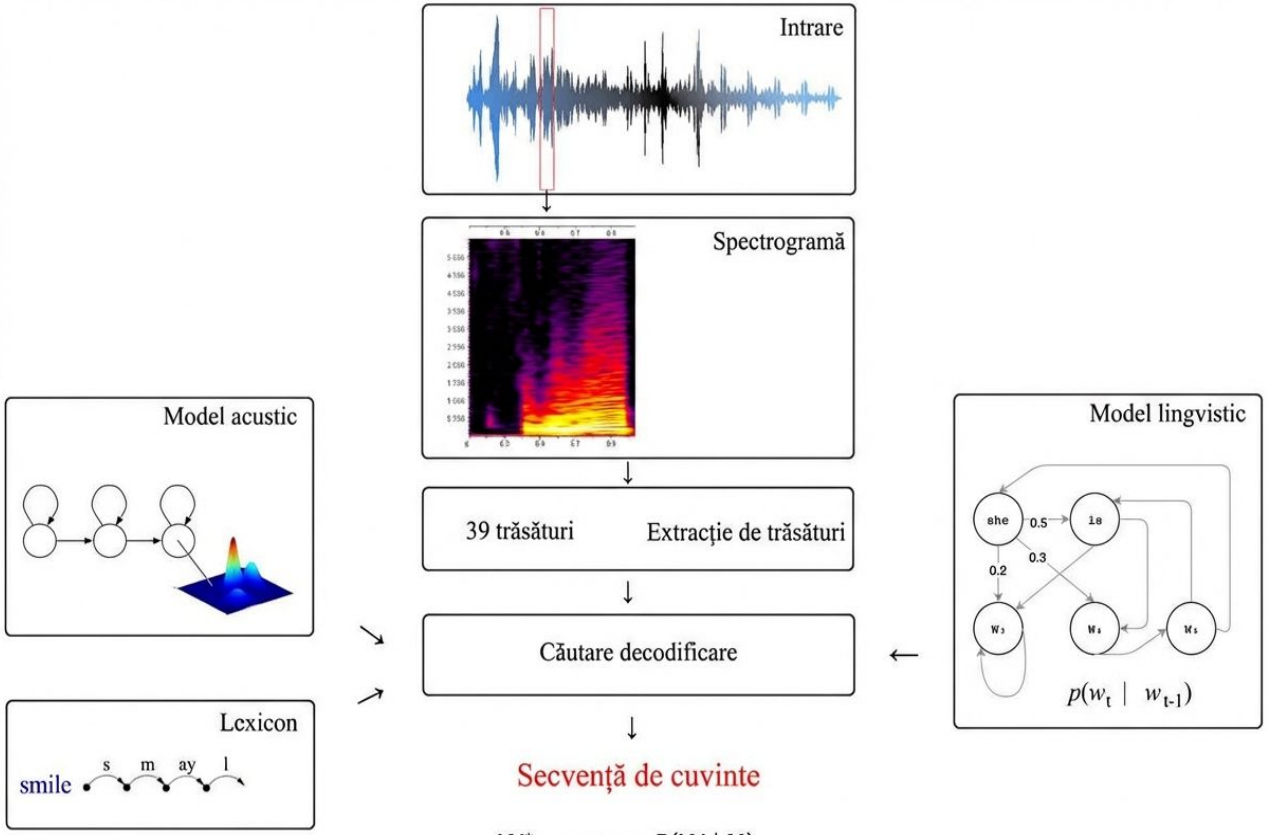
Modelarea acustică

Tudor Bumbu


Intrare
Spectrogramă
39 trăsături
Extracţie de trăsături
Căutare decodificare
Secvenţă de cuvinte
Model acustic
Lexicon
smile
s
m
ay
l
Model lingvistic
0.5
0.3
0.2
$p(w_t \mid w_{t-1})$
$\mathbf{W}^* = \arg\max_{\mathbf{W}} P(\mathbf{W} \mid \mathbf{X})$
$\mathbf{W}^* = \arg\max_{\mathbf{W}} p(\mathbf{X} \mid \mathbf{W}) \; P(\mathbf{W})$
secvenţă de cuvinte
model acustic
model lingvistic

Ce este modelul acustic?

Modelul acustic este o componentă într-un sistem de recunoaștere a vorbirii (ASR).

- Primește ca intrare o secvență de trăsături extrase din semnalul audio.
- Produce probabilități pentru unități lingvistice „mici”
  (foneme, sub-foneme sau stări) pentru fiecare moment de timp.

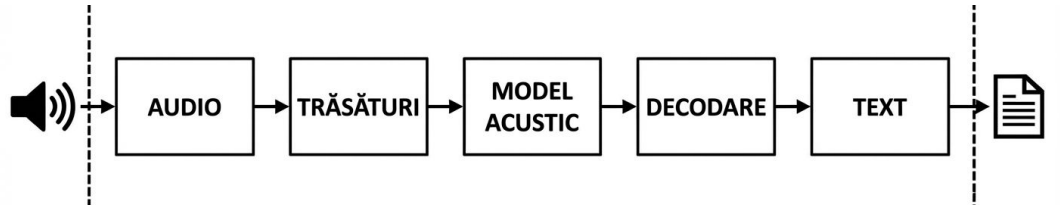
Modelul acustic decide, pentru fiecare clipă, ce sunet este cel mai probabil să se audă.

Decodor

Modelul acustic, de unul singur, nu produce textul final.

Folosim un Decodor care combină mai multe surse de informație:

- Ce „aude” modelul acustic (probabilitățile acustice).
- Restricțiile limbii (modelul lingvistic) care definește ce secvențe de cuvinte sunt corecte gramatical.
- Lexiconul (dicționarul de pronunții).

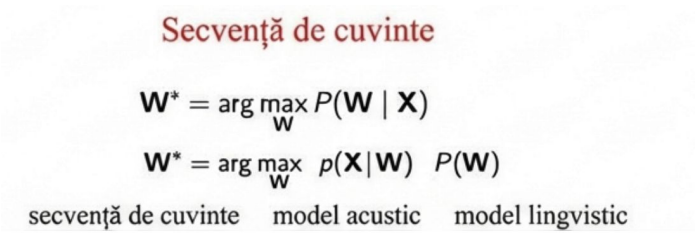


Model Markov Ascuns (HMM)

Vorbirea are loc în timp, iar cuvintele sunt formate dintr-o succesiune de foneme cu durate variabile. Pentru a modela această secvență, se folosește un Model Markov Ascuns (HMM).

- HMM descrie un proces în timp, cu stări ascunse care se succed și observații vizibile (trăsăturile audio).
- Oferă cadrul matematic necesar pentru alinierea temporală și tranzițiile între sunete.

În primele sistemele ASR, probabilitățile acustice erau modelate exclusiv statistic, cu Modele Gaussiene (GMM).



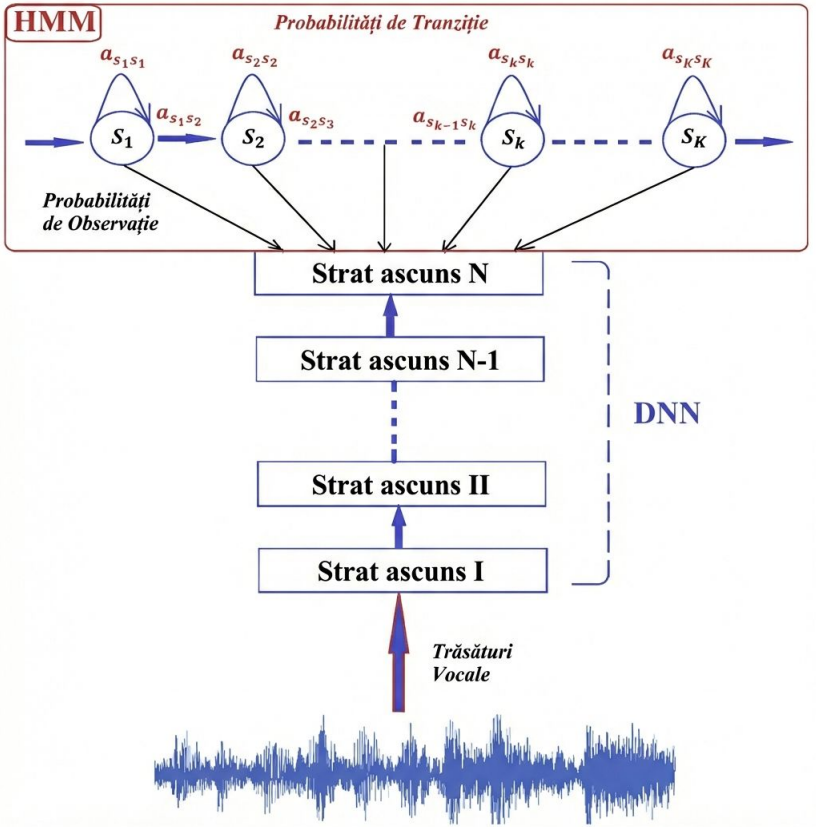
Modele Gaussiene și modelul DNN-HMM

Modelele Gaussiene (GMM) nu captează suficient de bine variabilitatea și complexitatea vocilor umane. O abordare modernă este modelul hibrid:

• Păstrăm HMM pentru partea de secvență (alinierea în timp).
• Înlocuim GMM cu o Rețea Neuronală Profundă (DNN).

La fiecare cadru, rețeaua DNN primește vectorul de trăsături și calculează probabilitatea pentru clase acustice sub-fonetice (numite „senone”).

Modelul DNN-HMM



Contextul

De ce nu putem clasifica fiecare cadru audio izolat, folosind o rețea neuronală clasică? Vorbirea are o structură dinamică, iar un fonem nu apare ca un element izolat.

• Coarticulația: Sunetele se influențează reciproc în timp.
• De exemplu, vocala „a” va suna diferit și va avea o formă diferită pe spectrogramă în funcție de consoanele din jur.

Pentru decizii corecte, modelul are nevoie de ***context***
(să rețină ce s-a auzit înainte și după).

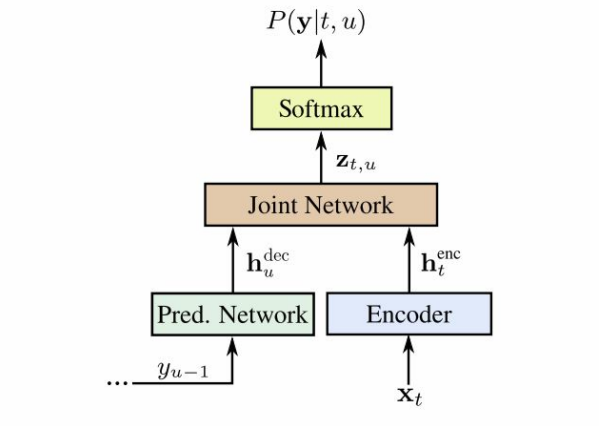
Pentru a modela dependențele temporale, utilizăm
Rețele Neuronale Recurente (RNN).

Rețele Neuronale Recurente (RNN)

Pentru a modela dependențele temporale, utilizăm Rețele Neuronale Recurente (RNN).

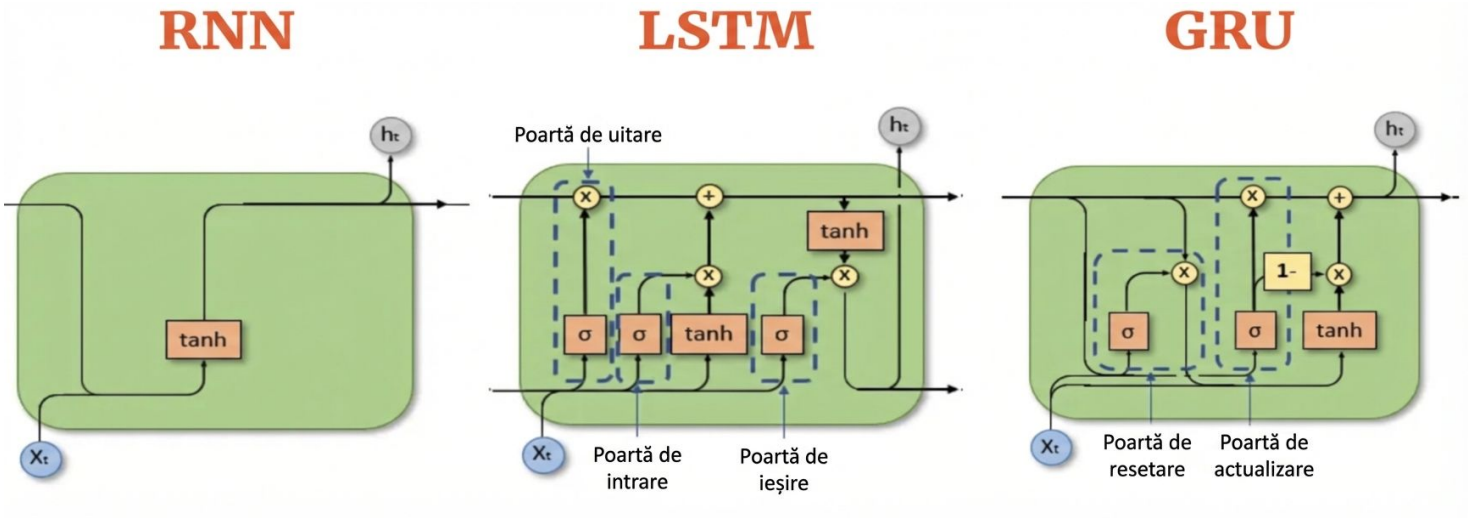
RNN introduc o formă de „memorie”:

La pasul curent, RNN folosește intrarea actuală și o „stare internă” (hidden state) care păstrează informații din pașii procesați anterior.



Memorie controlată cu LSTM și GRU

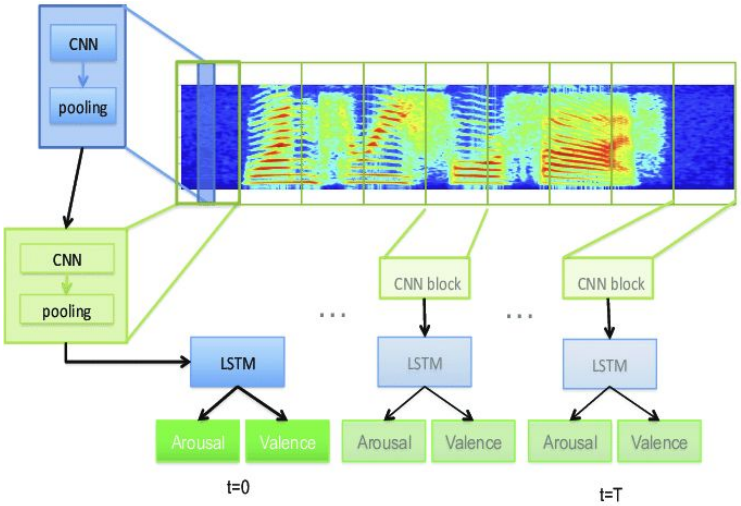
Soluția pentru limitările RNN-urilor o reprezintă arhitecturile cu „porți” (gates):
LSTM (Long Short-Term Memory) și ***GRU (Gated Recurrent Unit)***.



Arhitecturi combinate: CNN + LSTM

CNN (Rețele convoluționale)
Extrag excelent tiparele locale stabile din spectrogramă.

LSTM (Rețele recurente)
Prelucrează aceste trăsături extrase de CNN și modelează evoluția lor în timp.



Evaluarea unui sistem ASR

Cea mai folosită metrică pentru a evalua calitatea unui sistem de recunoaștere a vorbirii este ***WER*** (Word Error Rate - rata de eroare pe cuvinte).

WER măsoară numărul de „editări” necesare pentru a transforma transcrierea automată în textul corect:

- Substituții (S): Un cuvânt a fost înlocuit greșit cu altul.
- Ștergeri (D): Un cuvânt rostit a fost ignorat (omis).
- Inserții (I): Modelul a inventat un cuvânt în plus.

$$WER = (S + D + I) / N$$

(unde N este numărul total de cuvinte).

Cu cât WER este mai mic, cu atât sistemul e mai precis.