

Sistem de transcriere automată a textului în vorbire pentru limba Română folosind tehnici de învățare adâncă

Ionescu Sergiu Cristian

Coordonatori:

Prof. Univ. Dr. Ing. Corneliu Burileanu
As. Univ. Drd. Ing. Ana-Antonia Neacșu

Facultatea de Electronică, Telecomunicații și Tehnologia
Informației

Iulie 2021



Cuprins

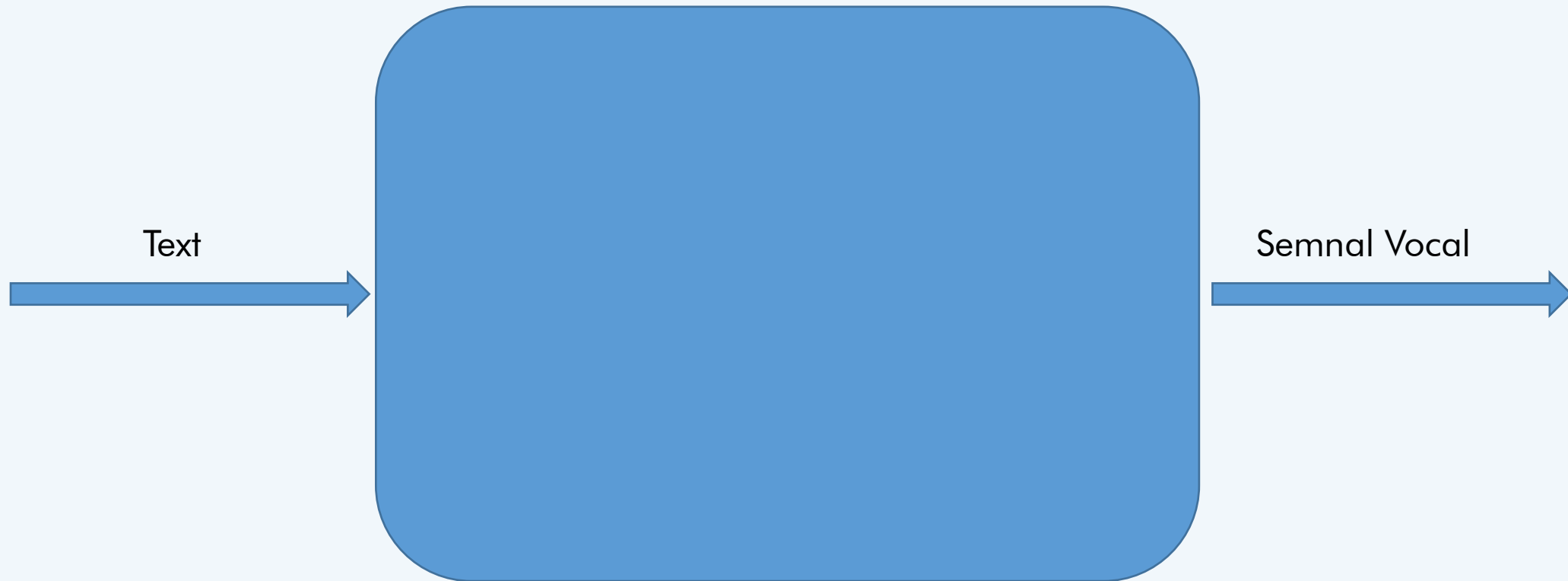
- Motivație
- Noțiuni introductive
- Testarea Bazelor de date
- Modelul propus
- Concluzii

Motivație

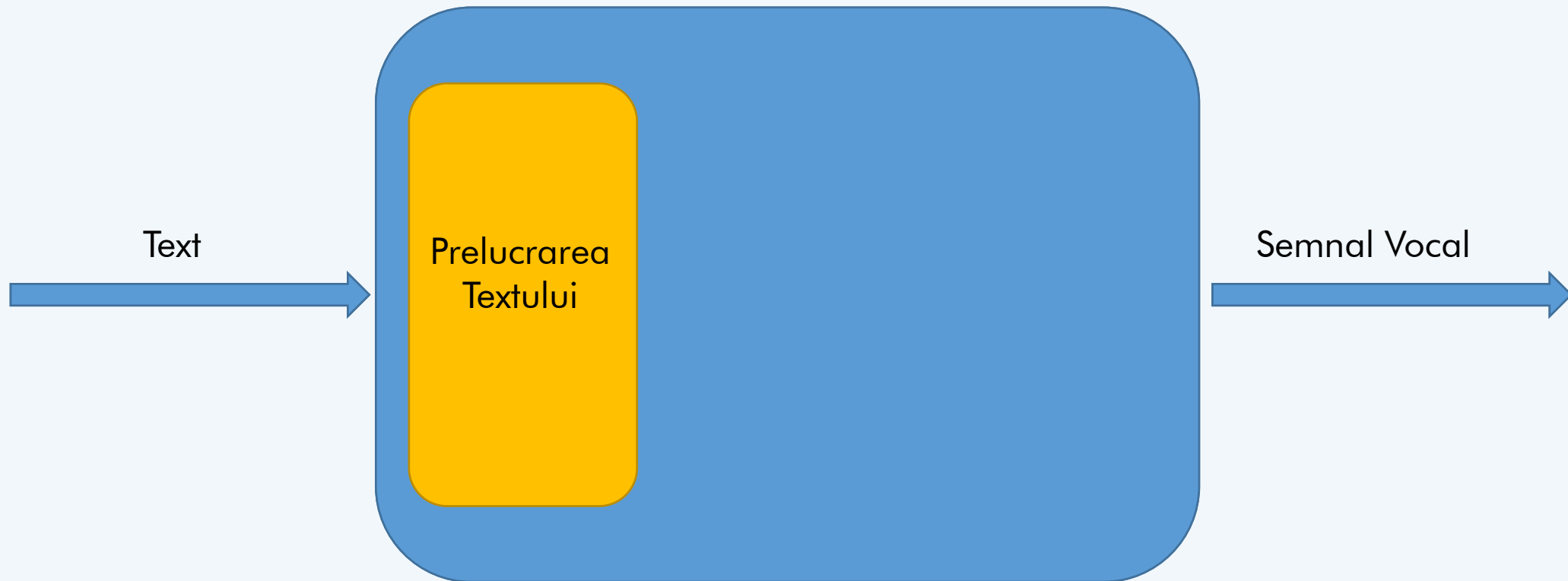
- Evoluția rețelelor neurale
- Domeniul de procesare al limbii române utilizând rețele neurale este unul nou și încă neexplorat.
- Implementarea unui astfel de sistem pe un robot

Noțiuni introductive

Sistemele Text to Speech



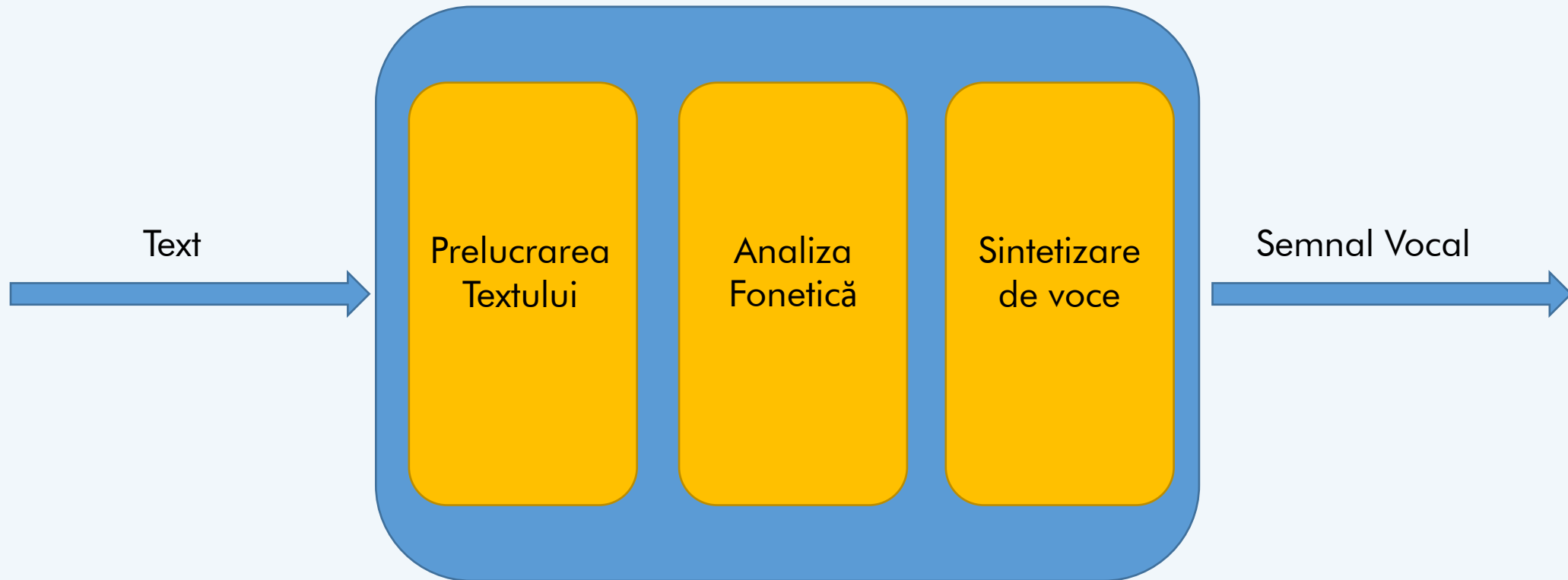
Sistemele Text to Speech



Sistemele Text to Speech



Sistemele Text to Speech



Prelucrarea textului

Prelucrarea textului presupune 2 etape:

- Normalizarea:
 - ☐ eliminarea caracterelor ce nu influențează pronunția
 - ☐ expandarea simbolurilor, abrevierilor și a numerelor
 - ☐ trecerea literelor mari în litere mici

Studentul are 221341 de foi !!! A ales să le lase acasă spunându-și "ce bine că nu le-am luat"...



studentul are două sute douăzeci și unu de mii trei sute patruzeci și unu de foi . a ales să le lase acasă spunânduși ce bine că nu leam luat .

Prelucrarea textului presupune 2 etape:

- Normalizarea:
 - ☐ eliminarea caracterelor ce nu influenteaza pronuntia
 - ☐ expandarea simbolurilor, abrevierilor si a numerelor
 - ☐ trecerea literelor mari in litere mici
- Reprezentarea vectorială a caracterelor (character embedding)

el_of t a

5 12 1 14 6 20 2

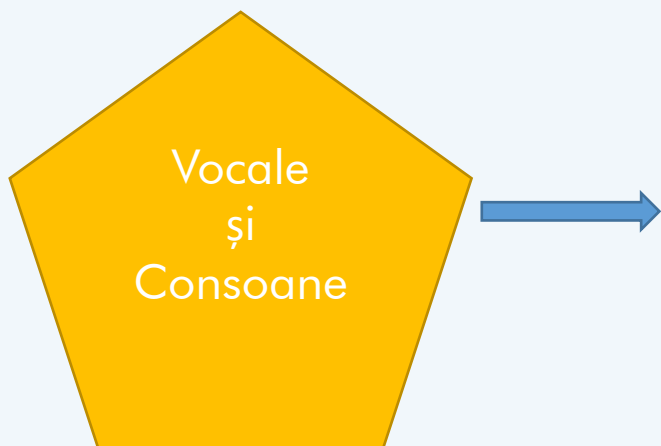
[0.2 0.3 0.6] [0.1 0.7 0.8] [0.3 0.1 0.4] [0.6 0.2 0.3] [0.7 0.8 0.1] [0.8 0.7 0.3] [0.1 0.2 0.3]

Analiza fonetică

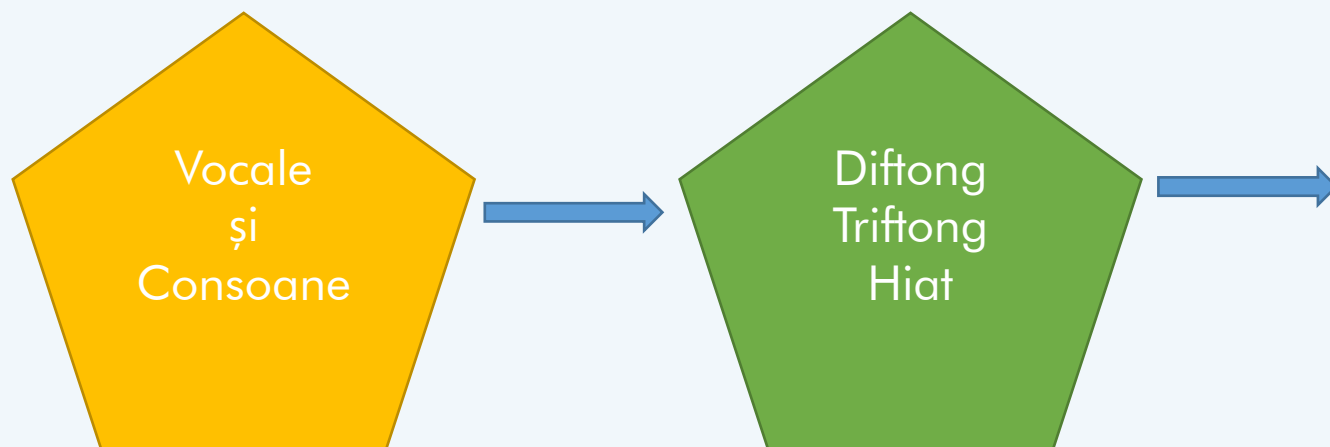
Ne propunem identificarea dependențelor între caractere.



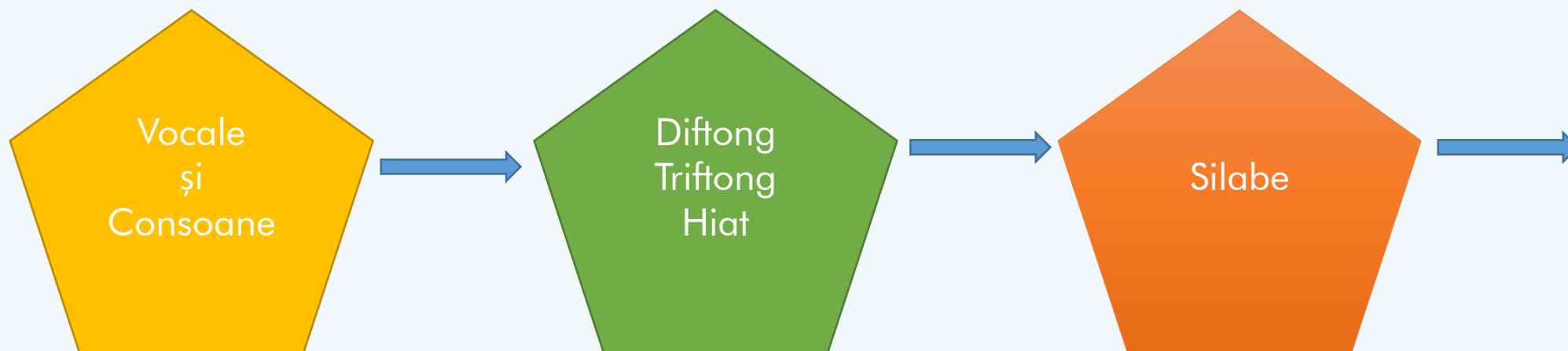
Ne propunem identificarea dependențelor între caractere.



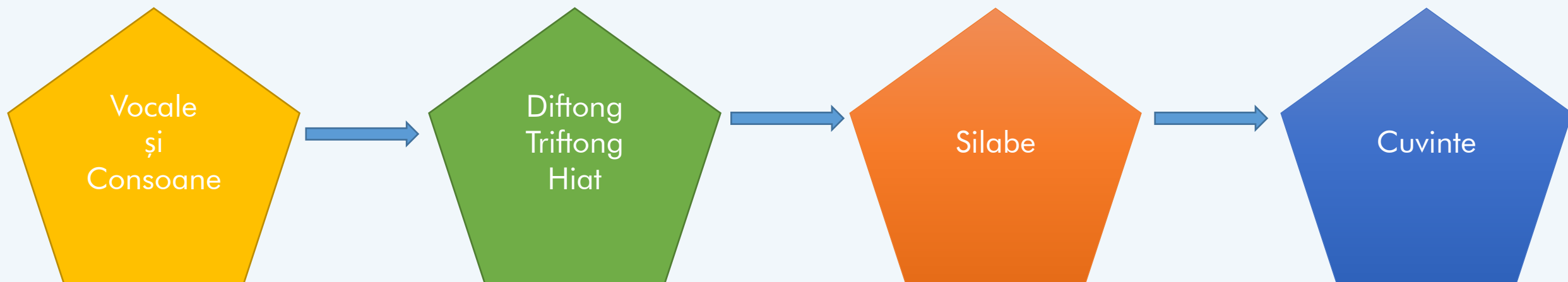
Ne propunem identificarea dependențelor între caractere.



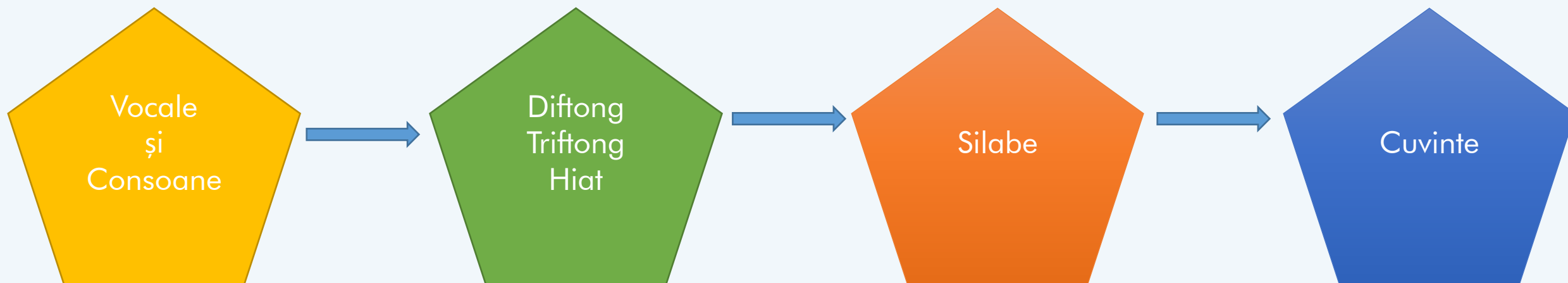
Ne propunem identificarea dependențelor între caractere.



Ne propunem identificarea dependențelor între caractere.



Ne propunem identificarea dependențelor între caractere.



Vom utiliza rețele convoluționale unidimensionale.

0.73

[0.2 0.3 0.6] [0.1 0.7 0.8] [0.3 0.1 0.4] [0.6 0.2 0.3] [0.7 0.8 0.1] [0.8 0.7 0.3] [0.1 0.2 0.3]

5 12 1 14 6 20 2

el_ofta

Kernel

$$\begin{bmatrix} 3 & 1 & 2 \\ -1 & 2 & 1 \\ 1 & 1 & -1 \end{bmatrix}$$



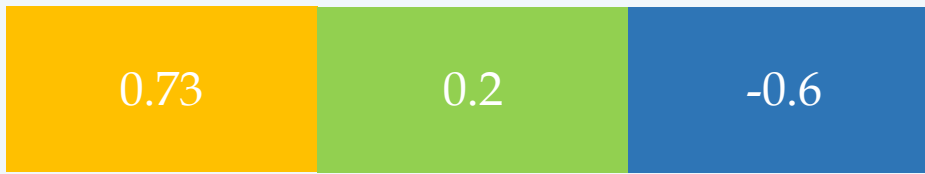
[0.2 0.3 0.6] [0.1 0.7 0.8] [0.3 0.1 0.4] [0.6 0.2 0.3] [0.7 0.8 0.1] [0.8 0.7 0.3] [0.1 0.2 0.3]

5 12 1 14 6 20 2

el_ofta

Kernel

$$\begin{bmatrix} 3 & 1 & 2 \\ -1 & 2 & 1 \\ 1 & 1 & -1 \end{bmatrix}$$



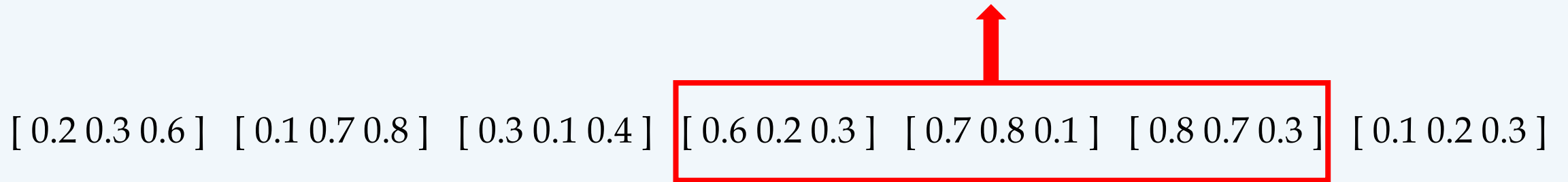
[0.2 0.3 0.6] [0.1 0.7 0.8] [0.3 0.1 0.4] [0.6 0.2 0.3] [0.7 0.8 0.1] [0.8 0.7 0.3] [0.1 0.2 0.3]

5 12 1 14 6 20 2

el_ofta

Kernel

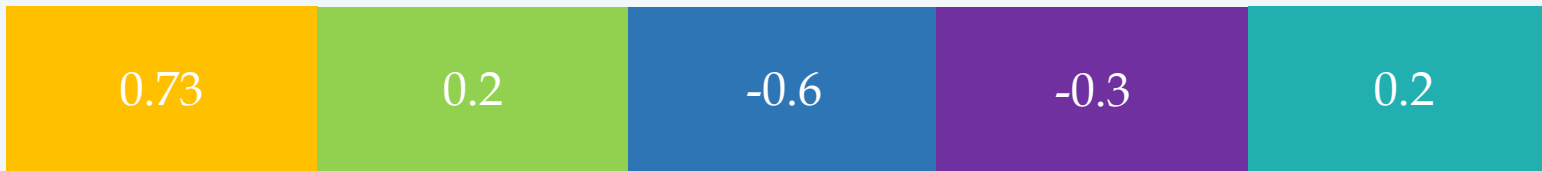
$$\begin{bmatrix} 3 & 1 & 2 \\ -1 & 2 & 1 \\ 1 & 1 & -1 \end{bmatrix}$$



5 12 1 14 6 20 2

el_ofta

Kernel $\begin{bmatrix} 3 & 1 & 2 \\ -1 & 2 & 1 \\ 1 & 1 & -1 \end{bmatrix}$

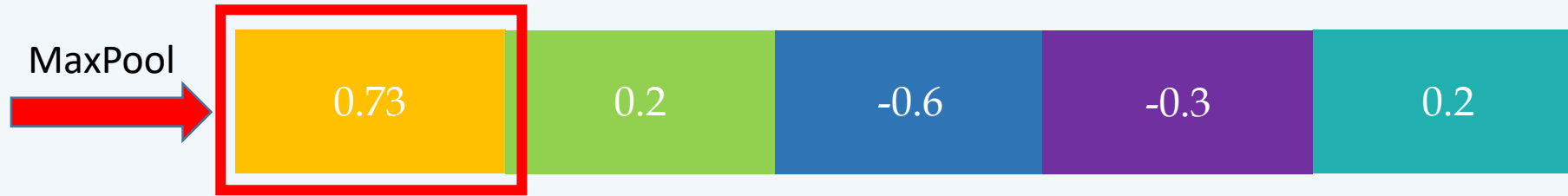


[0.2 0.3 0.6] [0.1 0.7 0.8] [0.3 0.1 0.4] [0.6 0.2 0.3] [0.7 0.8 0.1] [0.8 0.7 0.3] [0.1 0.2 0.3]

5 12 1 14 6 20 2

el_ofta

Kernel $\begin{bmatrix} 3 & 1 & 2 \\ -1 & 2 & 1 \\ 1 & 1 & -1 \end{bmatrix}$



[0.2 0.3 0.6] [0.1 0.7 0.8] [0.3 0.1 0.4] [0.6 0.2 0.3] [0.7 0.8 0.1] [0.8 0.7 0.3] [0.1 0.2 0.3]

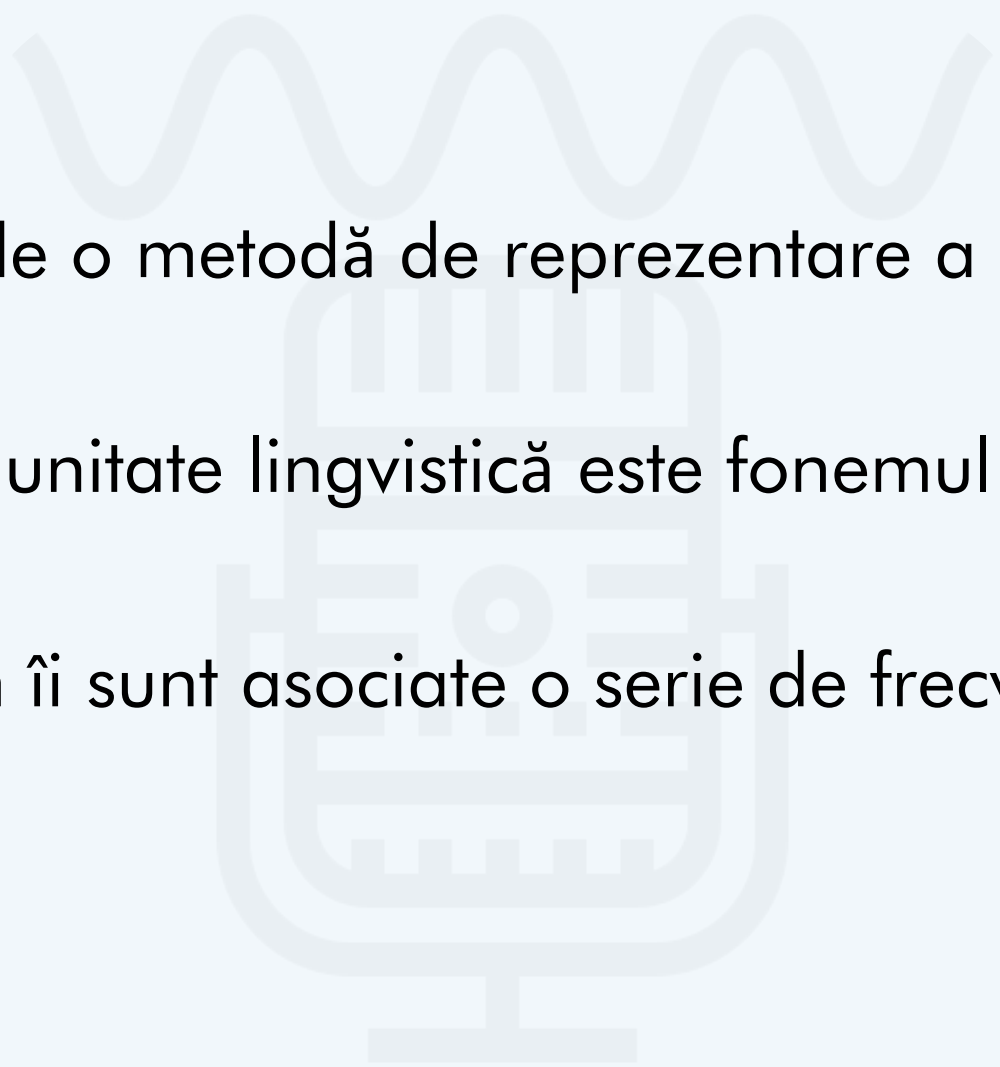
5 12 1 14 6 20 2

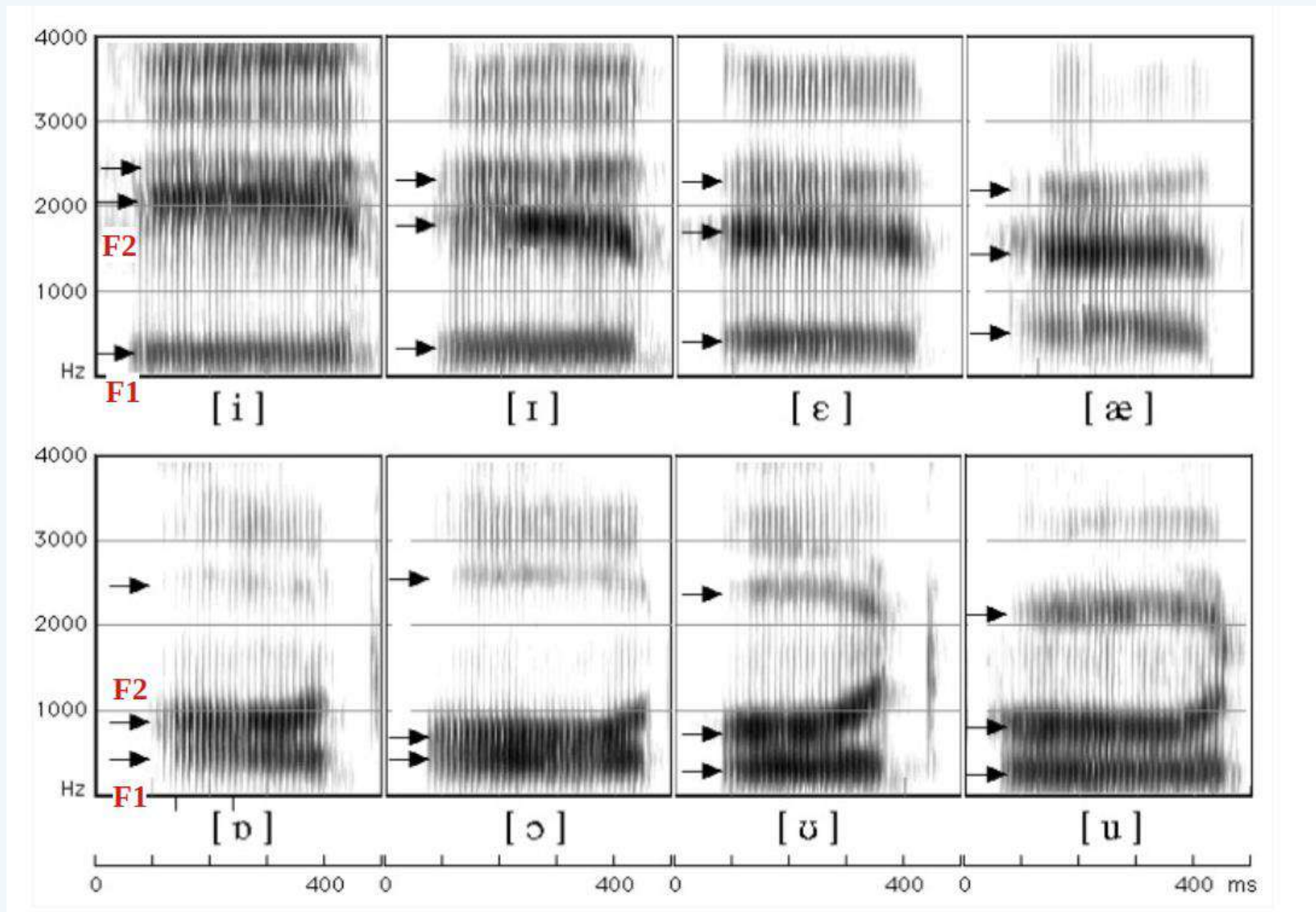
elofta

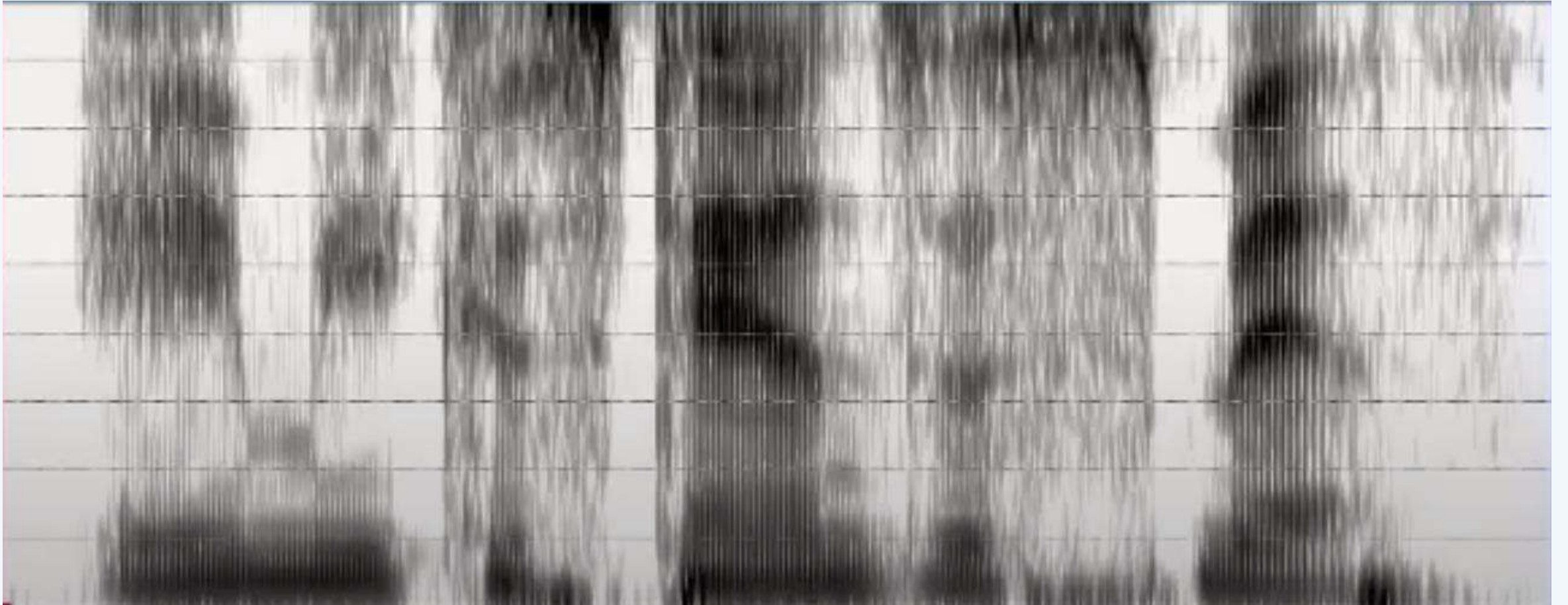
Kernel $\begin{bmatrix} 3 & 1 & 2 \\ -1 & 2 & 1 \\ 1 & 1 & -1 \end{bmatrix}$

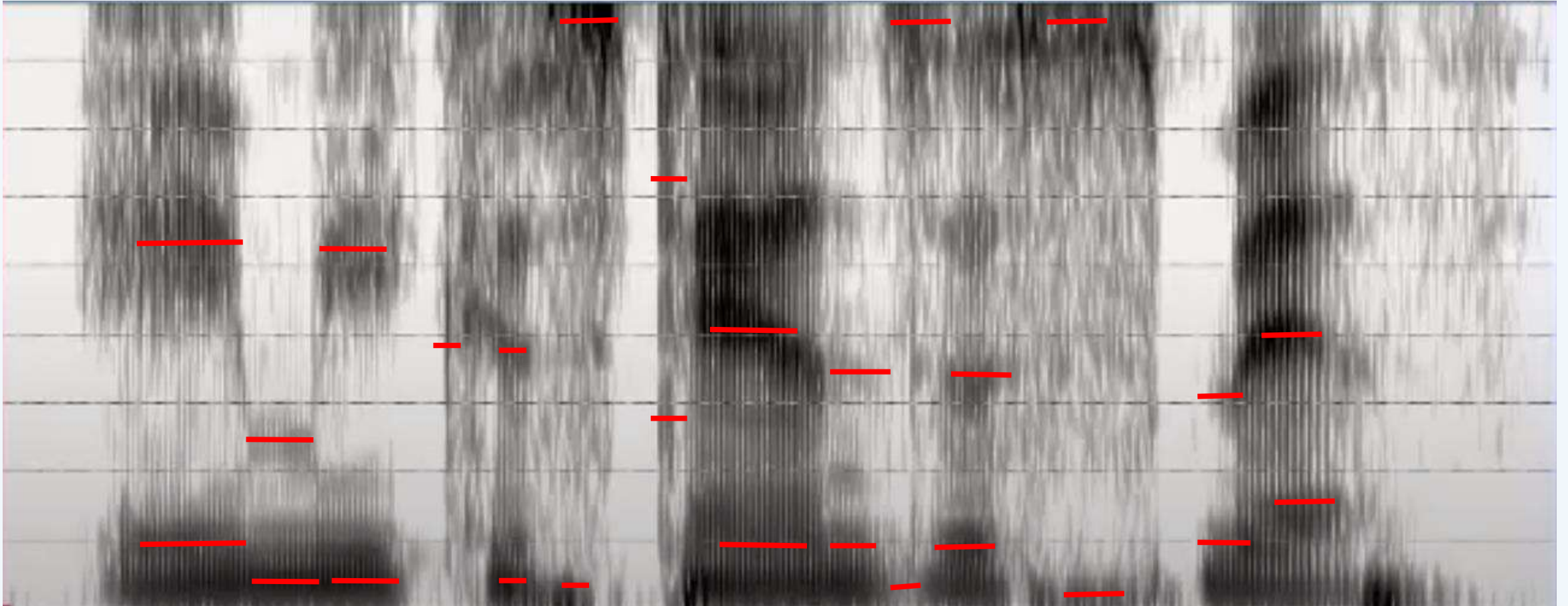


Sintetizarea de voce

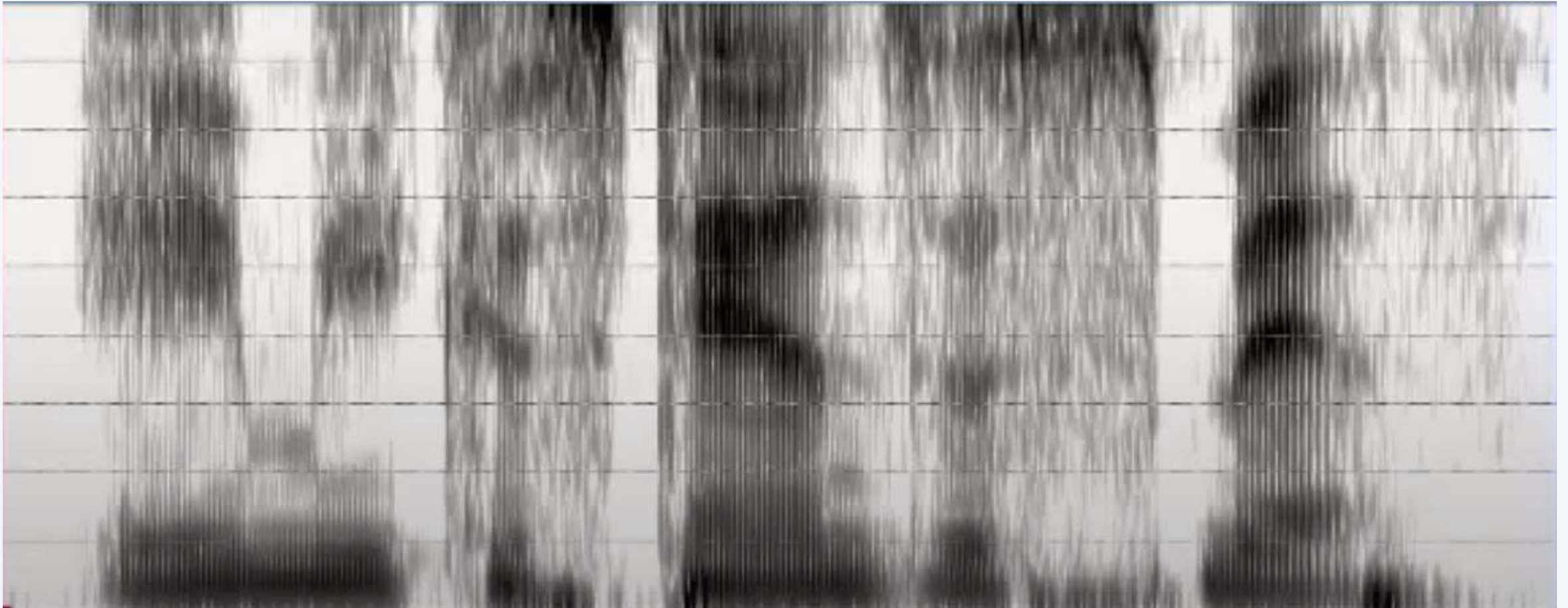
- 
- Avem nevoie de o metodă de reprezentare a sunetului
 - Cea mai mică unitate lingvistică este fonemul
 - Fiecărui fonem îi sunt asociate o serie de frecvențe formante







Amy Constanza Smith

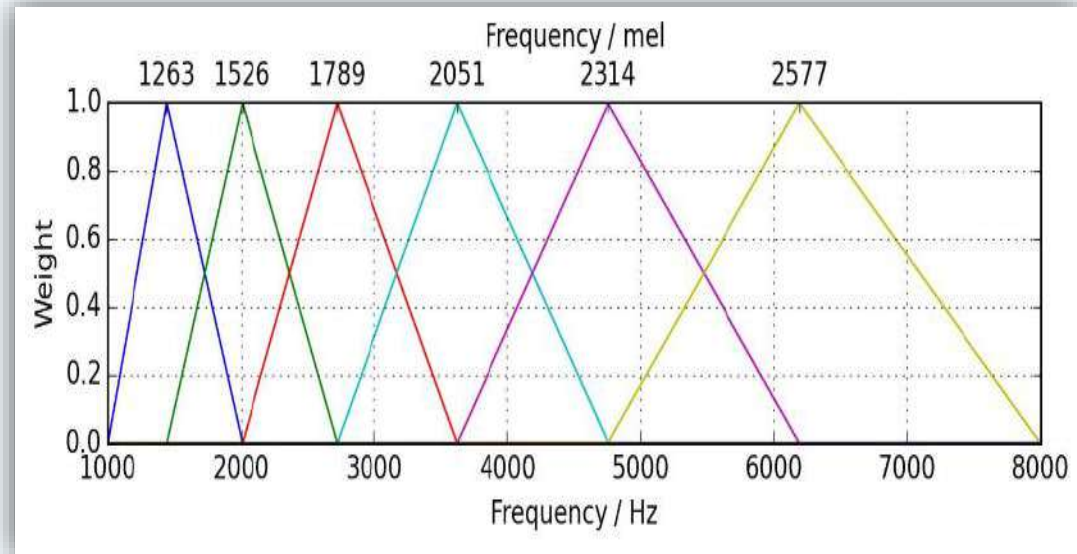


Reprezentarea este una liniară, în mod diferit față de cum percepem noi sunetele.

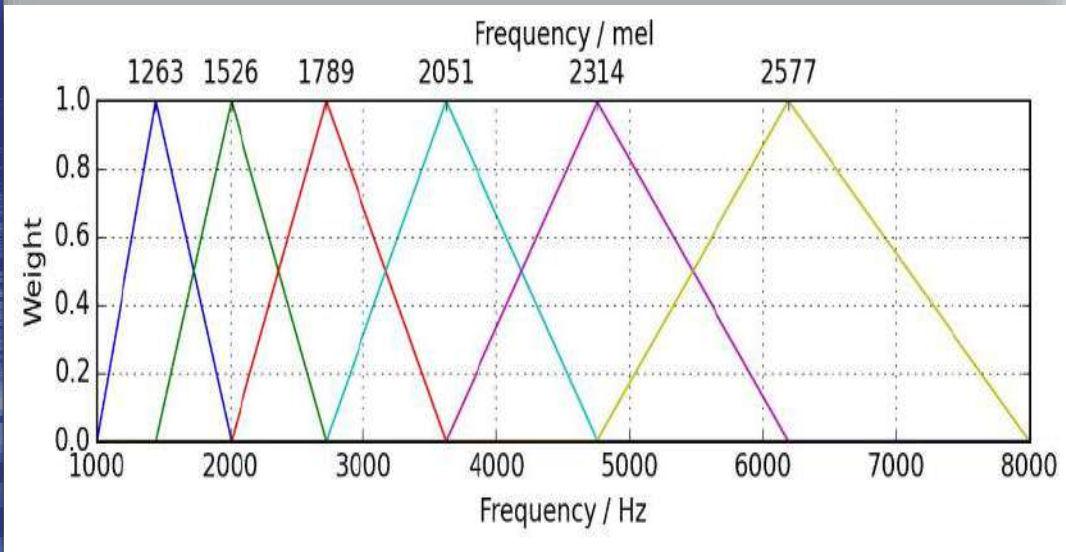
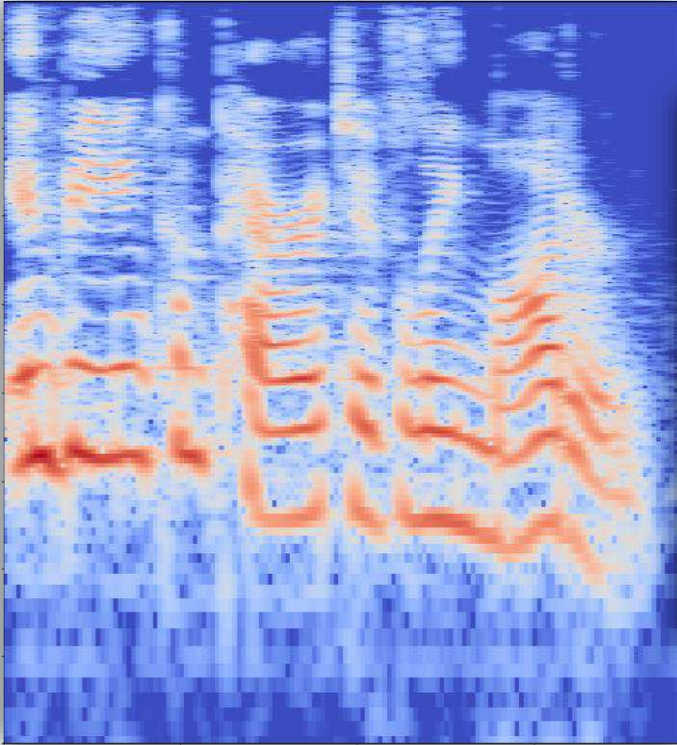
Spectrograma Mel

- dorim o reprezentare a frecvenței în timp bună
- dorim o reprezentare relevantă a amplitudinii în funcție de percepția noastră.
- dorim o reprezentare relevantă a frecvenței după modul în care și percepem (convertirea axei de frecvență)

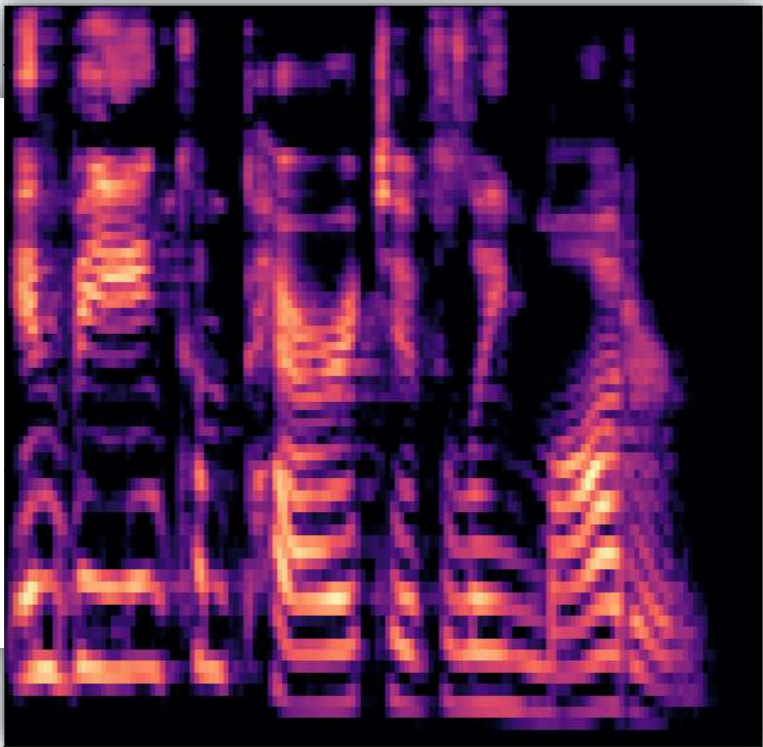
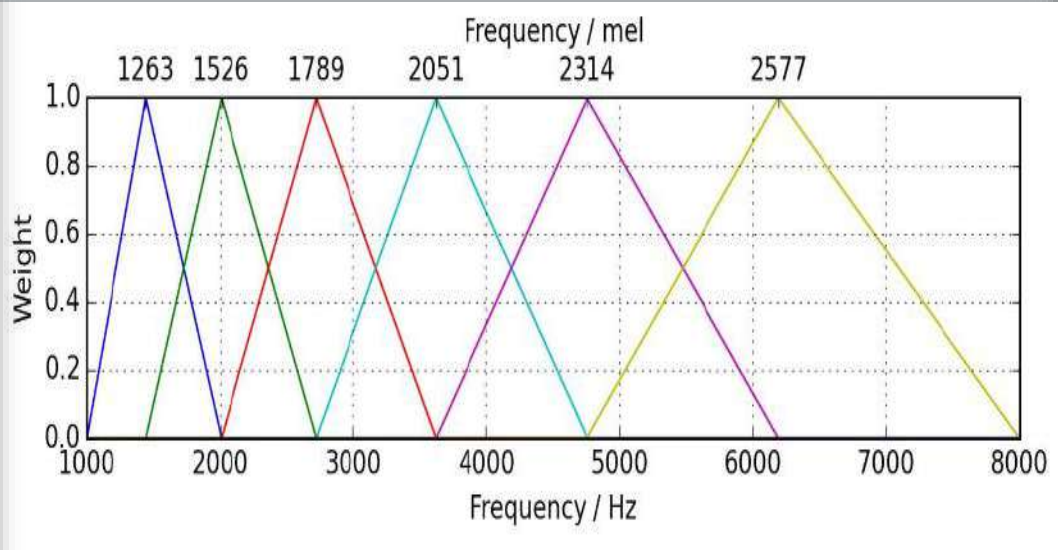
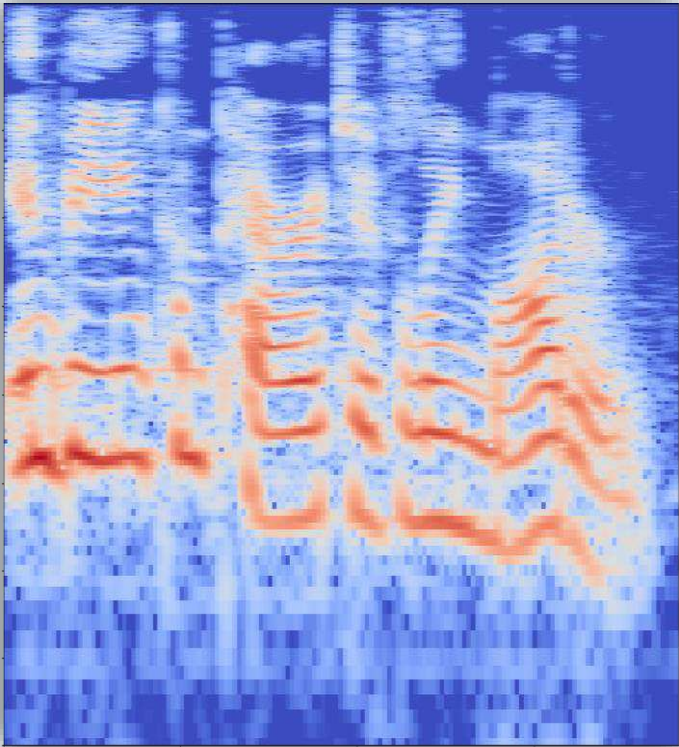
Proiectăm filtrele Mel

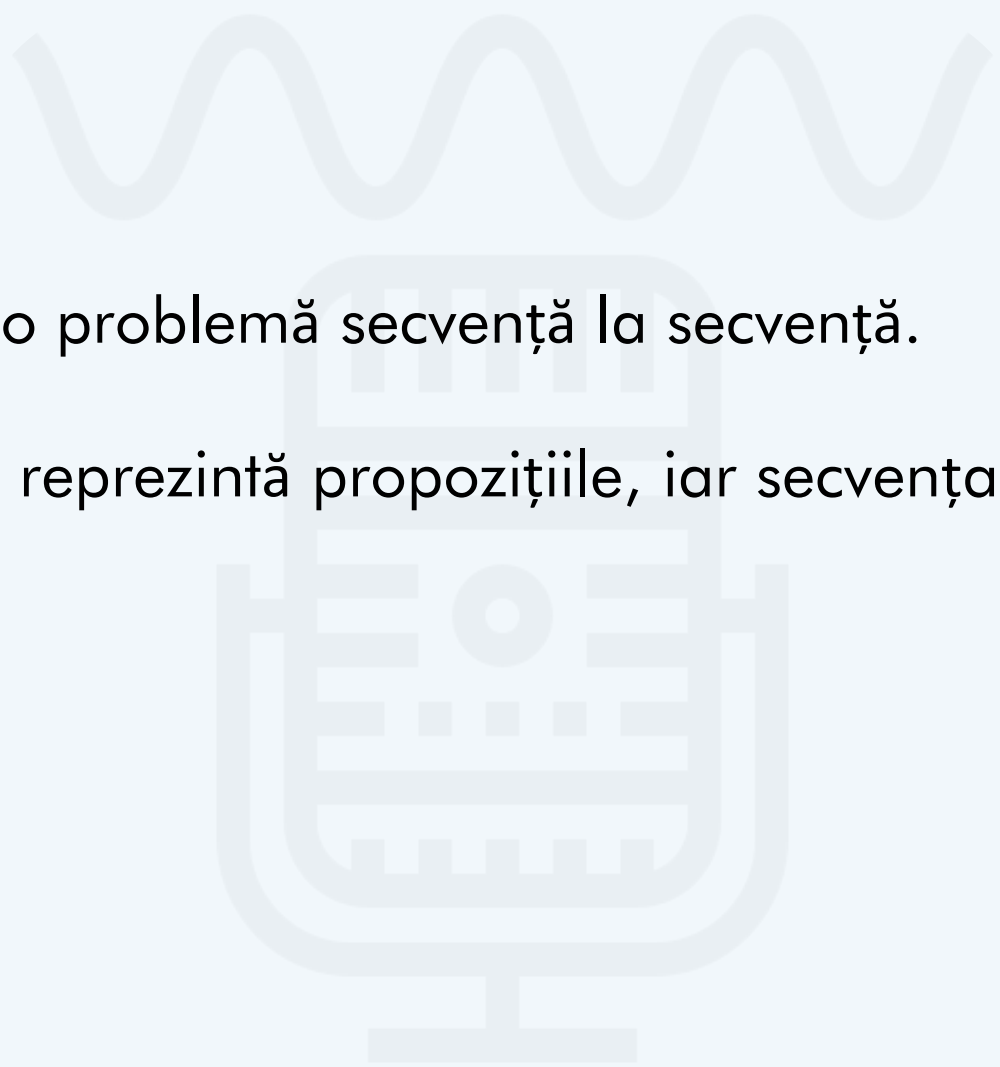


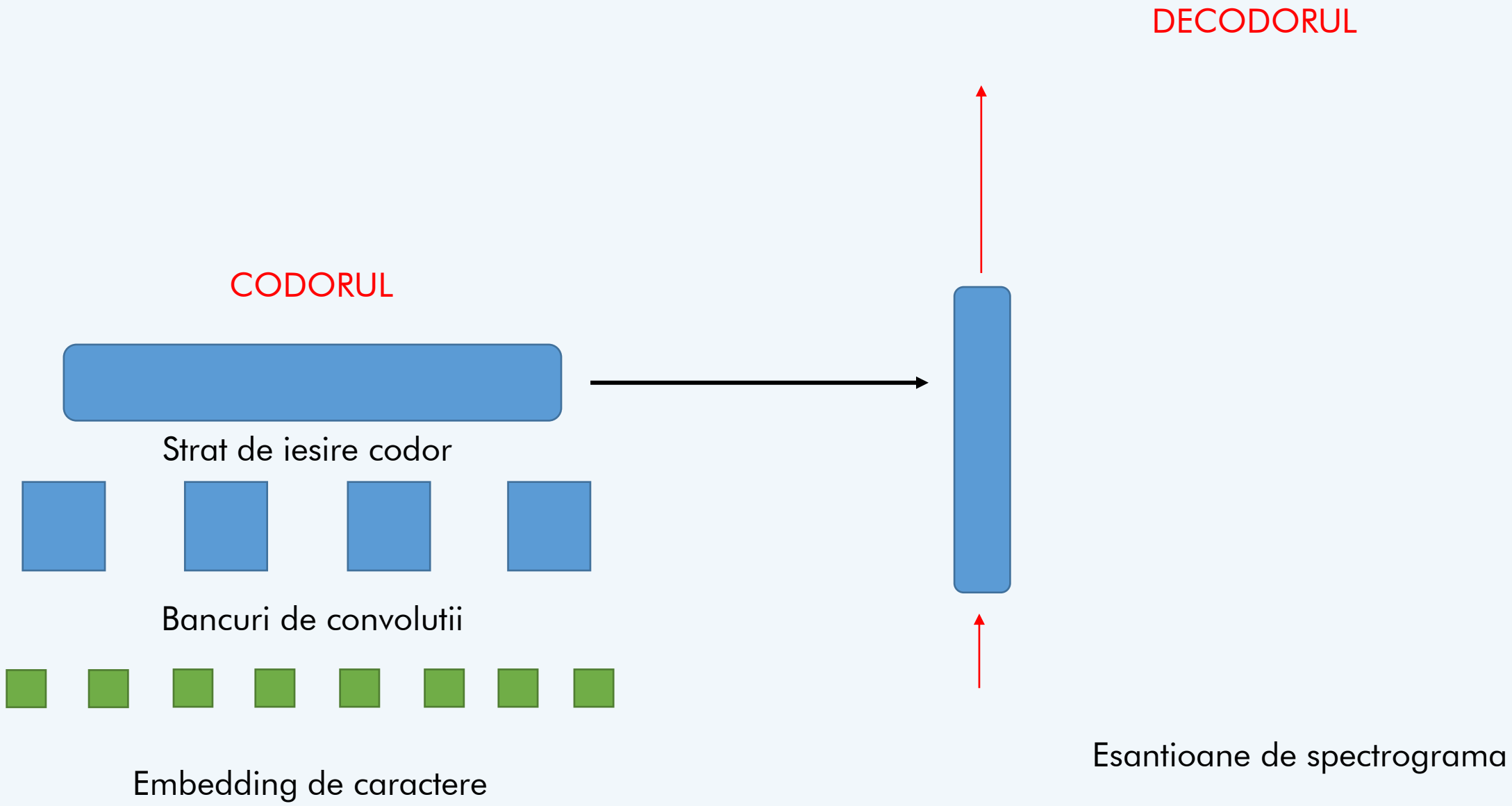
Le aplicăm pe Spectrograma Liniară

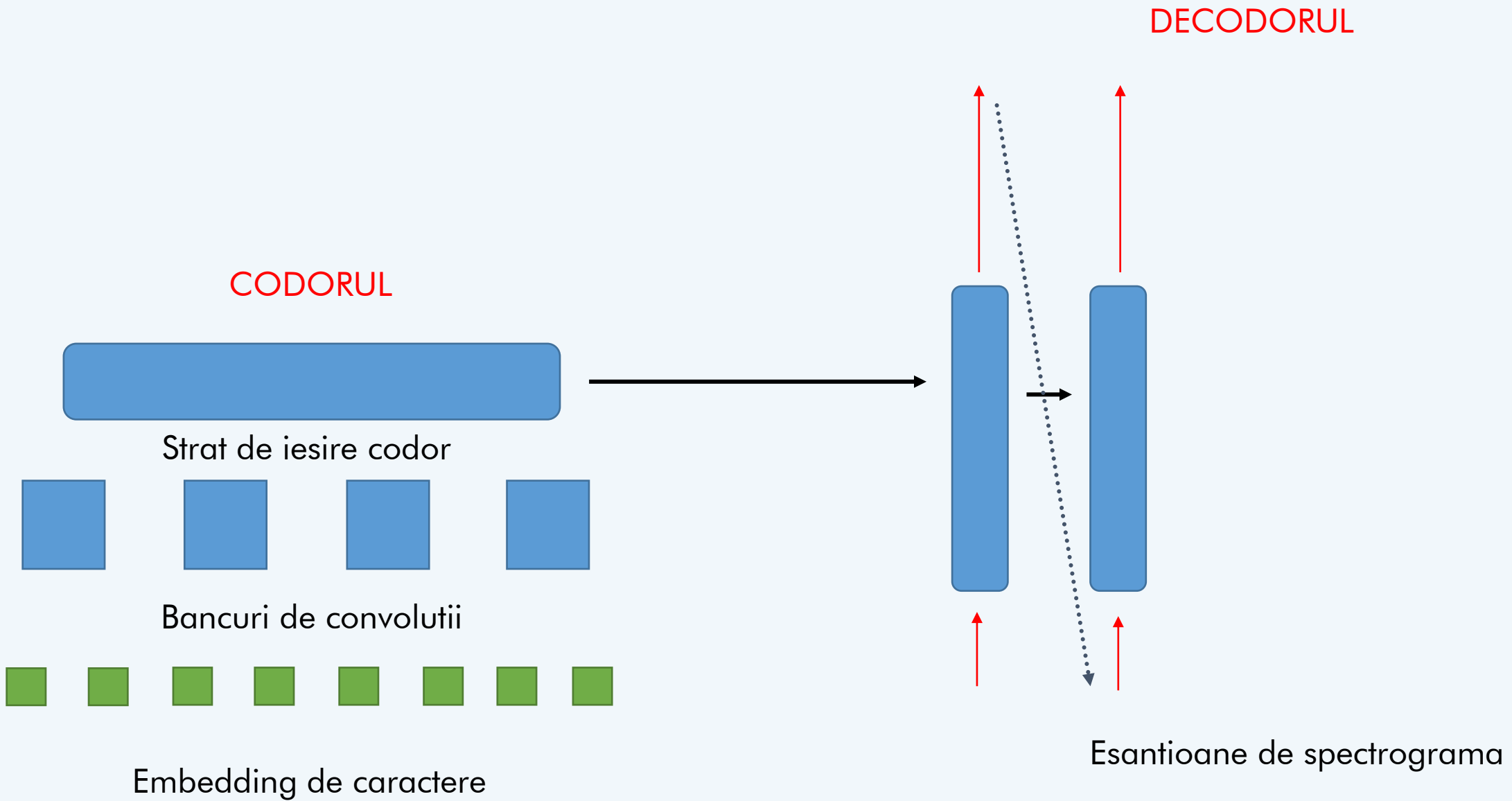


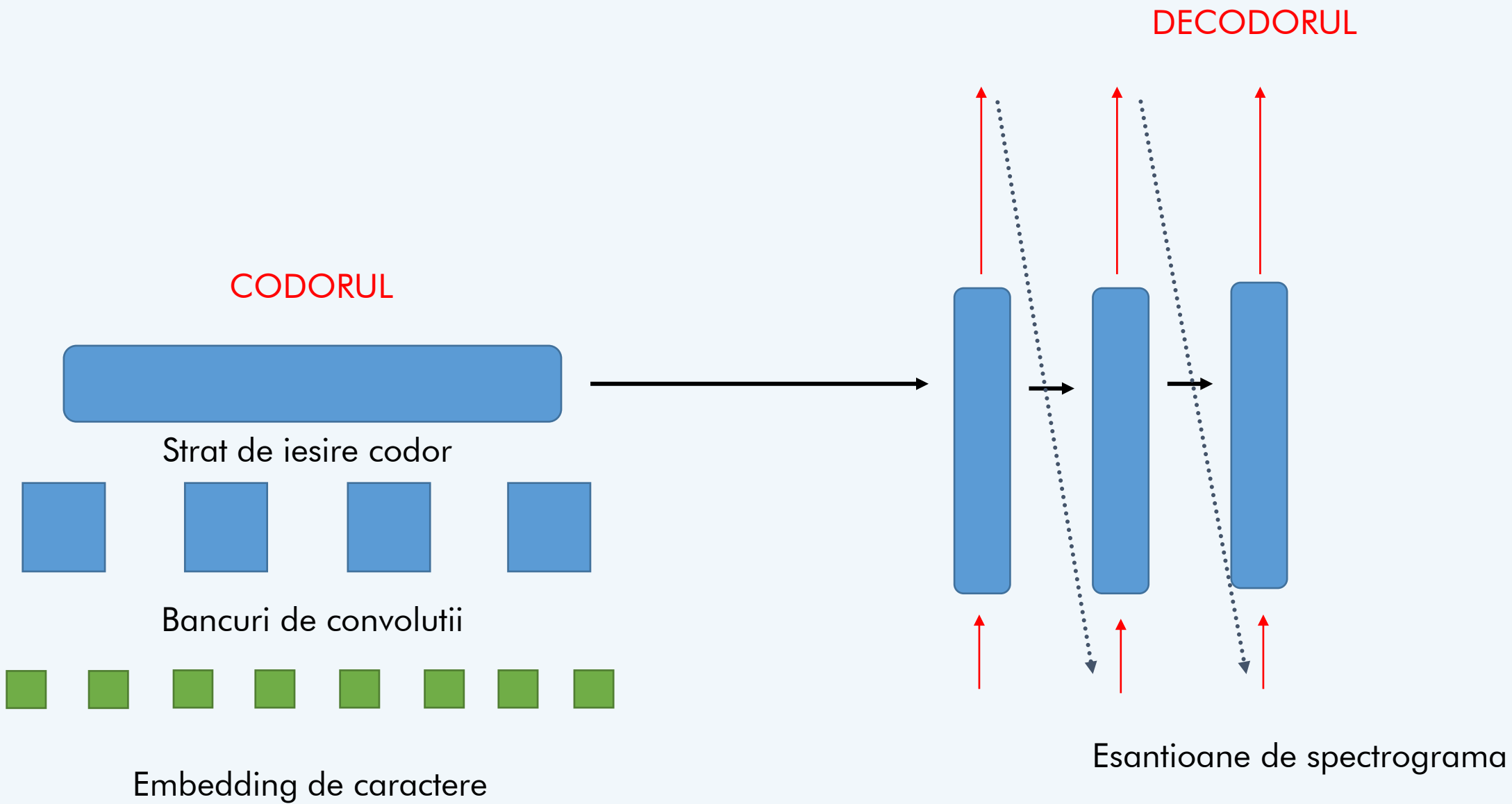
Obținem reprezentarea Mel

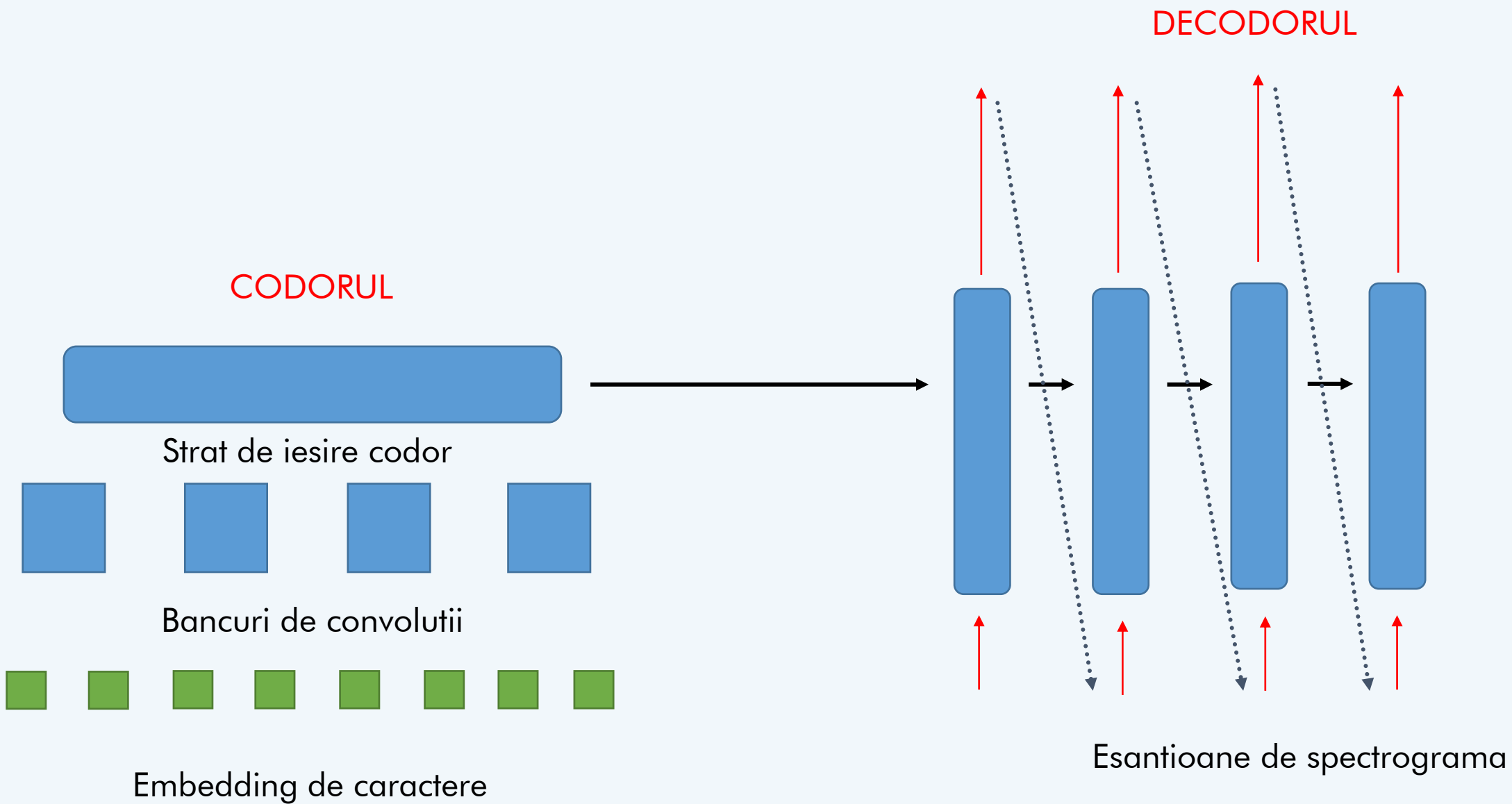


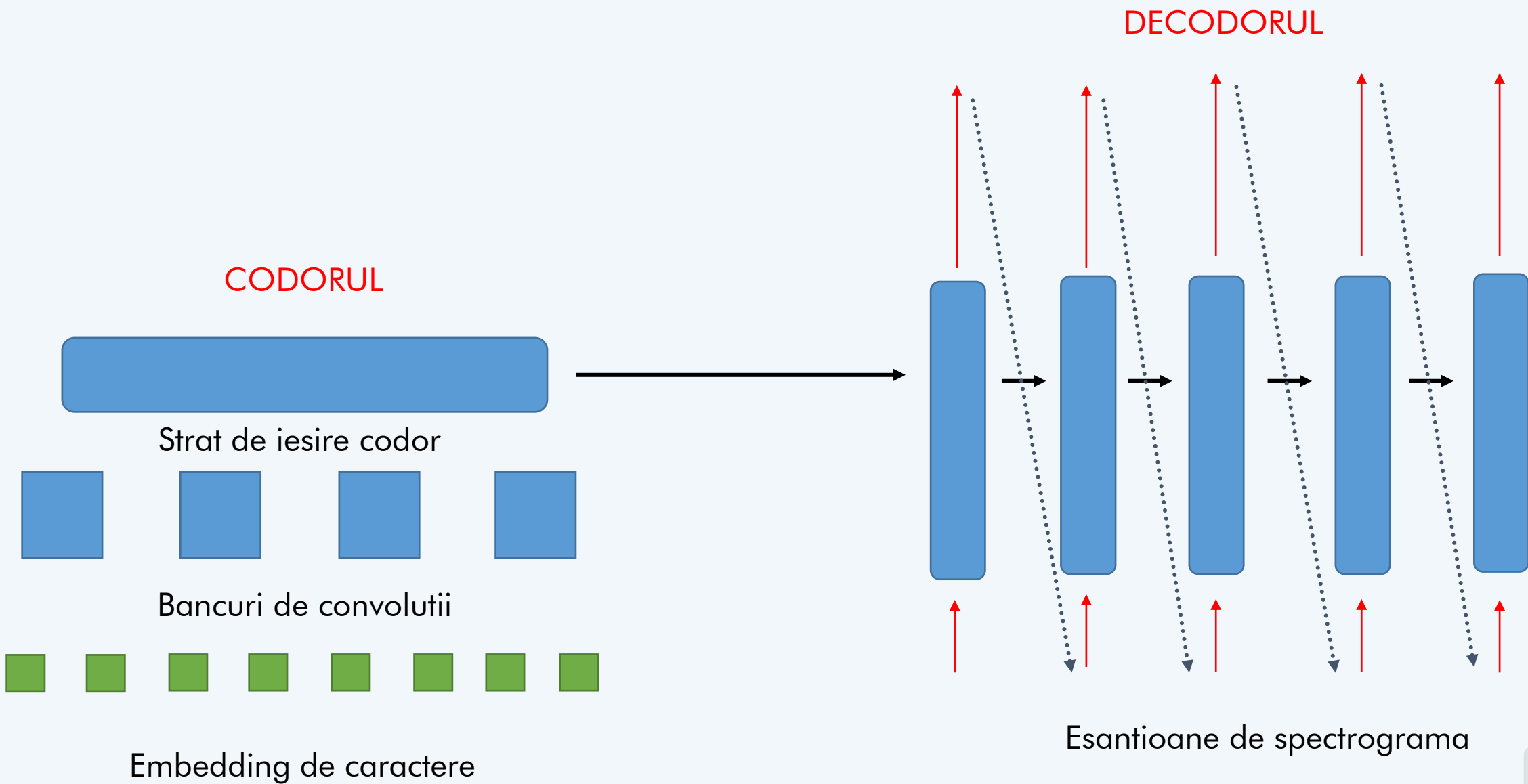
- 
- Text to Speech este o problemă secvență la secvență.
 - Secvența de intrare reprezintă propozițiile, iar secvența țintă reprezintă secvența de eşantioane.

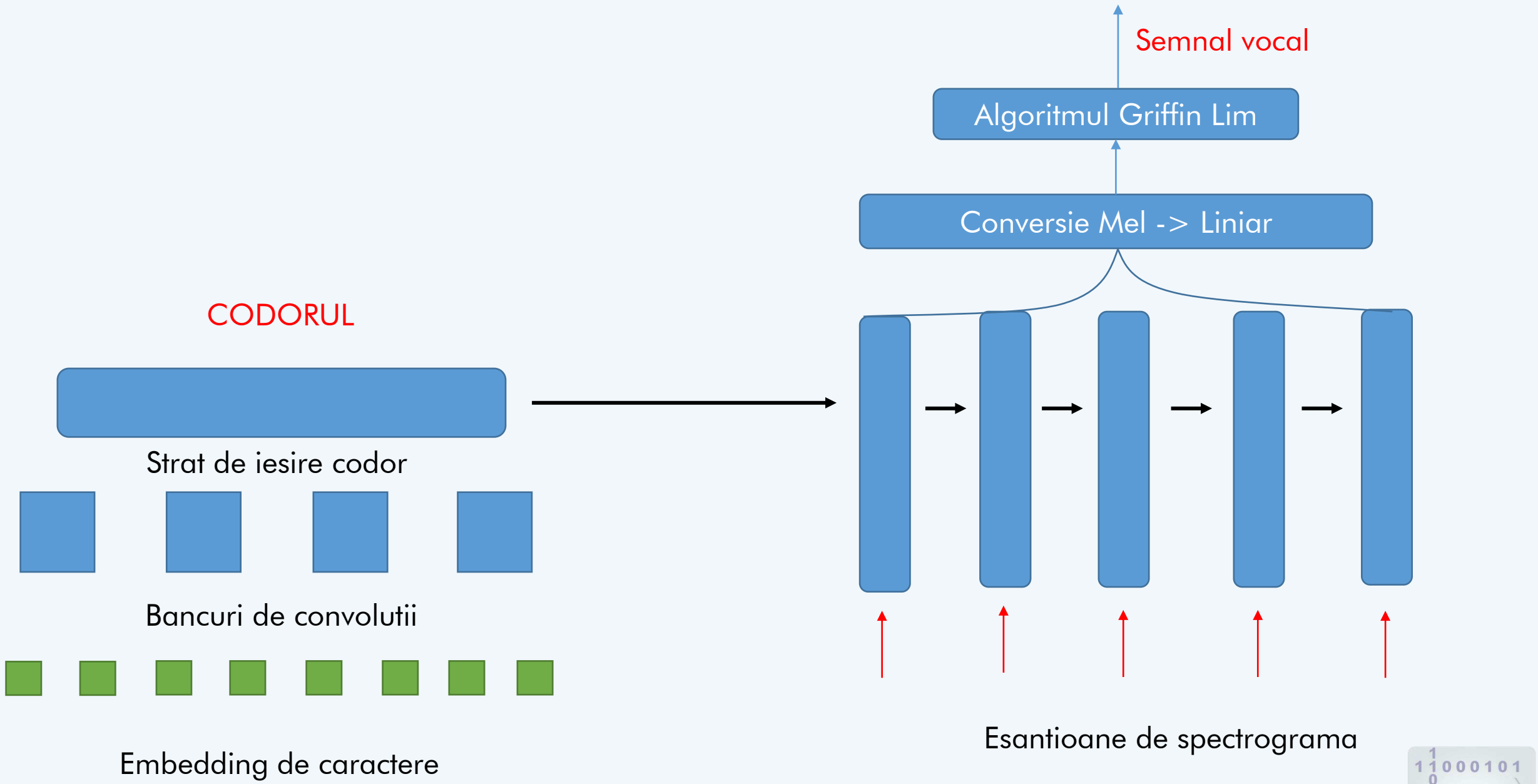




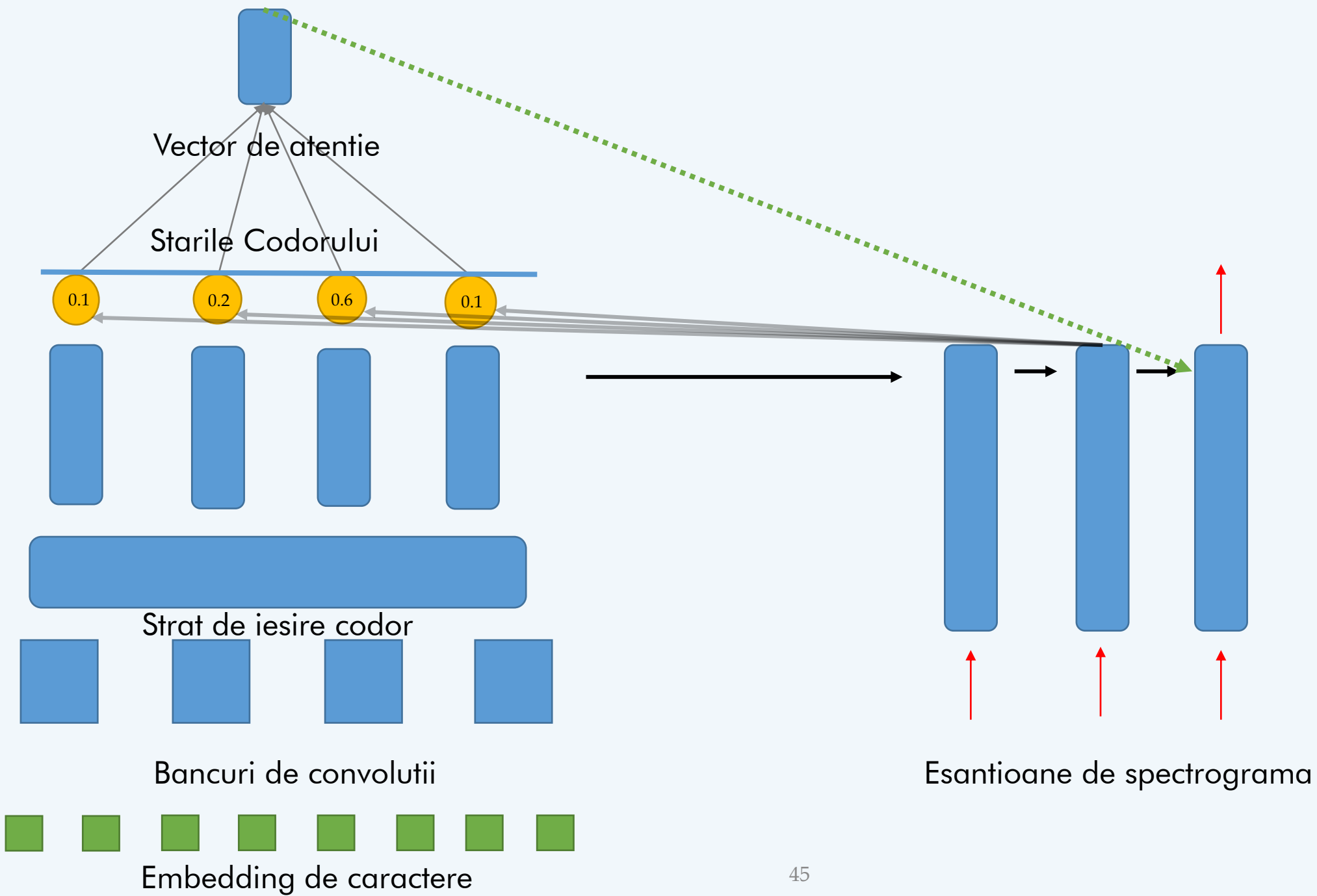








- Este nevoie sa asigurăm o aliniere corectă a eşantioanelor cu fragmentele de text
- Atenția este cea care ne va asigura acest lucru



ă

r

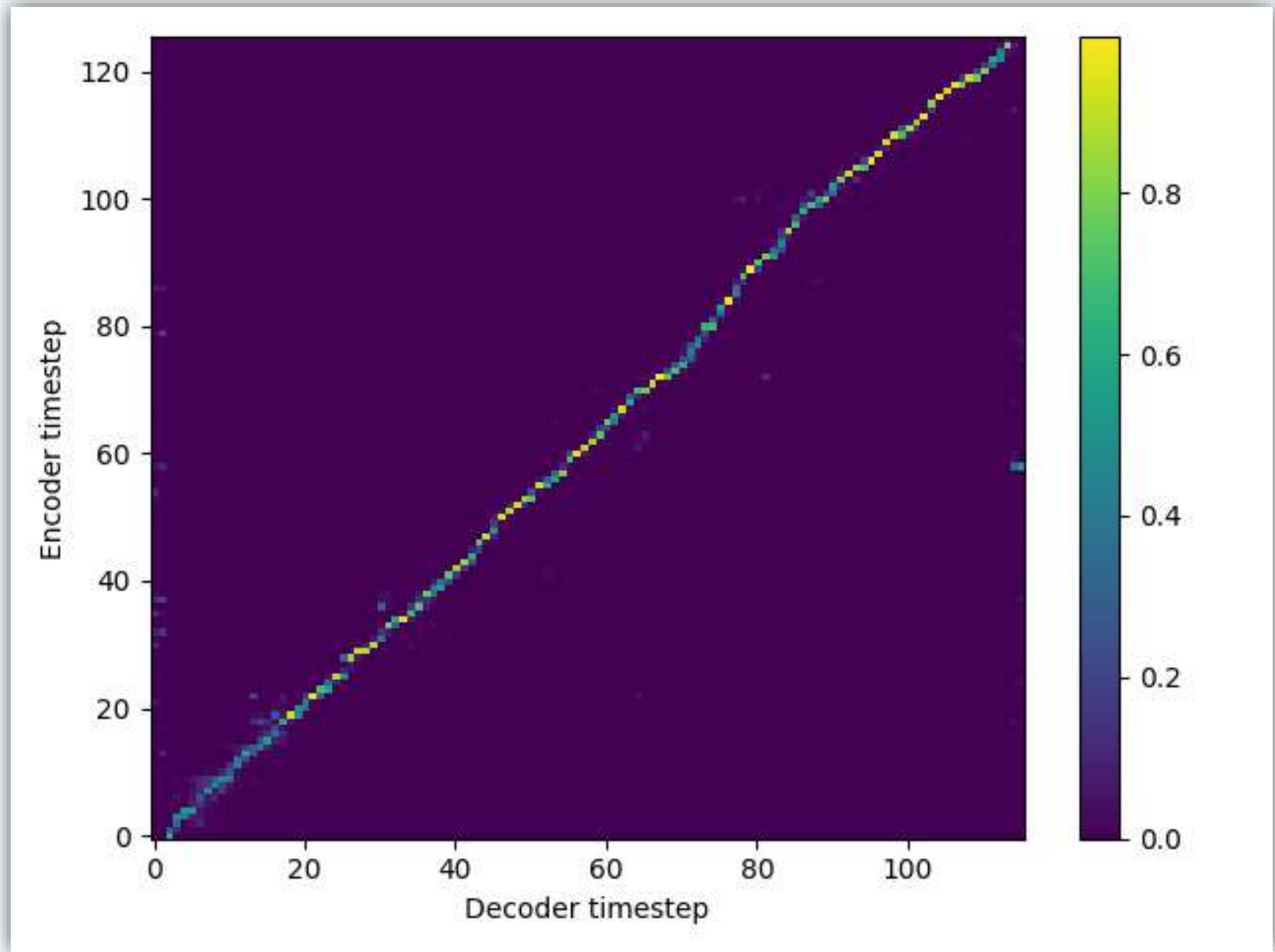
a

t

i

h

c



k'

i

t

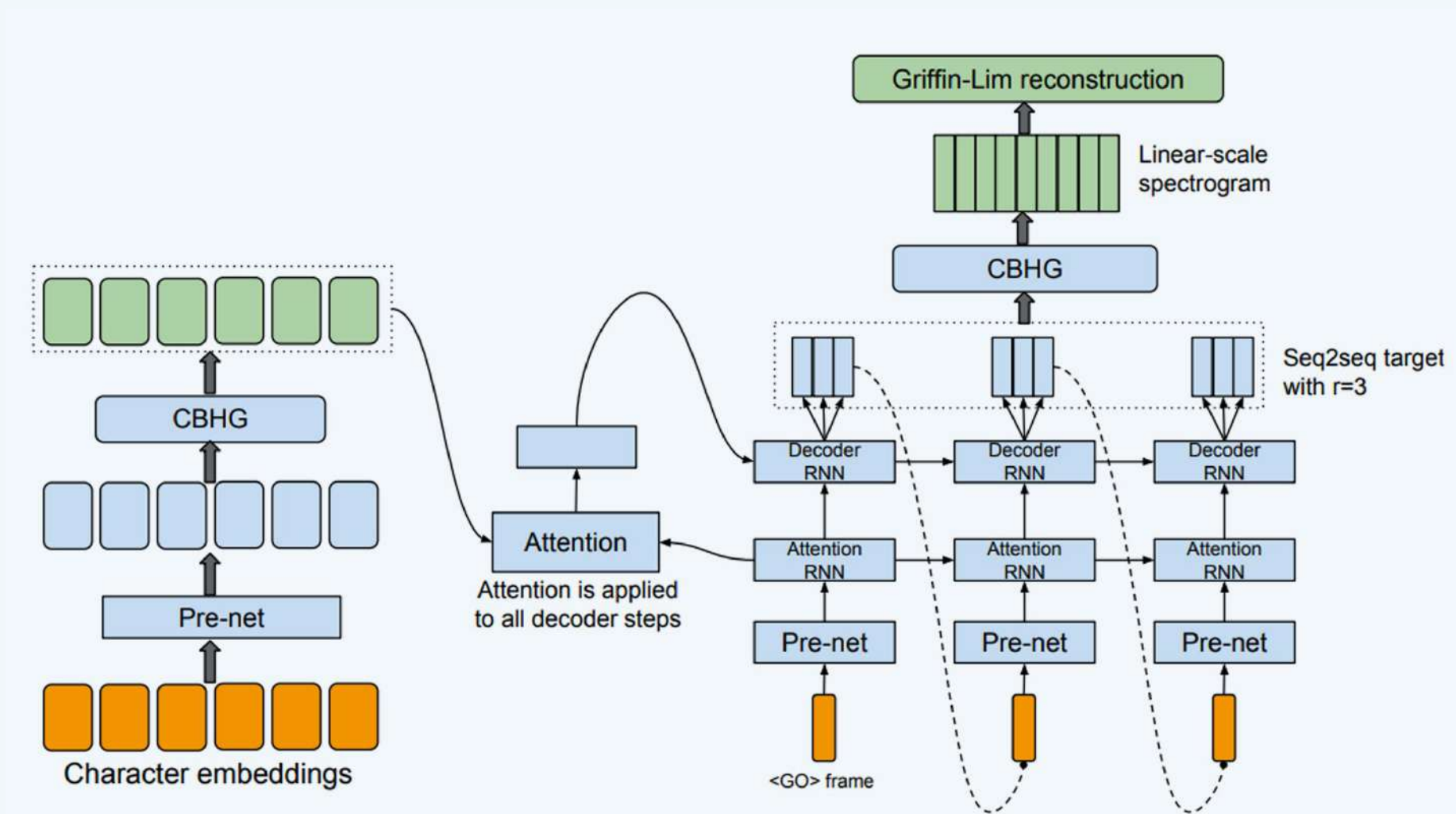
a

r

ə



Rețeaua Tacotron



Testarea bazelor de date

Baze de date

- Baza de date pentru limba engleză : LJSpeech ($f_s = 22050\text{Hz}$)
- Bazele de date pentru limba română :
 - RSC ($f_s = 16000\text{Hz}$)
 - RSS ($f_s = 48000\text{Hz}$)
 - baza de date proprie ($f_s = 44100\text{Hz}$)

Baza de date personală

- Corpusul a constatat din textul romanului Micul Print și articole de pe site-ul stiripentrucopii.ro

Baza de date personală

- Corpusul a constat din textul romanului Micul Prinț și articole de pe site-ul stiripentrucopii.ro
- Înregistrări între 2-20 secunde cu nivel dB similar

Baza de date personală

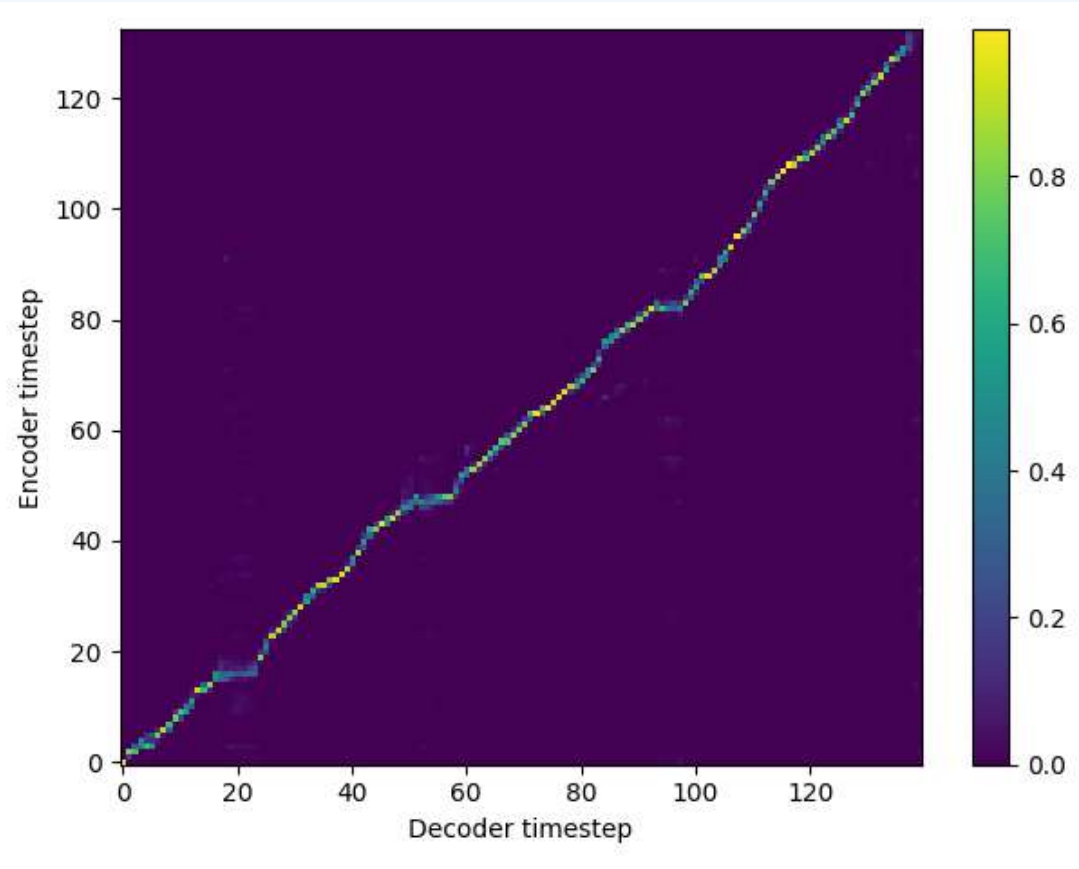
- Corpusul a constat din textul romanului Micul Print și articole de pe site-ul stiripentrucopii.ro
- Înregistrări între 2-20 secunde cu nivel dB similar
- Textele au fost normalizate

Evaluarea bazelor de date

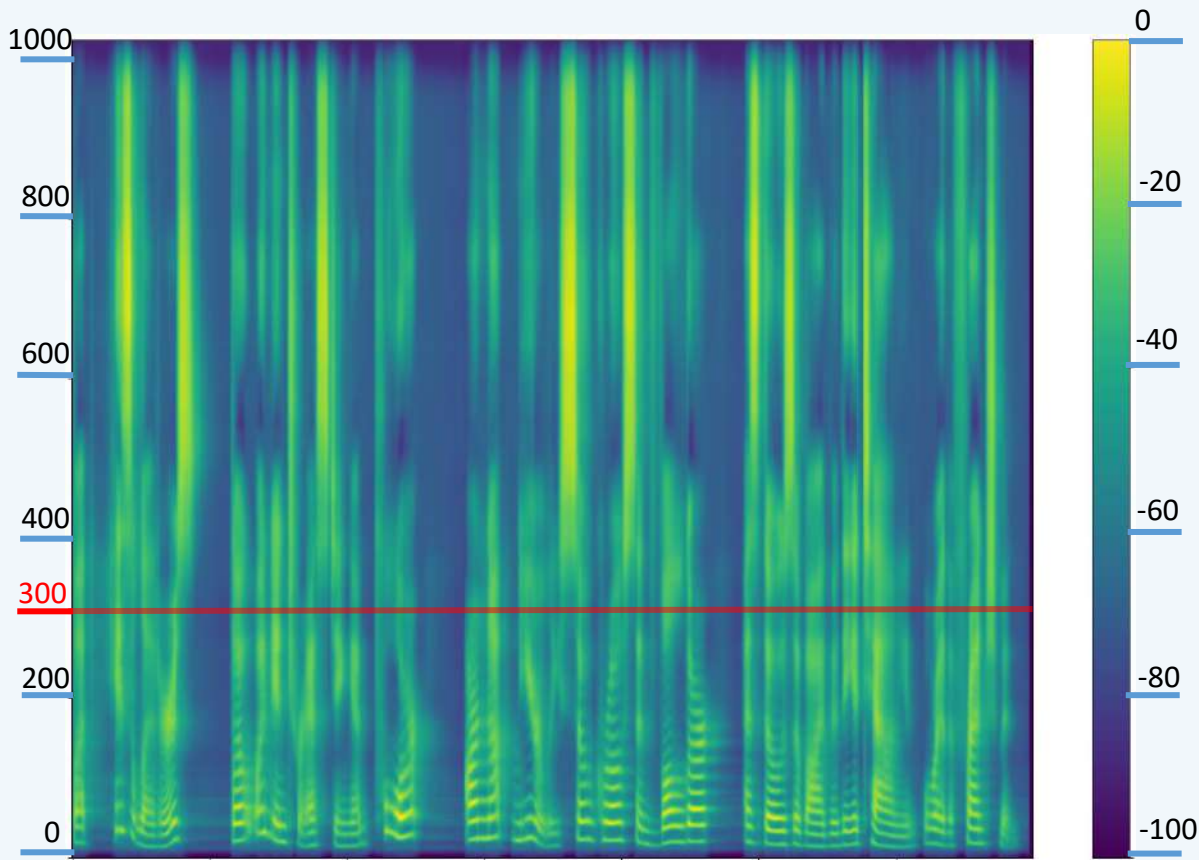
LJSPEECH



“but his letters, though uninteresting and poor, are not nearly so gross and vulgar as those of either the Italian or the Frenchman.”

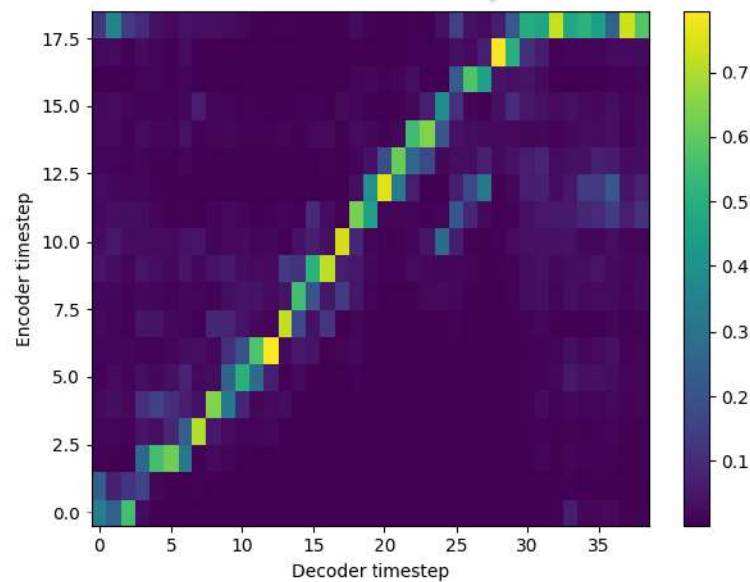


Cost	
Antrenare	Validare
0.0756	0.0856

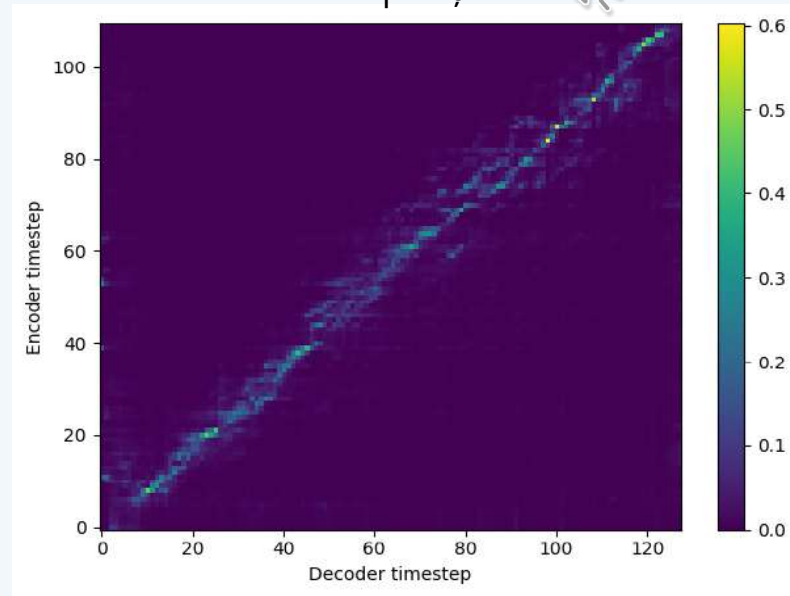


RSC

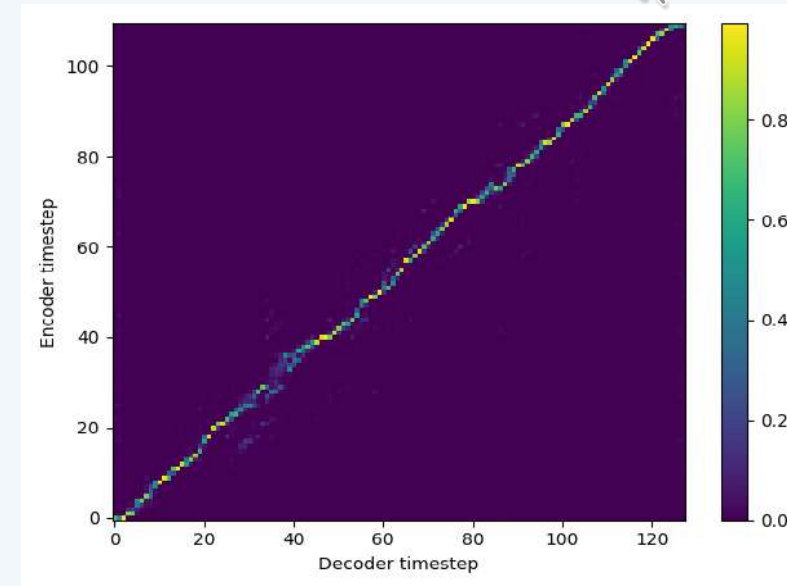
Cuvinte



Propoziții

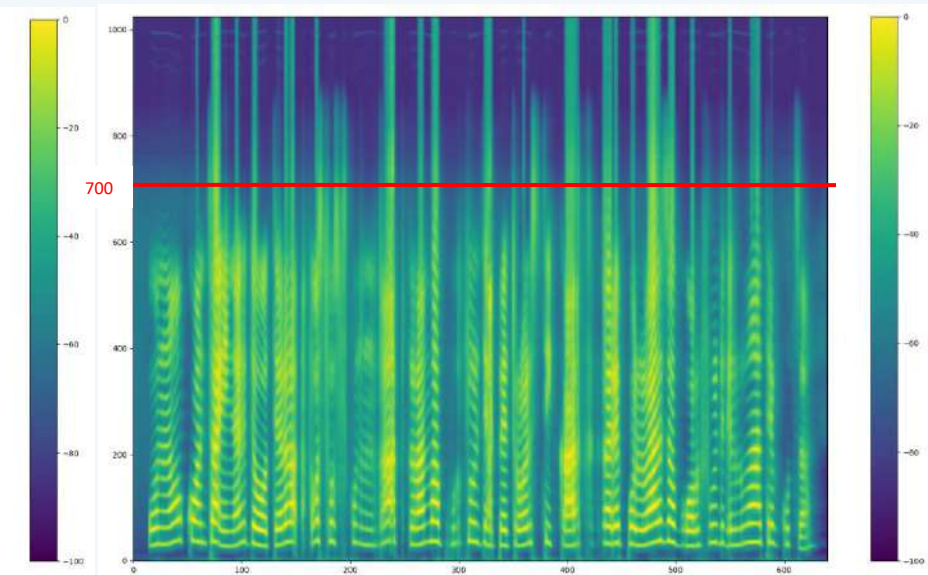
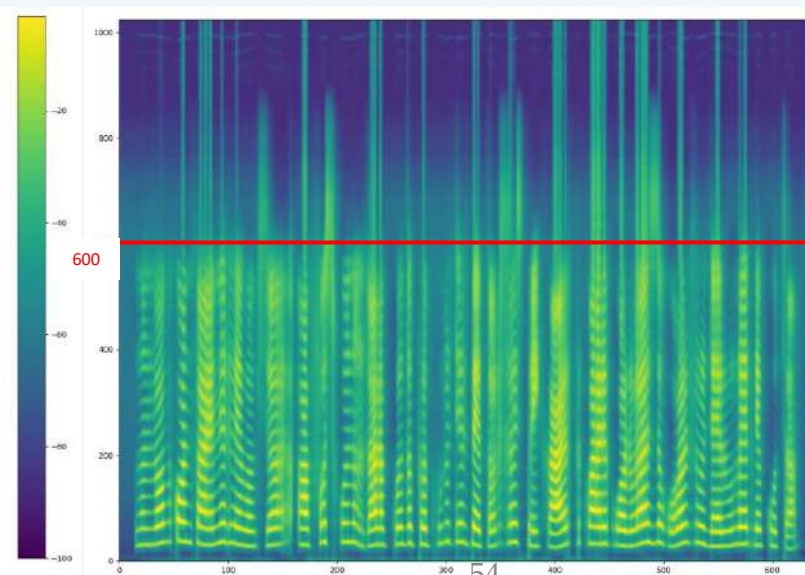
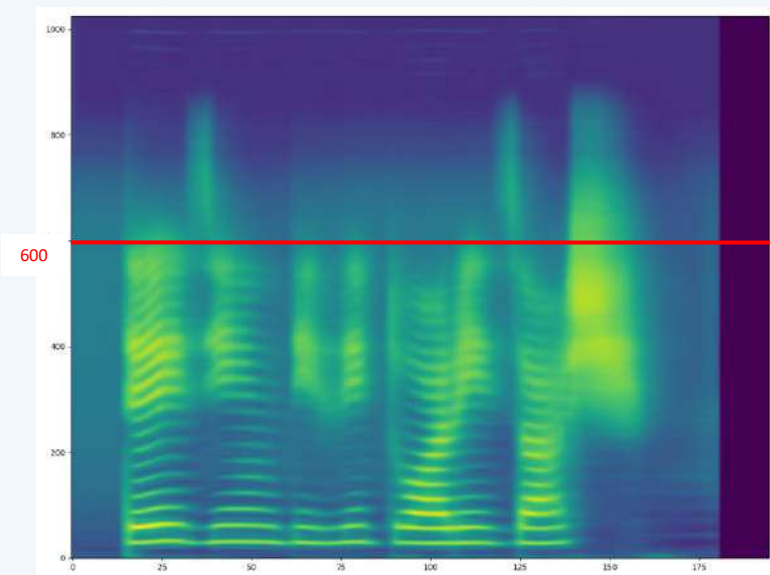


Cuvinte + propoziții



“dezindividualizași”

“un nou container de oțel a fost scufundat lângă puțul de petrol stricat care poluează golful de luna trecută”



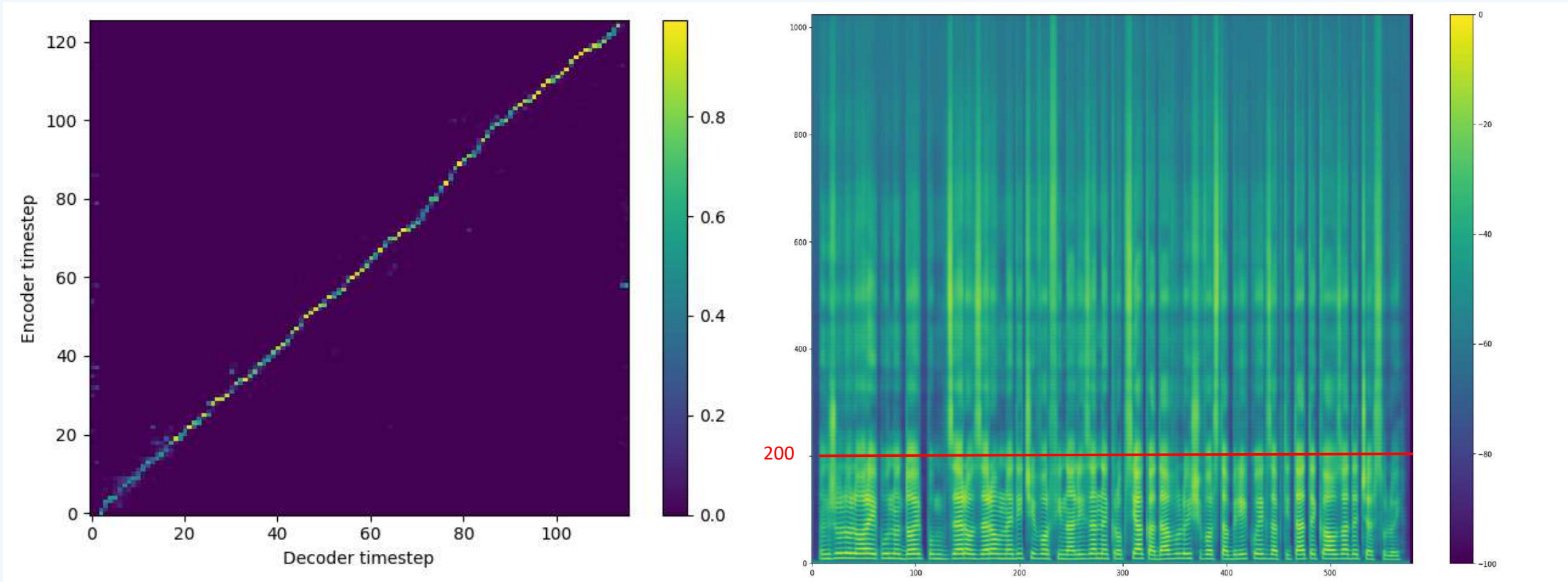
Cost		
	Antrenare	Validare
Cuvinte	0.0678	0.0755
Propozitii	0.0598	0.106
Cuvinte + Propozitii	0.0796	0.099

RSS

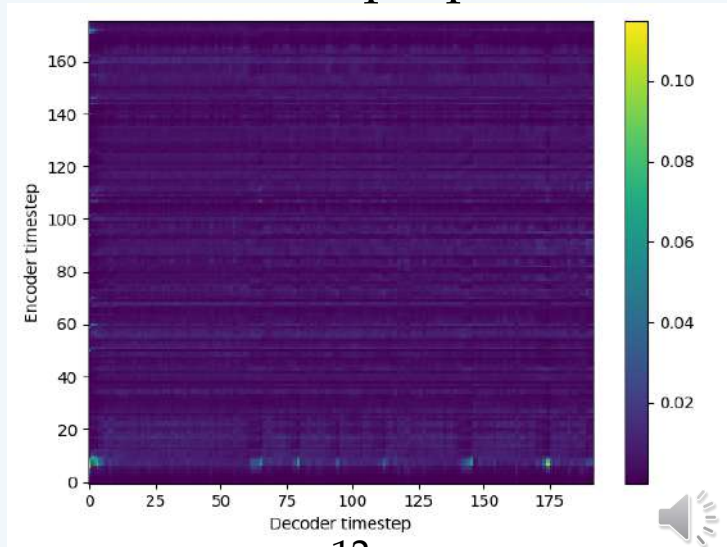


“În jurul orei unu doi punct zero zero medicii vor finaliza necropsia cadavrului și vor stabili cu exactitate cauza decesului.”

Cost	
Antrenare	Validare
0.0745	0.0818

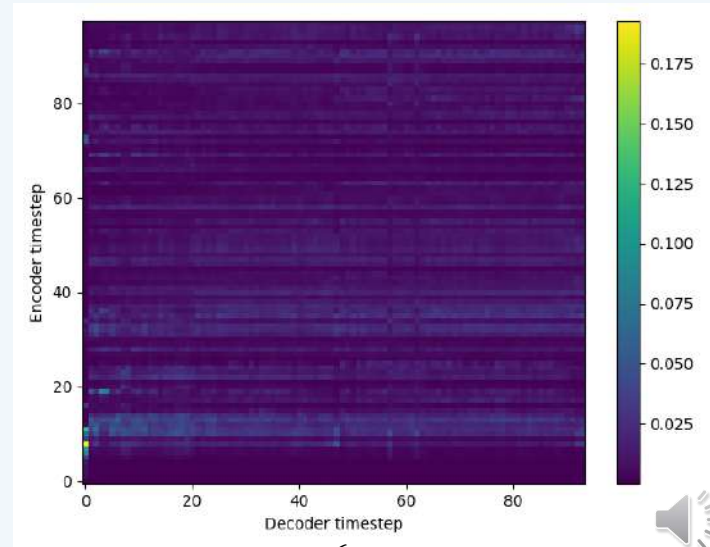


Baza de date proprie



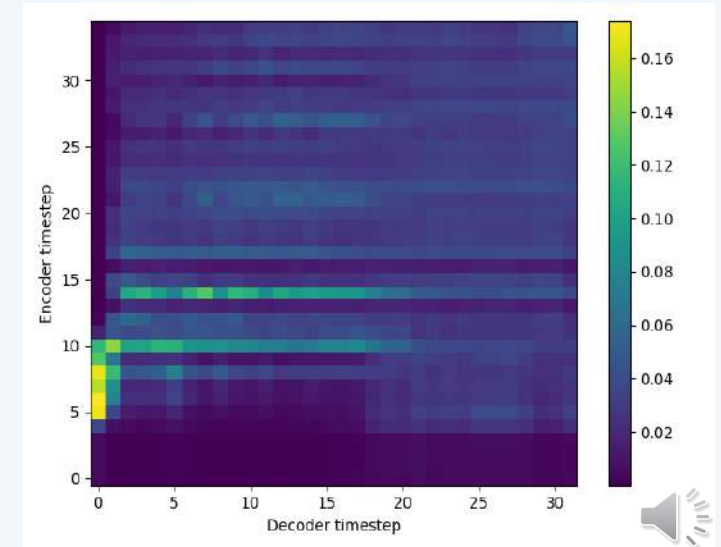
12s

*însă mi-am adus atunci aminte că studiasem
îndeosebi geografia , istoria , aritmetica și
gramatica și i-am spus omulețului acelaia că nu
mă pricep să desenez . el mi-a răspuns .*



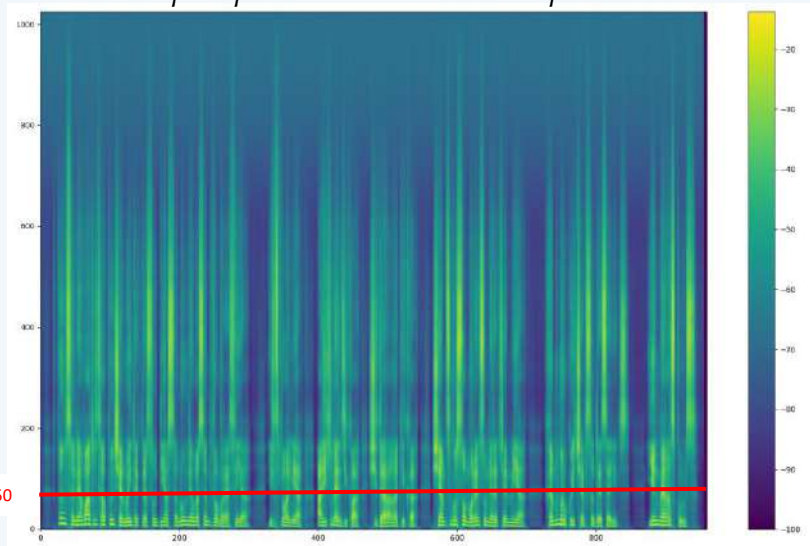
6s

*“am desenat atunci șarpele boa pe
dinăuntru , pentru ca astfel să poată pricepe
și oamenii mari .”*

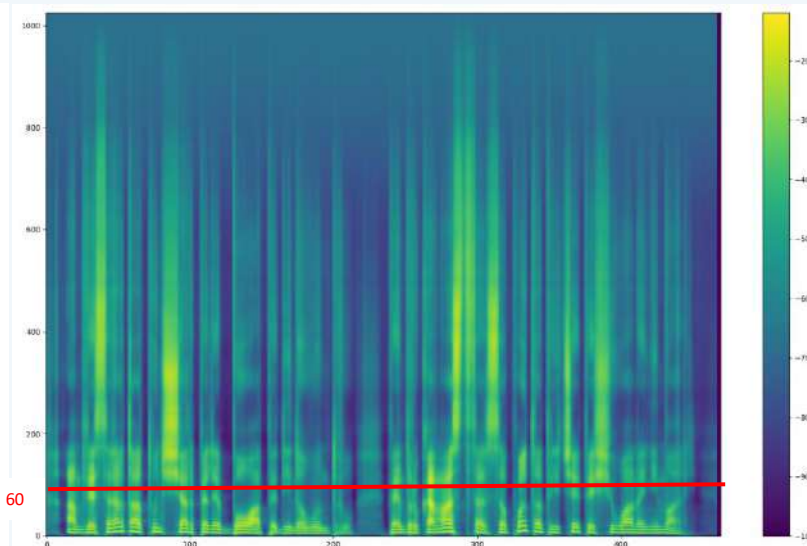


2s

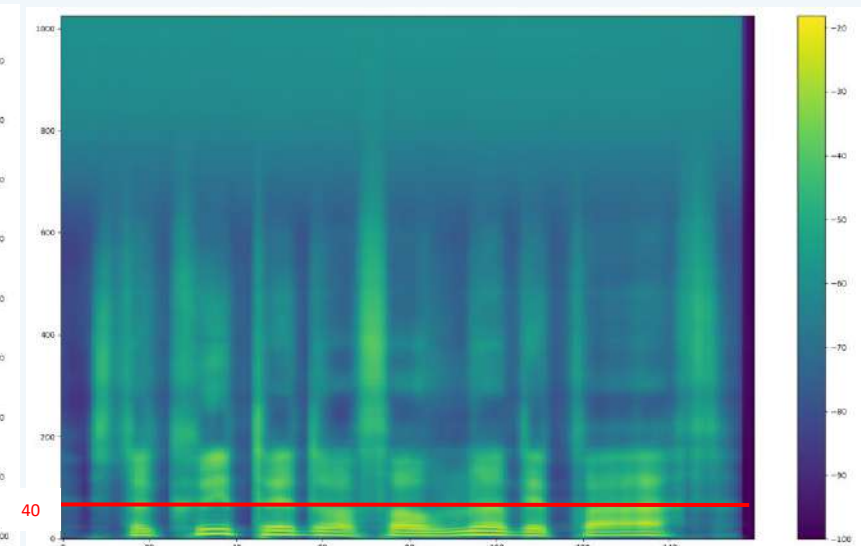
să cânte într-o formulă de taraf ,



50



60



40

Cost		
	Antrenare	Validare
12s	0.0519	0.0876
6s	0.0519	0.0804
2s	0.0519	0.1134

Modelul propus

Proiectarea propriei rețele

- Am divizat sarcina în două rețele separate

Proiectarea propriei rețele

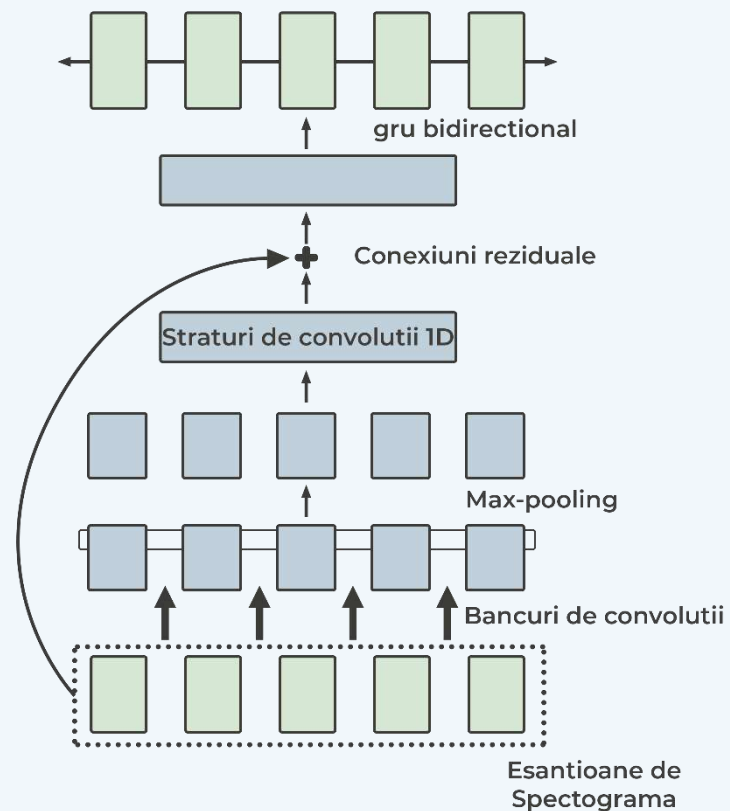
- Am divizat sarcina în două rețele separate
 1. conversia din spectrograma mel în spectrograma liniară

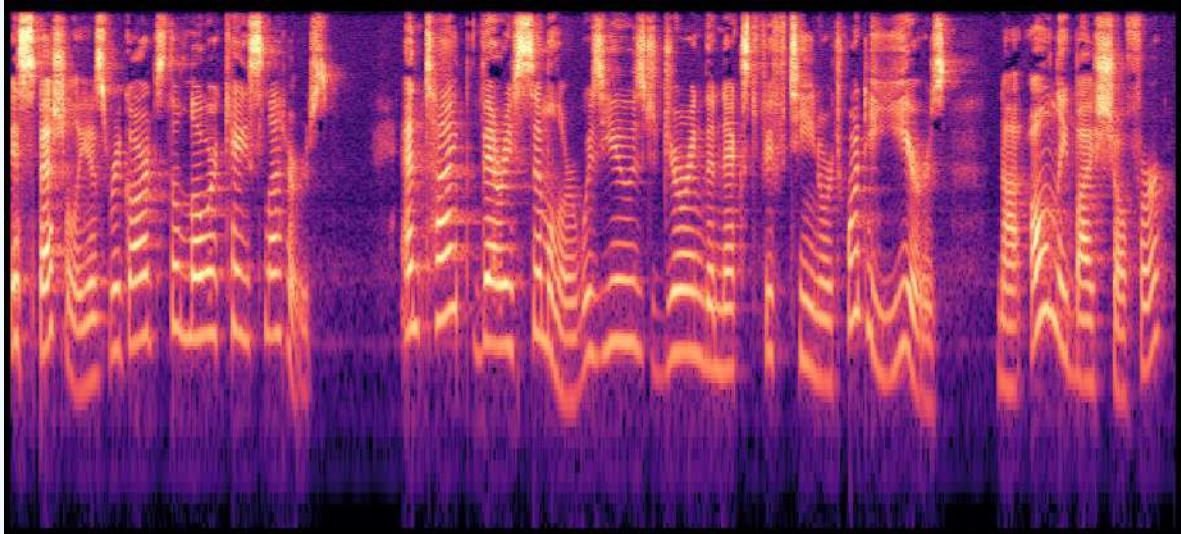
Proiectarea propriei rețele

- Am divizat sarcina în două rețele separate
 1. conversia din spectrograma mel în spectrograma liniară
 2. conversia textului în spectrograma mel

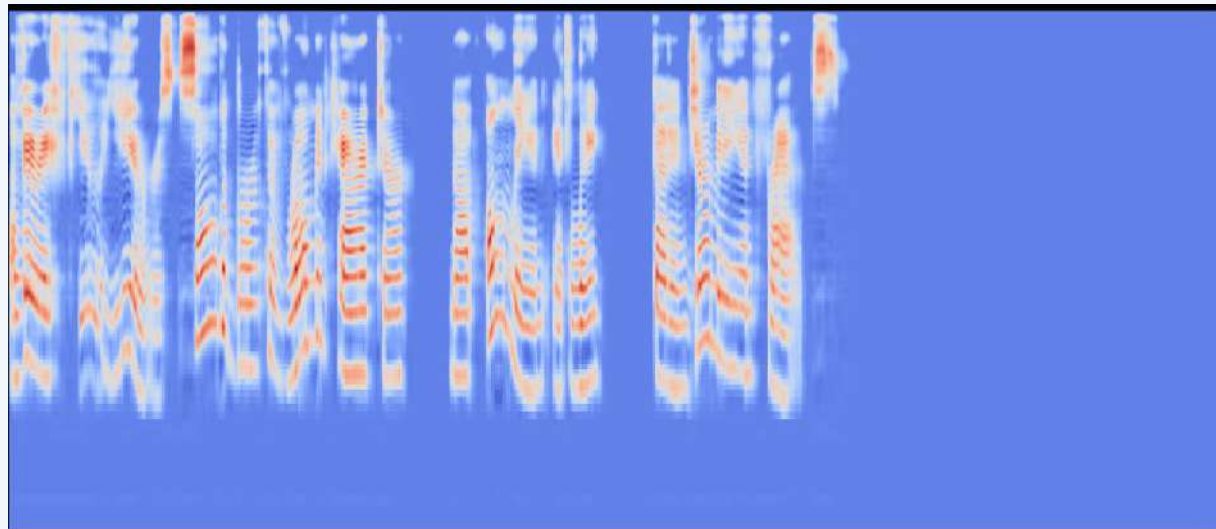
Analiză spectrală	pre-emphasis:0.97; frame_length: 50ms; frame_shift: 12.5ms; window_type: Hann
Character embedding	256
Codor (CBHG)	bancuri Conv1D: K=16, conv-k-128-ReLU MaxPooling: stride=1, width=2 Conv1D Proiecții:conv-3-128-ReLU → conv-3-128-ReLU Highway net:4, 128-ReLU GRU Bidirecțional:128
Codor preprocesare	Dense: 256-ReLU Dense: 128-ReLU
Decodor preprocesare	Dense:256-ReLU Dense:128-ReLU
Decodor RNN	2 GRU : 256 units
Atenție RNN	1 GRU Bidirecțional : 128
Post Procesare CBHG	bancuri Conv1D: K=16, conv-1-128-ReLU MaxPooling: stride=1, width=2 Conv1D Proiecții:conv-3-128-ReLU → conv-3-128-ReLU Highway net:4, 128-ReLU GRU Bidirecțional:128
Factor de reducere (r)	5

Conversie spectrograme

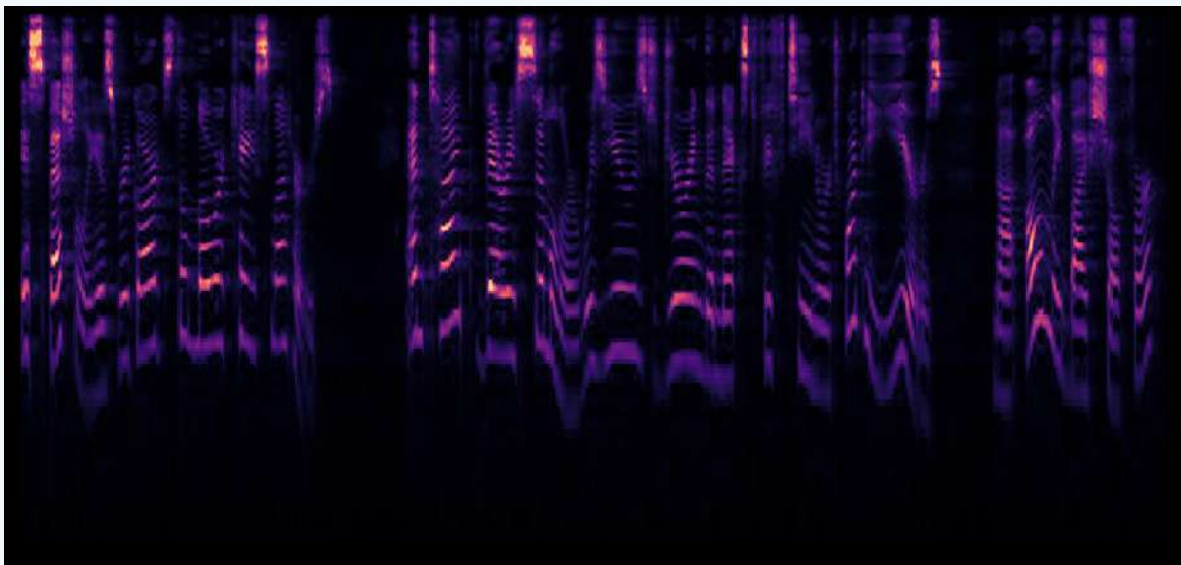




Engleză original



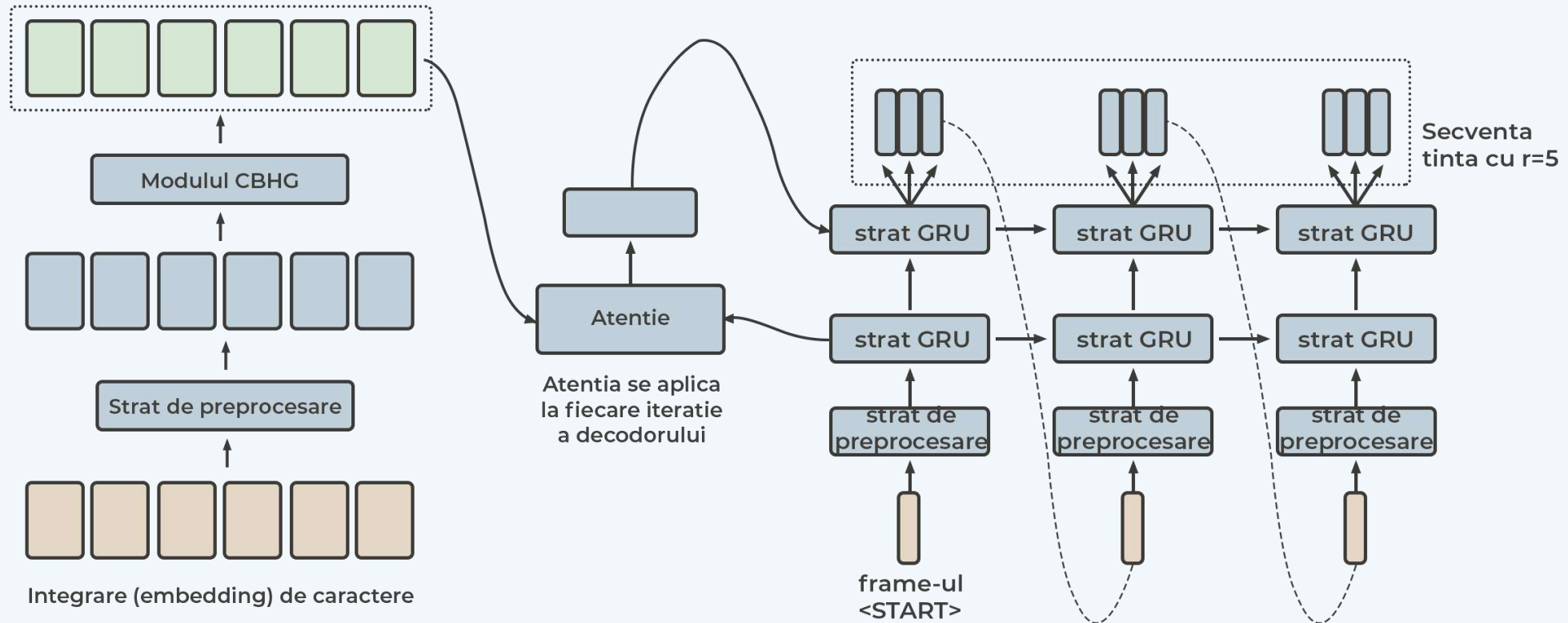
Engleză rețea

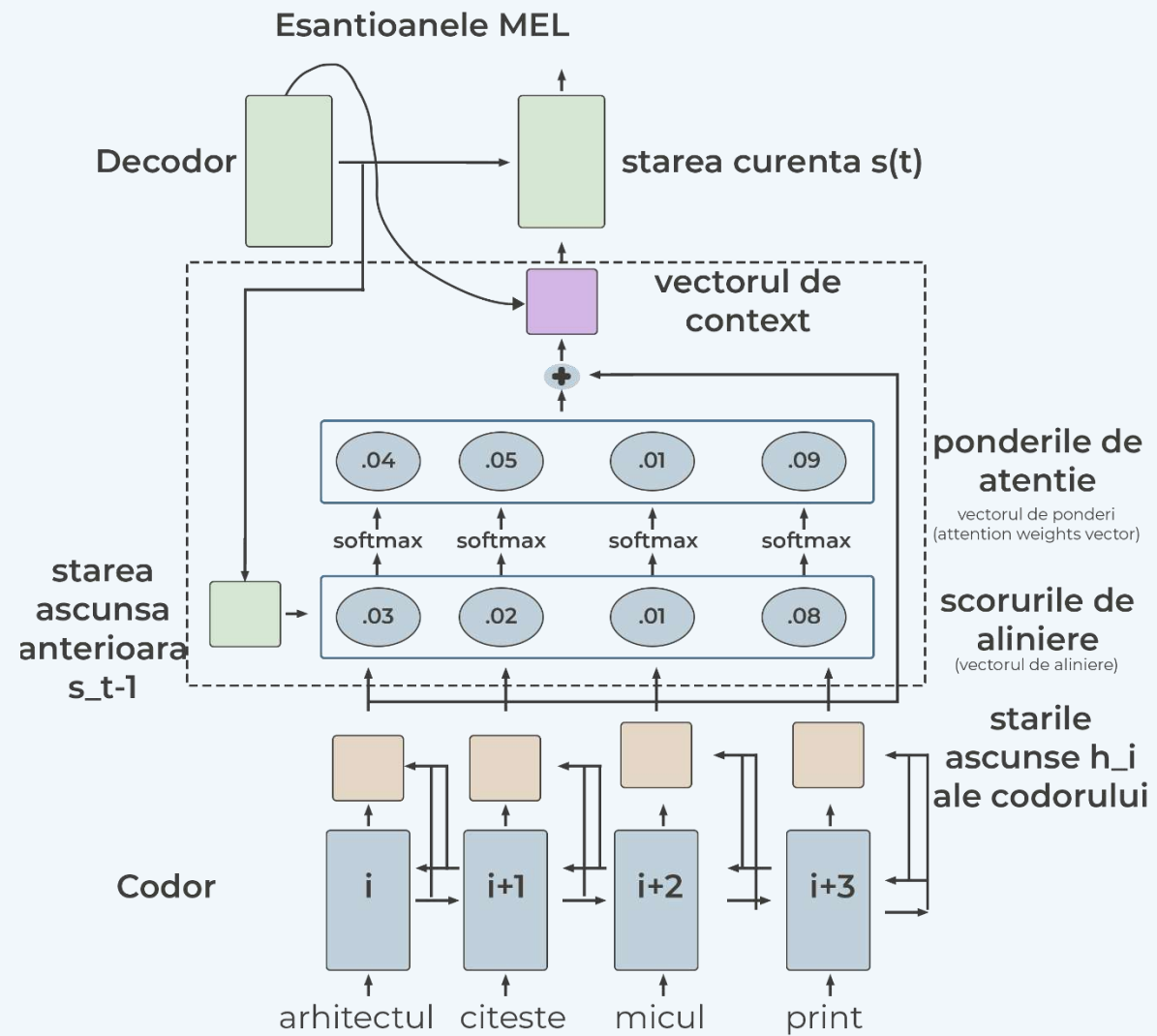


Engleză determinist

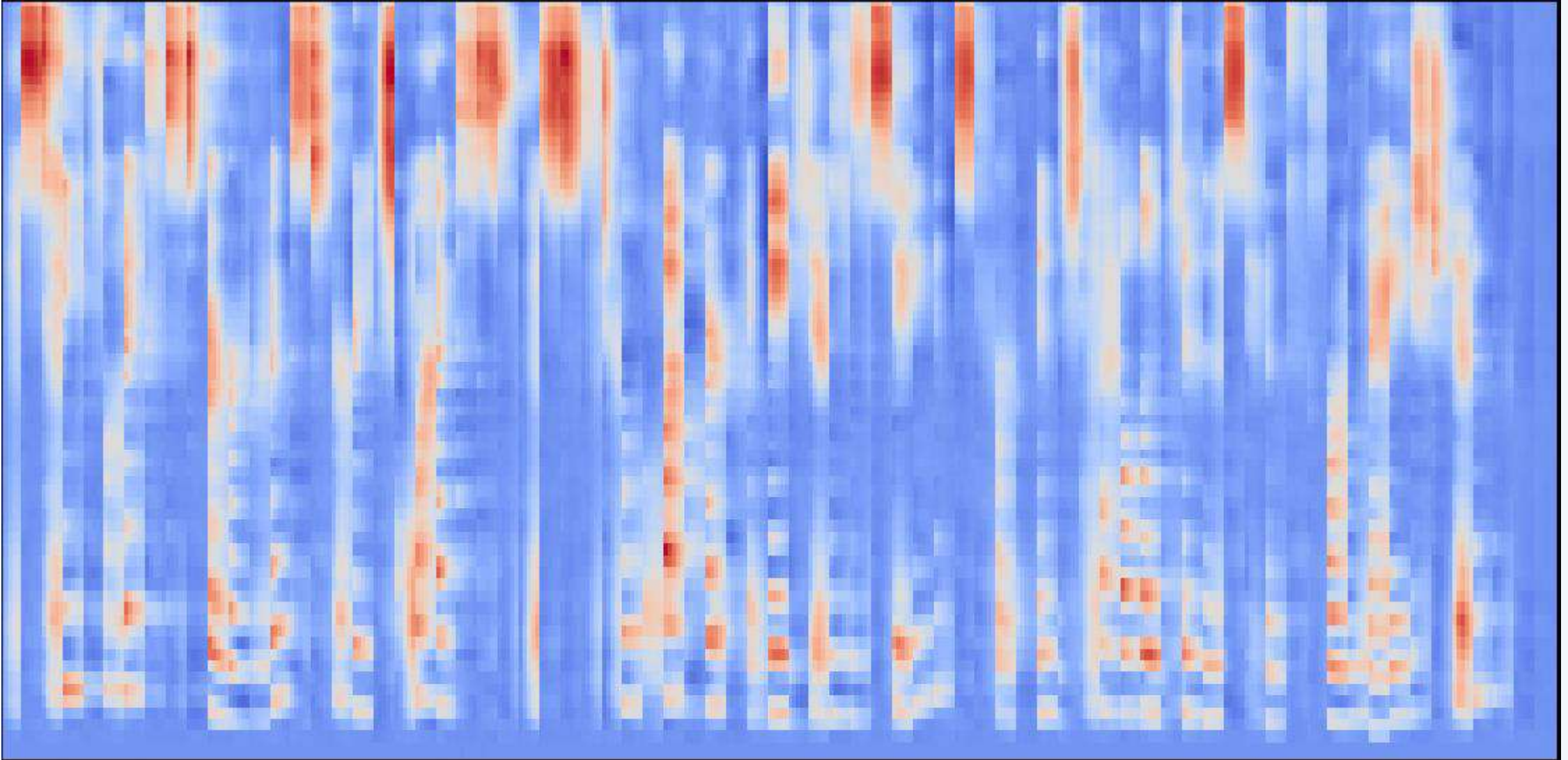
Cost		
Limba	Retea	Determinist
Engleza	0.0989	0.123
Romana	0.0751	0.0951

Conversia textului în spectrogramă MEL





Rezultate



Contribuții personale

- Realizarea unei baze de date proprii
- Prelucrarea și pregătirea bazei de date pentru antrenare, folosind multiple scripturi în Python
- Evaluarea și compararea mai mult baze de date
- Dezvoltarea modulelor pentru rețeaua de transcriere a textului în vorbire:
 - Codorul
 - Atenția
 - Decodorul



Vă mulțumesc !