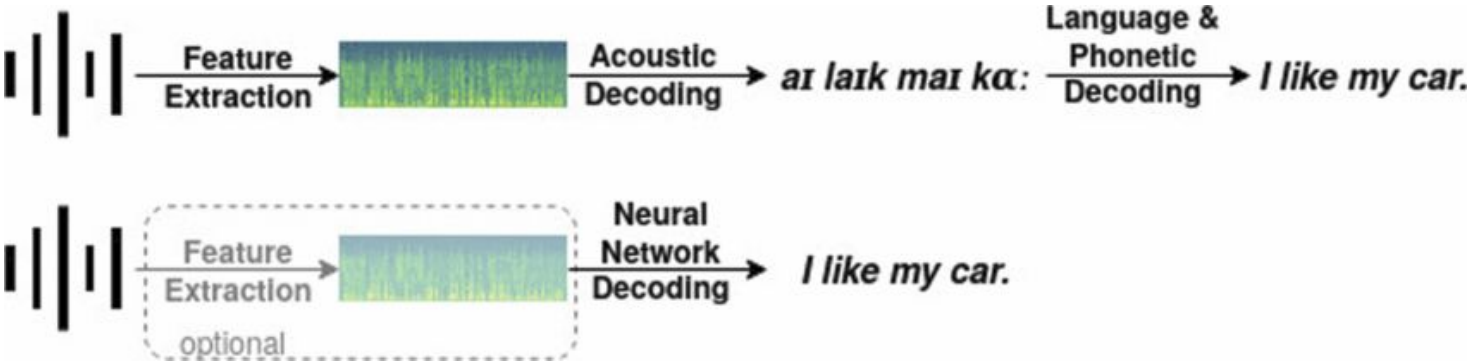
Paradigme moderne în ASR

Tudor Bumbu

Ce înseamnă ASR „end-to-end”?

Spre deosebire de sistemele clasice (modulare), ASR *end-to-end* încearcă să învețe maparea directă de la audio la text, folosind o singură rețea neuronală mare.

- Nu separăm sistemul în componente antrenate izolat (dicționar de pronunții, model de limbaj, aliniere forțată).
- Modelul învață regulile direct din date.



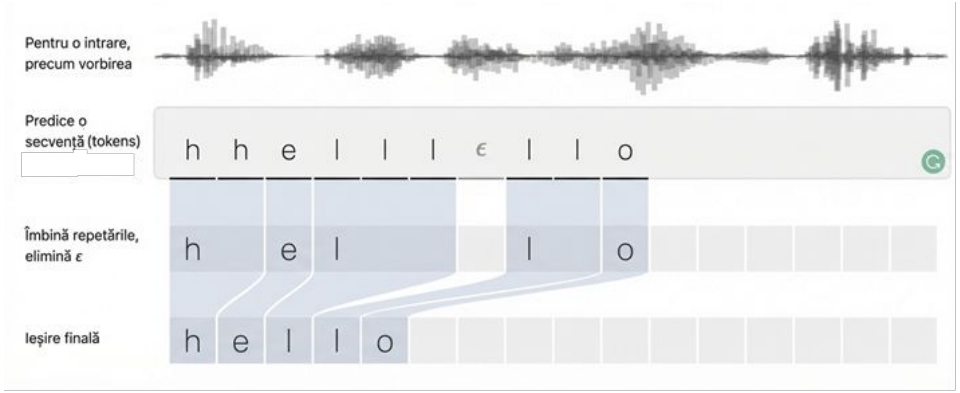
Alinierea cu CTC

Într-o abordare end-to-end, fișierul audio este lung, iar textul este scurt. Cum le potrivim?

CTC (Clasificarea Temporală Conexionistă) permite antrenarea fără a ști exact unde începe și se termină o literă.

• Introduce un simbol special numit „blank” (spațiu gol).

• Modelul folosește acest simbol pentru a acoperi pauzele și a rezolva automat problema alinierii.

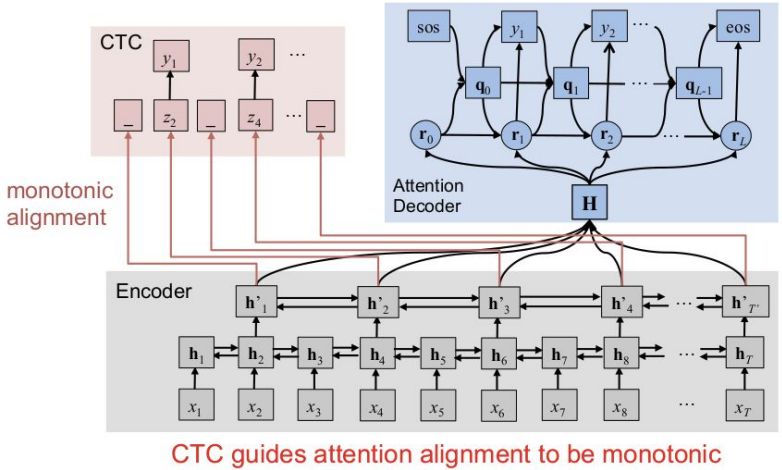


Encoder-Decoder cu Atenție

A doua paradigmă majoră end-to-end tratează recunoașterea vorbirii ca pe o „traducere” din audio în text.

- ***Encoderul*** citește semnalul audio și îl transformă într-o reprezentare latentă bogată.
- ***Decoderul*** generează textul pas cu pas (caracter cu caracter sau cuvânt cu cuvânt).
- ***Mecanismul de atenție*** indică decoderului „la ce moment exact din fișierul audio să se uite” atunci când generează următorul simbol.

Multitask learning: $\mathcal{L}_{\text{MTL}} = \lambda \mathcal{L}_{\text{CTC}} + (1-\lambda)\mathcal{L}_{\text{Attention}}$

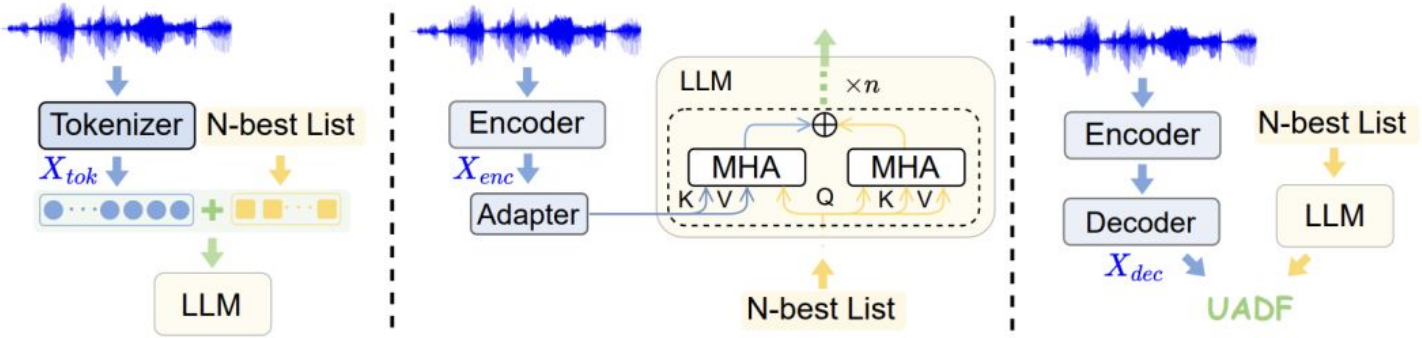


CTC guides attention alignment to be monotonic

Modele Mari și Învățare Auto-supervizată

Sistemele ASR moderne se bazează pe modele de dimensiuni masive, antrenate prin „învățare auto-supervizată” pe mii de ore de audio neetichetat din internet.

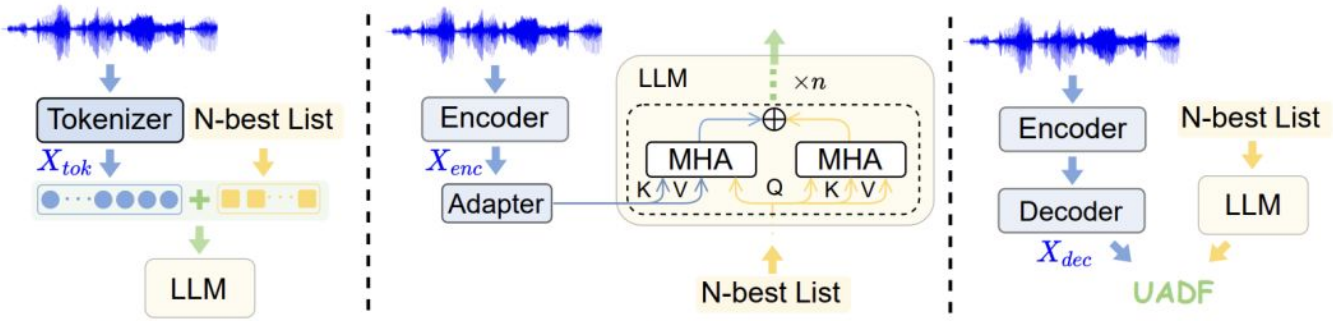
- Un exemplu reprezentativ este modelul open-source Whisper (OpenAI).
- Folosind arhitectura Transformer, acesta transcrie și traduce peste 100 de limbi cu o precizie ridicată.



ASR în regim de Streaming

Pentru anumite aplicații (subtitrări live, asistenți vocali), nu putem aștepta ca utilizatorul să termine propoziția.

- Sistemele de Streaming transcriu textul pe măsură ce omul vorbește, în timp real, cu o latență minimă.
- Folosesc arhitecturi dedicate (cum ar fi modelele Conformer).
- Cercetarea actuală integrează direct fluxuri audio cu modele LLM, pentru interacțiuni conversaționale imediate.



Ecosistemul ASR în aplicații reale

În practică, un sistem ASR integrează module auxiliare esențiale:

- VAD (Voice Activity Detection): Detectează automat momentele cu vorbire și elimină tăcerea.
- Modele lingvistice: Adaugă automat punctuația și majusculele pe textul transcris.
- Diarizare: Detectează și separă vorbitorii (cine și când a vorbit).
- Traducere simultană (Speech Translation): Traduce textul în timp real într-o altă limbă.