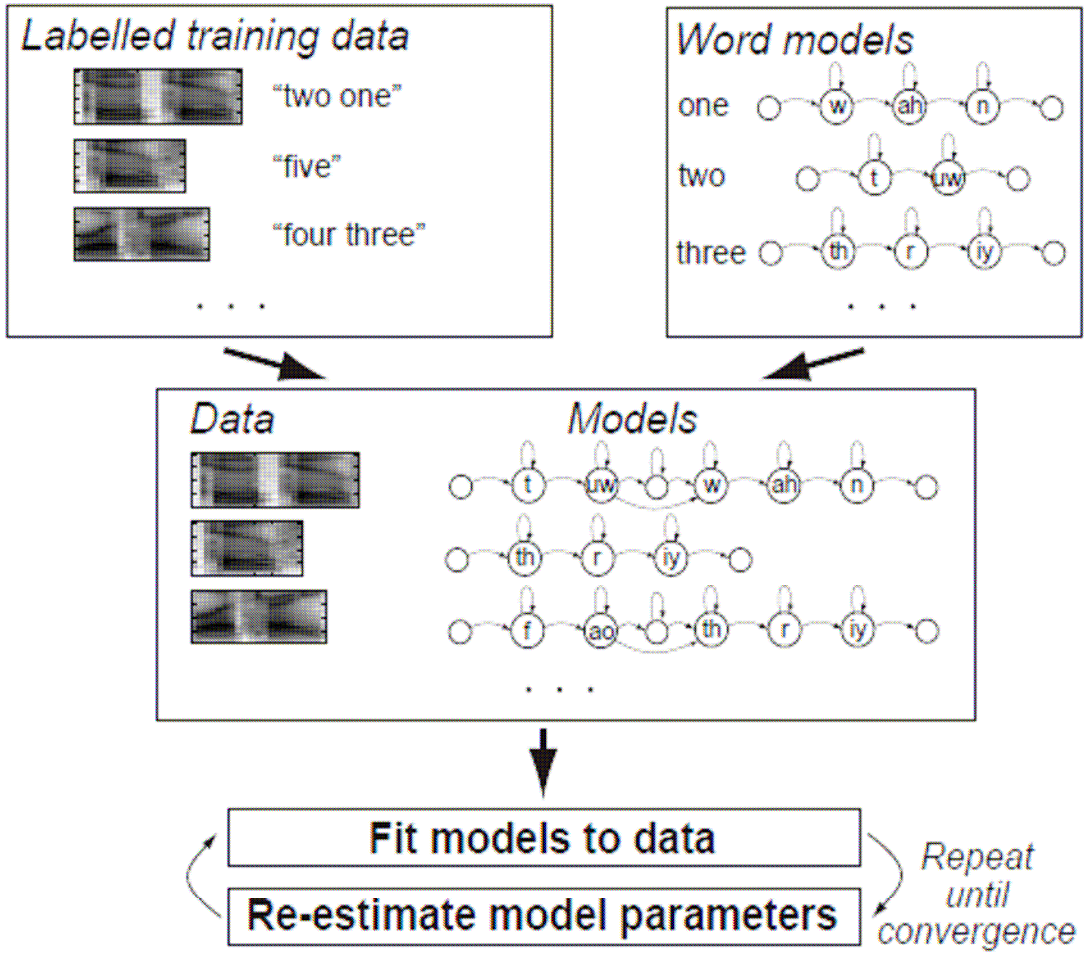
- In generale, vengono usati modelli HMM a tre stati per rappresentare fonemi o difoni, e parole, frasi etc. vengono rappresentate mediante concatenazione di questi

- Il modello HMM consente la composizione secondo strutture gerarchiche



ASR: riconoscimento di pattern

HMM: procedure di training



ASR: riconoscimento di pattern

HMM: procedure di training

- Le procedure di training permettono di calcolare i parametri dei modelli HMM (probabilità di transizione e parametri delle misture gaussiane)

- Le procedure di training consistono in raffinamenti successivi della stima a partire da valori iniziali di bootstrap (ad es., forward-backward, Baum-Welch).

ASR: riconoscimento di pattern

HMM: modelli del linguaggio

- A completamento dell'equazione di Bayes è necessario fornire un modello del linguaggio

- Il modello del linguaggio permette di stimare la probabilità di una parola, date le parole precedenti

- N-grammi:

$$P(w_k|w_{k-1}w_{k-2}w_{k-3}\ldots)$$

- La stime delle probabilità che descrivono gli N-grammi si basano su valutazione delle frequenze di occorenza delle parole nei dati di training

ASR e corpora vocali

Applicazioni del riconoscimento a corpora vocali

- Ricerca delle occorrenze di una parola nel corpus (riconoscimento di parole isolate)

- Allineamento testo-parlato, in caso di disponibilità di versioni testuali

- Conversione parlato-testo scritto per generazione automatica di trascrizione

ASR e corpora vocali

Alcuni Tools disponibili per ASR

- HMM Toolkit (HTK): framework in linguaggio C e linguaggi di scripting per lo sviluppo di applicazioni di ASR
- OGI CSLU Toolkit: framework per ASR, allineamento con il testo, etc.
- Sphinx: framework in linguaggio C e Java per lo sviluppo di applicazioni di ASR

Riferimenti bibliografici

J. R. Deller, J. G. Proakis, and J.H.L. Hansen, *Discrete-time processing of speech signals*, Prentice Hall, 1987.

Acknowledgements

Parte del materiale usato in queste slides è stato adattato dai materiali dal corso Speech Recognition di Michael Mandel, Dip. di Ingegneria Elettrica, Columbia University