Facultatea de Electronică,
Telecomunicații și
Tehnologia Informației
www.etti.utcluj.ro

Sistem de Recunoaștere Vocală Automată pentru Limba Română utilizând Rețele Neuronale DNN

Automatic Speech Recognition System for Romanian Language using Deep Neural Networks

ASRSV 15.12.2020
Andreea-Georgiana SARCA

UNIVERSITATEA TEHNICĂ DIN CLUJ-NAPOCA

Cuprins

1. Motivația
2. Stadiul Actual in ASR
3. Modele Clasice vs. Trendul Actual
4. Arhitectura Rețelelor Neuronale în DeepSpeech
5. Implementare: Adaptarea pentru Limba Română
6. Rezultate experimentale
7. Concluzii

1. Motivația

- Necesitatea unui sistem de recunoaștere vocală automată în limba română pentru a fi integrat în aplicații de securitate și aplicații de tip interacțiune om-mașină
- Există modelele vechi care utilizează Modelele Markov clasice, dar trendul este de a utiliza Deep Learning
- Sunt puține studii în acest domeniu, unul dintre ele fiind realizat de către Universitatea Politehnică din București
- Cererea mare a pieței pentru a avea un sistem de recunoștere vocal propriu, personalizat în conformitate cu necesitățile utilizatorilor

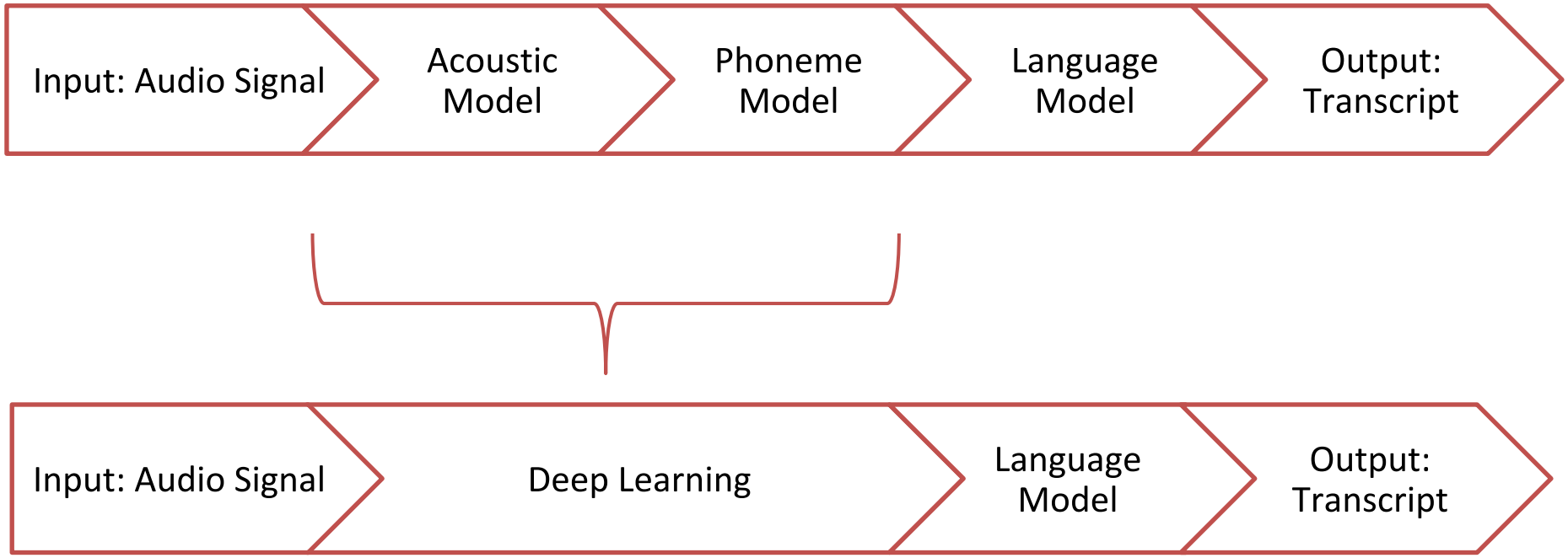
2. Stadiul actual în ASR

- Există sisteme comerciale inclusiv pentru limba română care îndeplinesc funcția de Speech to Text
- Asistenții digitali: Siri, Google Assistant, Amazon Alexa

De ce ASR pentru limba română?

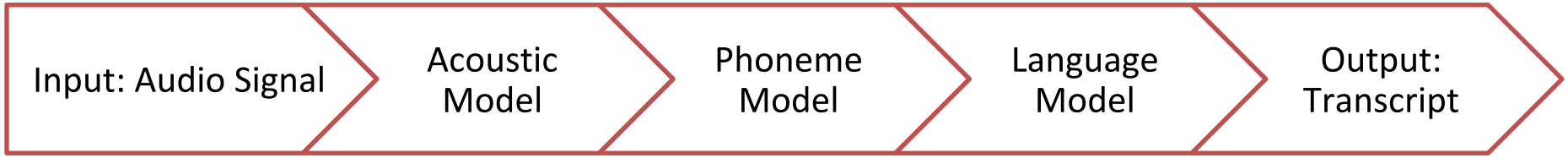
3. Modelul Clasic vs. Trendul actual

Classic ASR



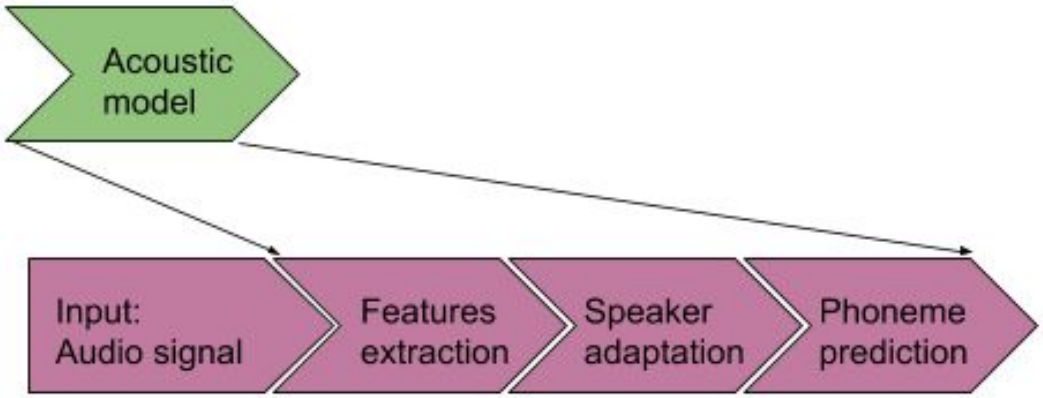
DeepSpeech

Modelul Clasic



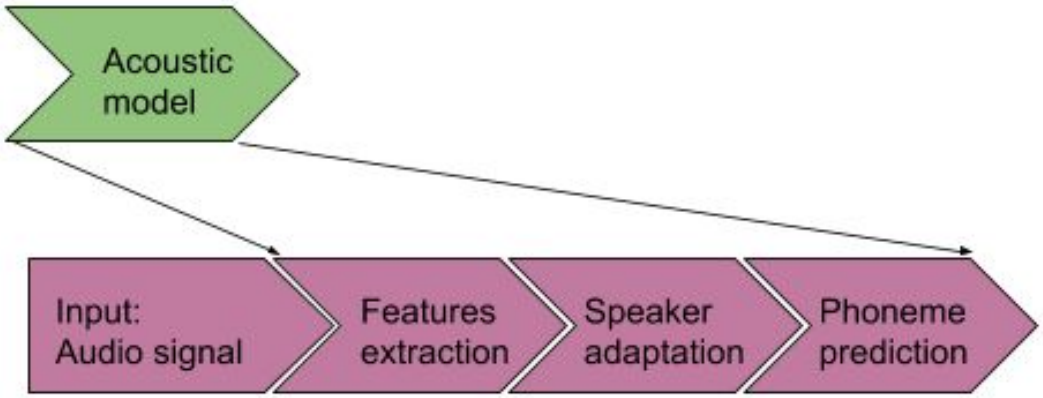
- **Modelul acustic** este cel mai important bloc dintr-un sistem de recunoaștere vocală. Acesta este responsabil pentru **recunoașterea fonemelor** care sunt procesate mai apoi de către **Modelul fonetic.**
- **Modelul fonetic** procesează fonemele precedent determinate și formează **cuvinte**.
- **Modelul de limbă** este necesar pentru a asigura **contextul și coexistența cuvintelor** pentru a forma **transcrierea finală**.

Modelul Clasic - Modelul Acustic

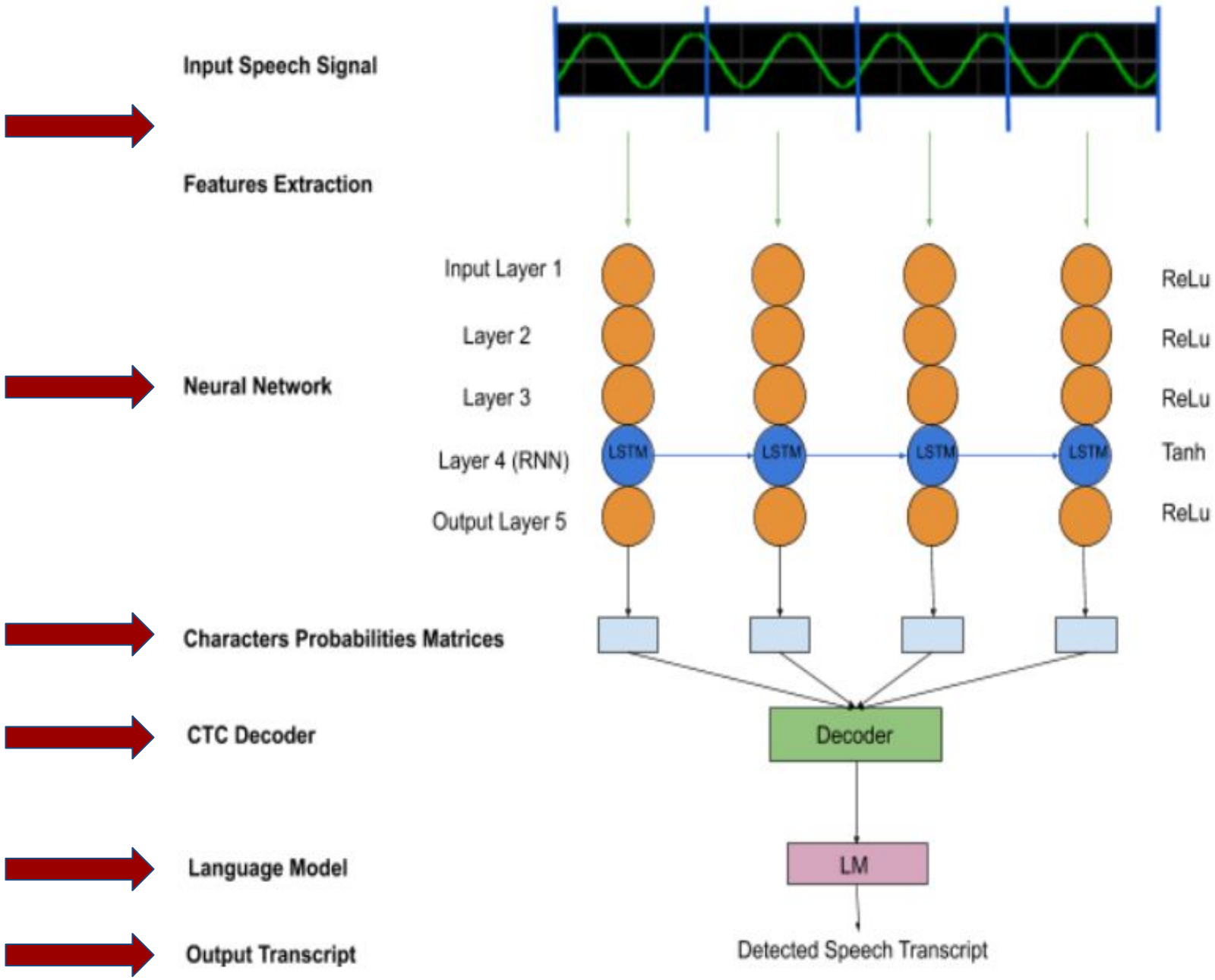


- Modelul Acustic este cheia oricărui sistem de recunoaștere vocală.
- Acesta este compus din un **bloc de extragere a caracteristicilor vorbirii**, un **bloc de adaptare a vorbitorului** și un **bloc de predicție fonetică**.
- **Extragerea caracteristicilor vorbirii** presupune transformarea în domeniul frecvență a semnalului audio.
- Deoarece **audibilitatea umană este neliniară**, aceasta este reprezentată pe **Scara Mel**.
- Din această reprezentare neliniară se extrage Spectrograma semnalului și 39 de Coeficienți Mel Cepstrali, cu ajutorul cărora se identifică cuvintele monosilabice.

Modelul Clasic - Modelul Acustic

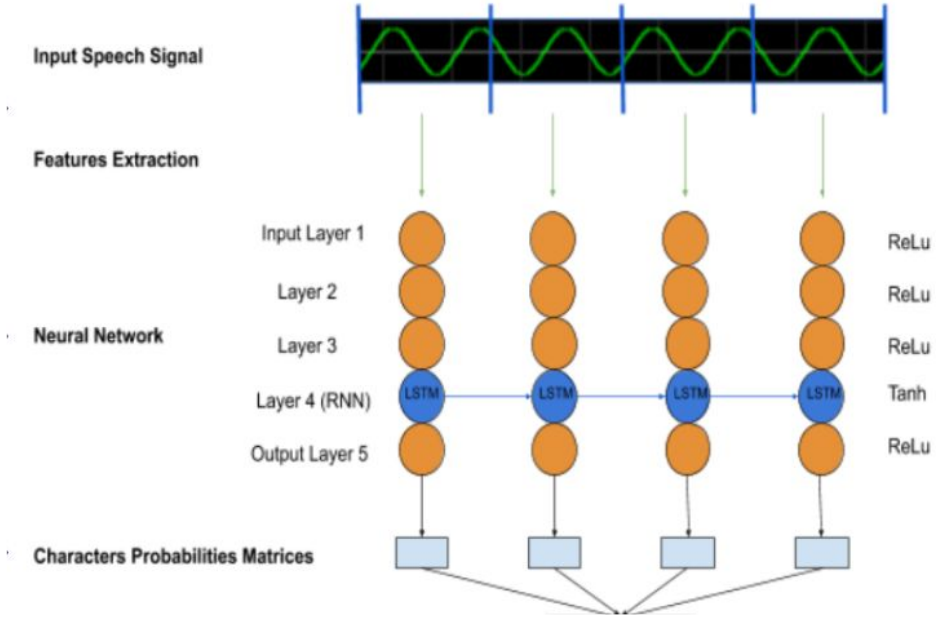


- **Adaptarea vorbitorului** presupune **normalizarea datelor** obținute în blocul anterior deoarece **variabilitatea** specifică fiecărui vorbitor prezintă o **sursă importantă de zgomot și erori.**
- Procedura prin care se realizează această **normalizare** utilizează **metode statistice avansate de uniformizare** a datelor. Totuși, chiar și după acest proces, **amprenta personală și accentul vorbitorului** pot fi încă parțial sesizate.
- **Predictorul de foneme** este deseori construit utilizând Deep Neural Networks. Folosind informațiile extrase din blocurile anterioare, în această etapă **se prezic fonemele din vorbire**. Din punct de vedere **computațional predicția de foneme necesită cea mai mare complexitate**, astfel **modelul acustic fiind cel mai costisitor** dintre toate blocurile principale implicate în recunoaștere.


Input Speech Signal
Features Extraction
Neural Network
Input Layer 1
Layer 2
Layer 3
Layer 4 (RNN)
Output Layer 5
LSTM
LSTM
LSTM
LSTM
ReLu
ReLu
ReLu
Tanh
ReLu
Characters Probabilities Matrices
CTC Decoder
Decoder
Language Model
LM
Output Transcript
Detected Speech Transcript

Modelul DeepSpeech

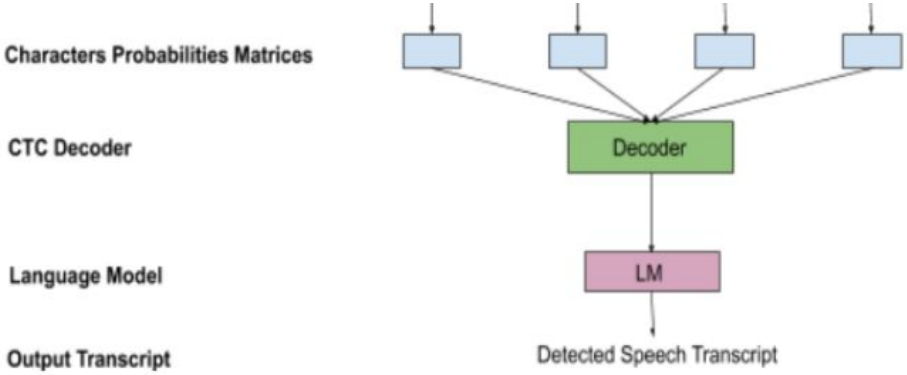
- Sistemul este construit pentru a **recunoaște vorbirea la nivel de caracter**, utilizând **5 straturi de rețele neuronale**.
- **Caracteristicile extrase** din Spectrograma semnalul audio sunt **Coeficienții Mel Cepstrali**, care mai apoi sunt introduși în sistemul de recunoaștere.



- În modelul **DeepSpeech** sunt **4 straturi** (Stratul de Intrare, Stratul 2, Stratul 3 și Stratul de Ieșire) care utilizează **conceptul clasic de reacție pozitivă**, folosind **funcția de activare de tip ReLu**.
- **Stratul 4 este stratul recurent**, construit cu **Rețea Neuronală Recurentă și LSTM**, neuronii find activați cu **funcție Tanh**.
- Rezultatele obținute din aceste straturi neuronale sunt reprezentate sub forma unor **matrici de probabilități a ocurenței fiecărui caracter pentru fiecare segment de timp analizat.**

Modelul DeepSpeech

- Până a oferi transcrierea vocii identificate, matricea de probabilități mai trebuie introdusă în **Decodorul de Clasificare Temporală Conecționistă (CTC)**.
- Acest decodor are rolul de a forma **cuvinte** din caracterele obținute în blocurile precedente.
- Rețeaua neuronală tinde să identifice mai multe caractere decât există cu adevărat, din cauza schimbărilor de intonație și alte fluctuații ale vocii. Din acest motiv este necesară o **coliziune a caracterelor** pentru a forma cuvinte.
- După ce s-au format cuvintele urmează introducerea acestora în **Modelul de Limbă**, răspunzător de **verificarea contextuală a cuvintelor** (dacă doua cuvinte pot coexista în limba prezentă).
- **Modelul de limbă** este foarte important pentru **corectitudinea contextuală** a cuvintelor din propoziția formată. Atât în limba engleză cât și în limba română există numeroase cuvinte **omofone** (cuvinte care se scriu diferit dar se pronunță identic), această categorie fiind principala sursă de erori în momentul în care se compară transcrierile.
- Pentru sistemul **DeepSpeech** modelul de limbă este realizat cu ajutorul instrumentului **KenLM**, care poate fi ajustat pentru limba dorită.



Word Error Rate (WER)
Character Error Rate (CER)

Necesitate: O metrică obiectivă care măsoară erorile introduse de către decodorul CTC și de către modelul de limbă

$$WER = \frac{S + I + D}{N} \cdot 100$$

- S numărul de substituții (cuvinte care au fost înlocuite)
- I numărul de inserții (cuvinte care au fost introduse)
- D numărul de ștergeri (cuvinte care au fost șterse)
- N numărul total de cuvinte din transcrierea originală

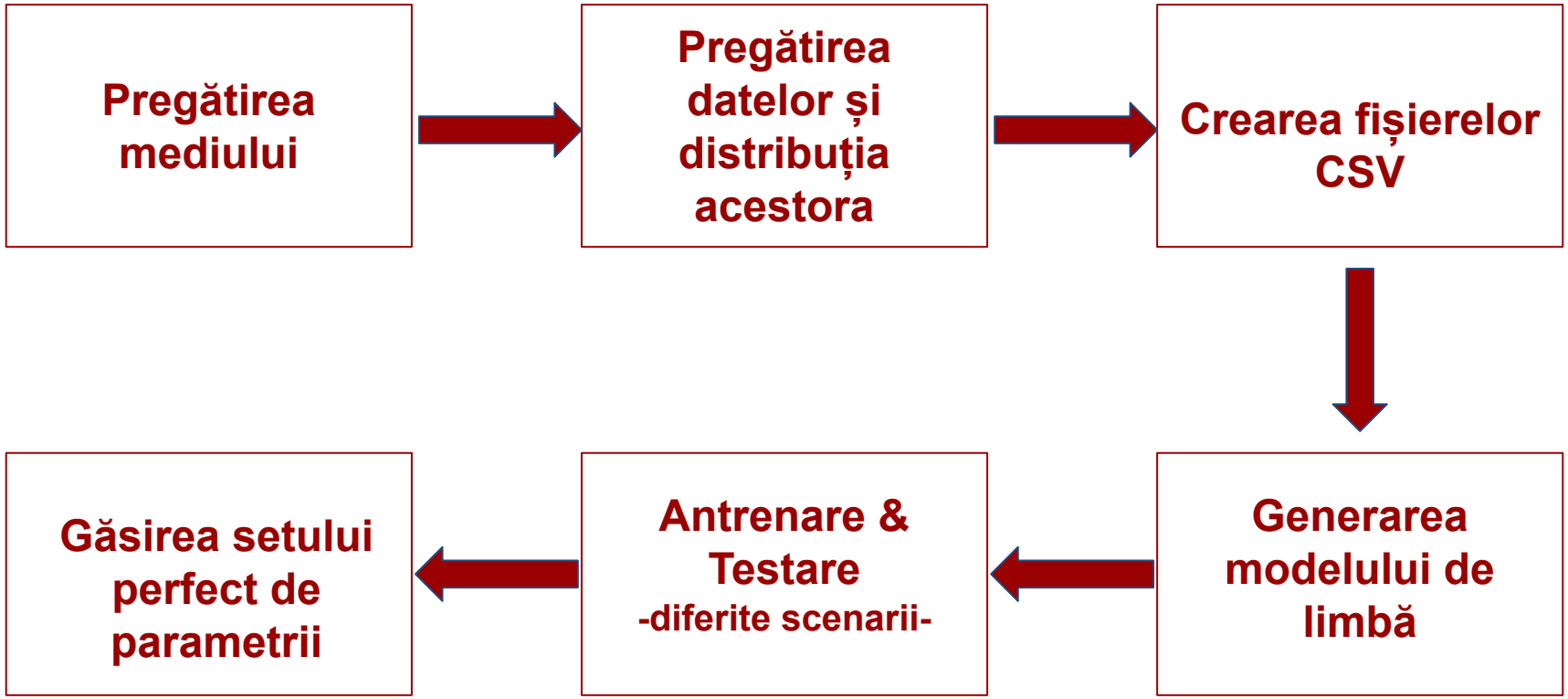
WER: 0.166667, CER: 0.107692

- wav: file:///home/andreea/DeepSpeech/data/test/sam_rnd3_320.wav

- src: "borcanele le închidem cu un capac ermetic și le așezăm în pivniță"
- res: "borcanele le închidem cu un capac ermetic și le așezăm linia"

Original transcript	borcanele le închidem cu un capac ermetic și le așezăm în pivniță		
Detected transcript	borcanele le închidem cu un capac ermetic și le așezăm linia		
Word Error Rate		Character Error Rate	
Sw (substitutions)	1	Sc (substitutions)	2
Iw (insertions)	-	Ic (insertions)	-
Dw (deletions)	1	Dc (deletions)	5
Nw(total words)	12	Nc (total characters)	65
WER	16.6%	**CER**	10.7692%

5. Implementare



Adaptarea pentru Limba Română

1. **Corpusul SWARA versiunea 1**
- 17 vorbitori
- 21 ore și 19 minute
- De înaltă calitate, dar de dimensiune scăzută

2. **RSC**
- 156 vorbitori
- 100 ore
- De dimensiune mare, dar date imperfecte

3. **Corpusul SWARA versiunea 1+2**
- 17 vorbitori + 29 vorbitori
- 21 ore și 19 minute + aprox. 70 ore
- De înaltă calitate, dimensiune scăzută în continuare (LibriSpeech 1000 ore limba engleză)

Modelul de limbă: KenLM

1. Transcrierile SWARA versiunea 1 (5032 cuvinte)
2. Peste 20.000 de articole Wikipedia

Corpusul SWARA

SWARA VERSIUNEA 1	SWARA VERSIUNEA 2
17 Vorbitori 21 ore și 19 minute ~1500 / ~1000 fișiere audio / vorbitor 17794 fișiere audio / corpus 1500 propoziții diferite maximum 10 secunde / audio Fs=48 khz / Fs=16khz Fs=16 khz -> 2.3 GB cuvinte unice: 5032	29 vorbitori ~ 60 ore ~ 1797 fișiere audio / vorbitor 51839 fișiere audio / corpus 297 (audiobook) + 1500 propoziții diferite maximum 10 secunde / audio Fs=48 khz / Fs=16khz Fs=16 khz -> 9.2 GB cuvinte unice: 5849

6. Rezultate Experimentale

CORPUS	Vorbitori	Modelul de limbă	WER
SWARA	1	KenLM cu transcrieri	12.9 %

CORPUS	Vorbitori	Modelul de limbă	WER
SWARA	1	KenLM cu transcrieri	12.9 %
SWARA	2 F antrenare / M test	KenLM cu transcrieri	41.64 %

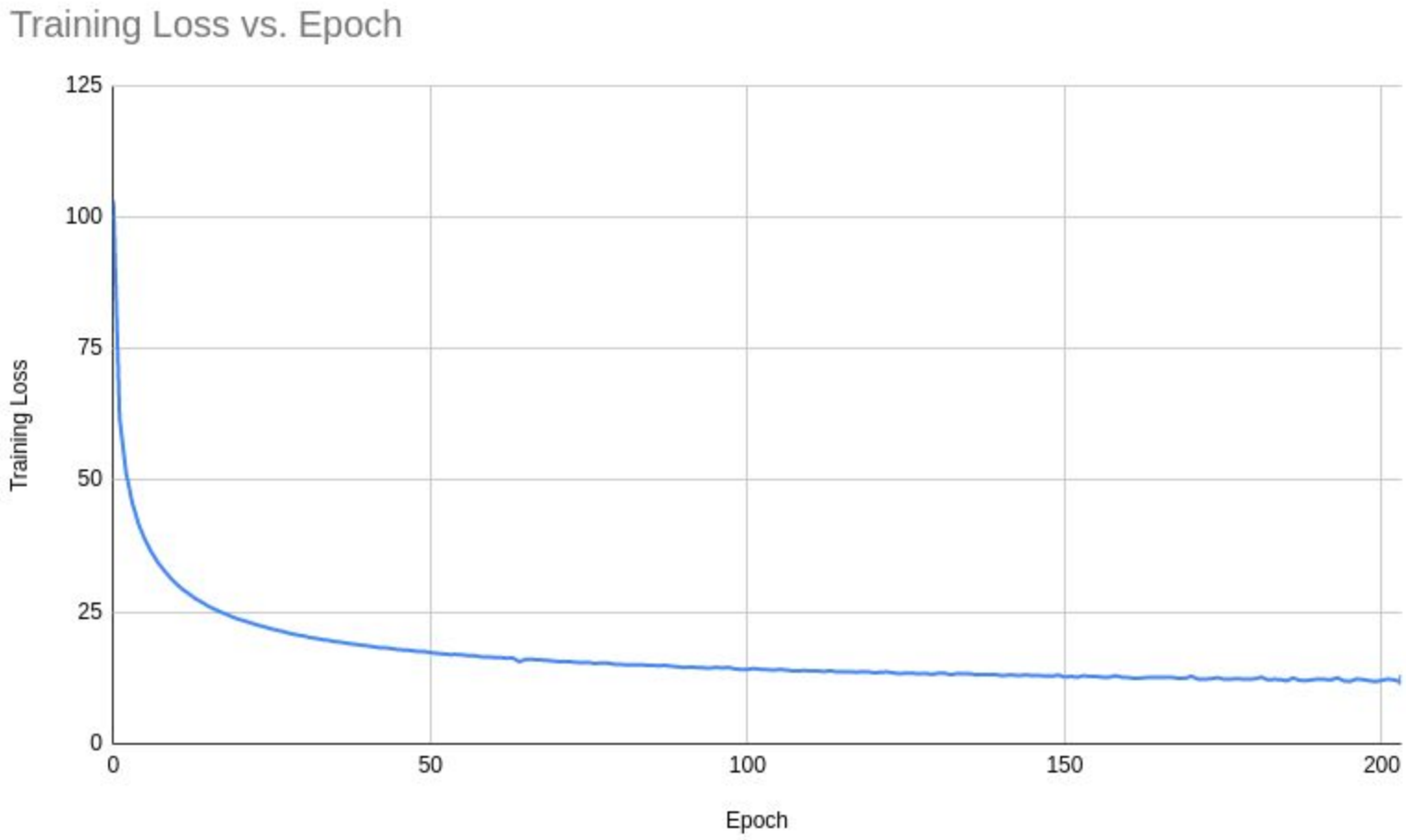
CORPUS	Vorbitori	Modelul de limbă	WER
SWARA	1	KenLM cu transcrieri	12.9 %
SWARA	2 F antrenare / M test	KenLM cu transcrieri	41.64 %
SWARA	17	KenLM cu transcrieri	19.6 %

CORPUS	Vorbitori	Modelul de limbă	WER
SWARA	1	KenLM cu transcrieri	12.9 %
SWARA	2 F antrenare / M test	KenLM cu transcrieri	41.64 %
SWARA	17	KenLM cu transcrieri	19.6 %
SWARA	1	KenLM cu articole	38.75 %

CORPUS	Vorbitori	Modelul de limbă	WER
SWARA	1	KenLM cu transcrieri	12.9 %
SWARA	2 F antrenare / M test	KenLM cu transcrieri	41.64 %
SWARA	17	KenLM cu transcrieri	19.6 %
SWARA	1	KenLM cu articole	38.75 %
SWARA	17	KenLM cu articole	28.3 %

CORPUS	Vorbitori	Modelul de limbă	WER
SWARA	1	KenLM cu transcrieri	12.9 %
SWARA	2 F antrenare / M test	KenLM cu transcrieri	41.64 %
SWARA	17	KenLM cu transcrieri	19.6 %
SWARA	1	KenLM cu articole	38.75 %
SWARA	17	KenLM cu articole	28.3 %
RSC antrenare SWARA testare	156 antrenare 17 testare	KenLM cu articole	47 %

Eroarea de antrenare



7. Concluzii

1. Experimentele au fost realizate pentru o varitate largă de scenarii și configurații (în jur de 1000 ore de antrenări)

2. Ce mai bună configurație: Corpusul SWARA pentru antrenare și testare, Modelul de limbă construit doar cu transcrieri

3. A fost demonstrată necesitatea unui set de date de înaltă calitate și de dimensiuni mari

4. Extindere artificială a corpusului SWARA: prin adăugare de zgomot de fundal

5. Minimizarea timpului de antrenare: folosirea GPU în loc de CPU