



SPEECH RECOGNITION IN BENGALI

Reconnaissance de la parole en bengali

Décembre 2023
Clara Yaïche
Étudiante en alternance NXP - OC
Parcours Machine Learning

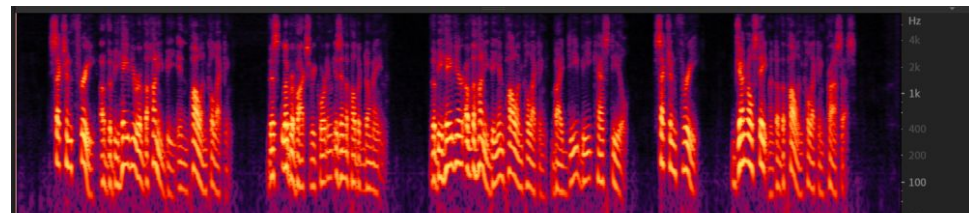
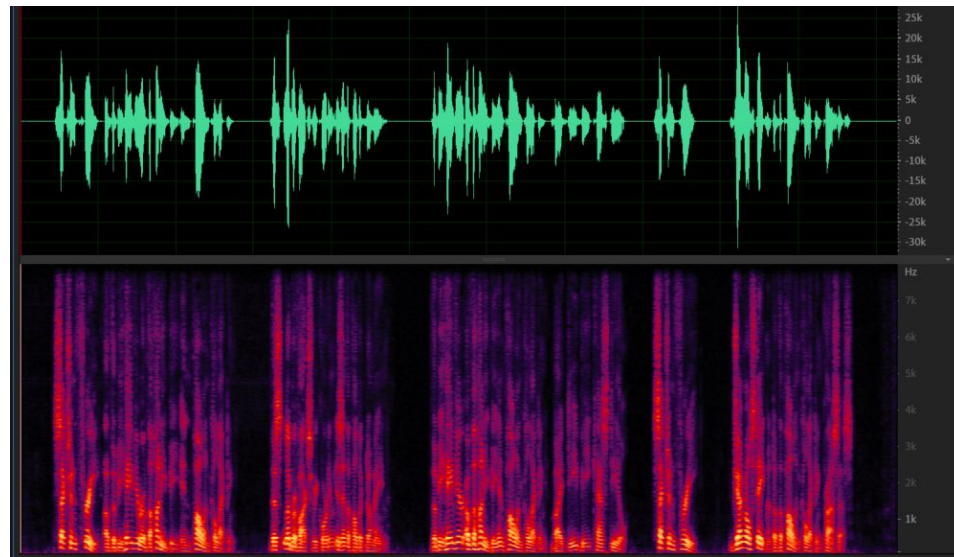
ASR - STT



Automatic Speech Recognition - An Overview,
Microsoft Research de Preethi Jyothi



500 Hertz ~ “e” (magnet)

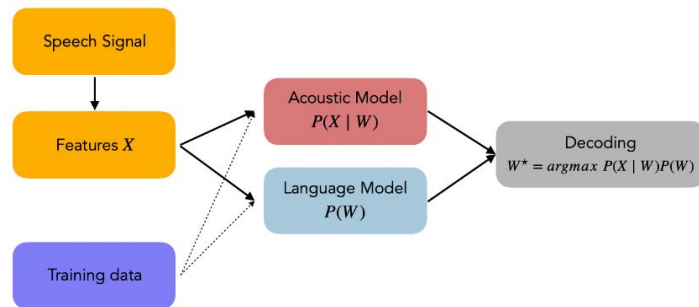


Deux techniques

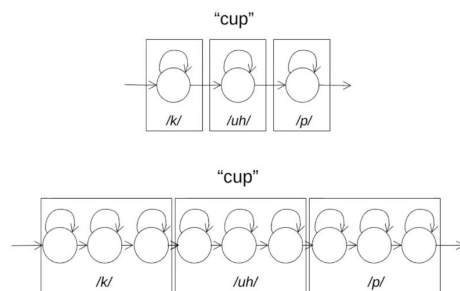
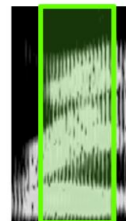
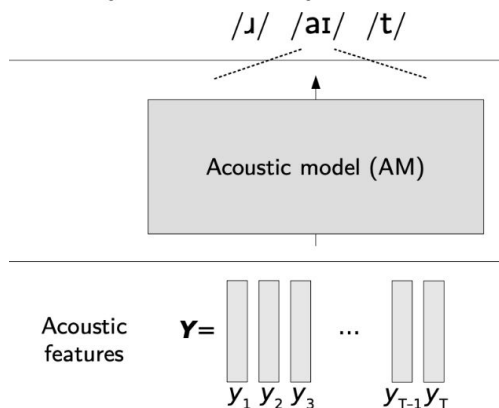


Approches Statistiques en deux parties

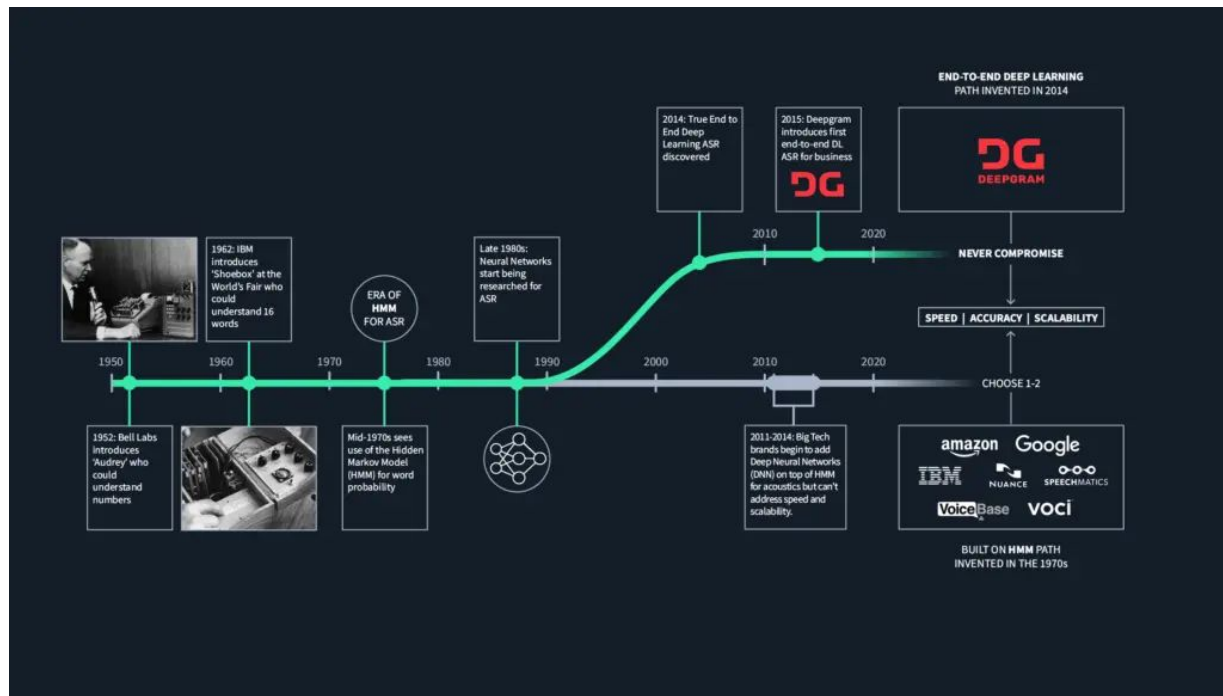
Attention : ASR
n'effectue pas de la
compréhension du
langage, uniquement de
la transcription.
Il a cependant besoin
d'une certaine
compréhension de la
langue pour comprendre
un enchainement
probabiliste des mots et
des sons.



Acoustic Model + Lexicon également



Deux techniques : End-to-End





BENGALI.AI · RESEARCH CODE COMPETITION · 2 MONTHS AGO

Bengali.AI Speech Recognition

Recognize Bengali speech from out-of-distribution audio recordings

[Overview](#) [Data](#) [Code](#) [Models](#) [Discussion](#) [Leaderboard](#) [Rules](#)

	098	099	09A	09B	09C	09D	09E	09F
0	৭ 0980	ঐ 0990	ঔ 09A0	র 09B0	ী 09C0		ঋ 09E0	ঌ 09F0
1	ঁ 0981		ড 09A1		ঢ 09C1		ণ 09E1	ত 09F1
2	় 0982		ঢ 09A2	ল 09B2	ণ 09C2		ত 09E2	থ 09F2
3	ং 0983	ও 0993	ণ 09A3		ত 09C3		থ 09E3	দ 09F3
4		ঔ 0994	ত 09A4		দ 09C4			ধ 09F4



alphasyllabaire
(son=phonème)

monocamérale
(majuscules = minuscules)

1^{re} langue du **Bangladesh**

2^e langue en **Inde**
Langue **maternelle** de
200 millions
de personnes

09

voyelles

39

consonnes

But de la compétition

Word error rate can then be computed as:

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

where

- S is the number of substitutions,
- D is the number of deletions,
- I is the number of insertions,
- C is the number of correct words,
- N is the number of words in the reference ($N=S+D+C$)

source : wikipedia

Grande diversité de **dialectes** et de caractéristiques prosodiques

data “out-of-distribution”



CLARAYAICHE06 · 44S AGO · 2 VIEWS

BengaliAI_STT_exploratory_part1

Python · [Bengali.AI Speech Recognition](#), [bengali-ai-asr-competition](#)

[Notebook](#) [Input](#) [Output](#) [Logs](#) [Comments \(0\)](#) [Settings](#)

[lien du notebook](#)

La base de données d'entraînement

Fichiers mp3

1180 heures, 26 GB

32 kHz, Float

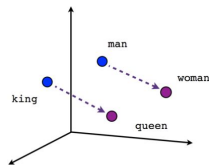
H/F

Data cleaning :

Keep only wer <70% on others ASR
equivalent to noise and sound level cleaning

Preparation :

audio : Mel-Log Spectrogram
text : tokenizer (+ embedding)



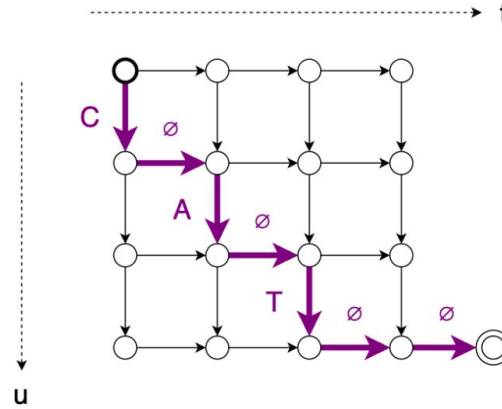
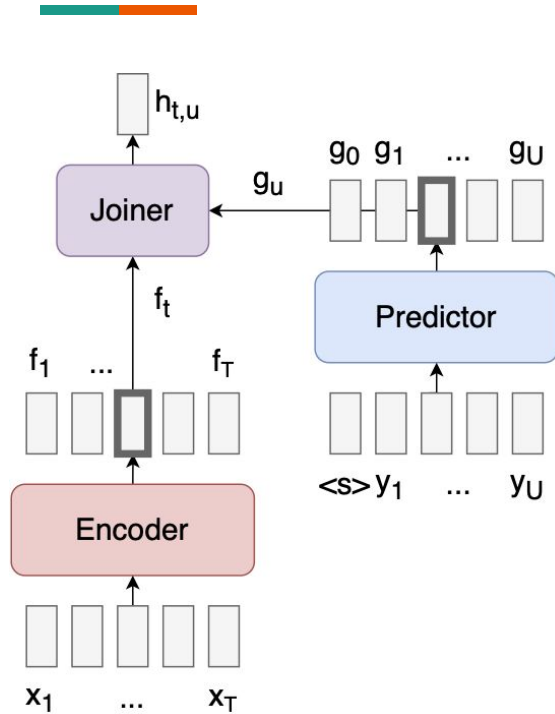
```
{"<s>": 1, "<pad>": 0, "</s>": 2, "<unk>": 3, "\u0987": 4,
```

	id	sentence	split
0	000005f3362c	ও বলেছে আপনার ঠিকানা!	train
1	00001dddd002	কোন মহান রাষ্ট্রের নাগরিক হতে চাও?	train
2	00001e0bc131	আমি তোমার কষ্টটা বুঝছি, কিন্তু এটা সঠিক পথ না।	train
3	000024b3d810	নাচ শেষ হওয়ার পর সকলে শরীর ধুয়ে একসঙ্গে ভোজন...	train
4	000028220ab3	হুমম, ওহ হেই, দেখো।	train

```
train_dataset = CustomAudioDataset(data_folder,  
                                   vocabulary_location,  
                                   train_dataframe)  
  
train_dataloader = DataLoader(train_dataset,  
                              collate_fn=collate_fn,  
                              batch_size=BATCH_SIZE, num_workers=4)
```

*code en annexe

RRN-T : Transducer



La fonction de coût prend en compte tous les alignements possibles.

Implémentation en Pytorch mais problème de backpropagation

Whisper : OpenAI

Multitask training data (680k hours)

English transcription

- 🗣️ "Ask not what your country can do for ..."
- 🗣️ Ask not what your country can do for ...

Any-to-English speech translation

- 🗣️ "El rápido zorro marrón salta sobre ..."
- 🗣️ The quick brown fox jumps over ...

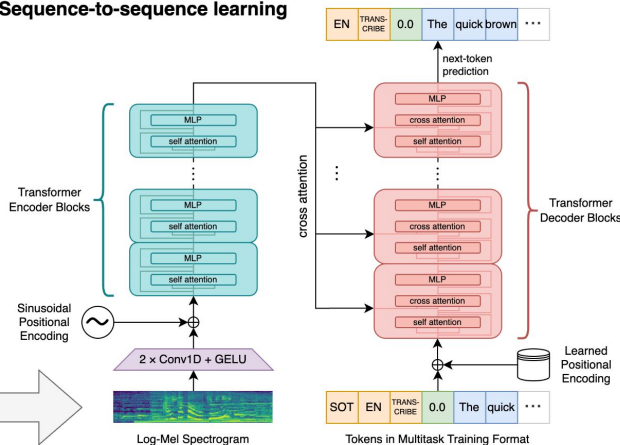
Non-English transcription

- 🗣️ "언덕 위에 올라 내려다보면 너무나 넓고 넓은 ..."
- 🗣️ 언덕 위에 올라 내려다보면 너무나 넓고 넓은 ...

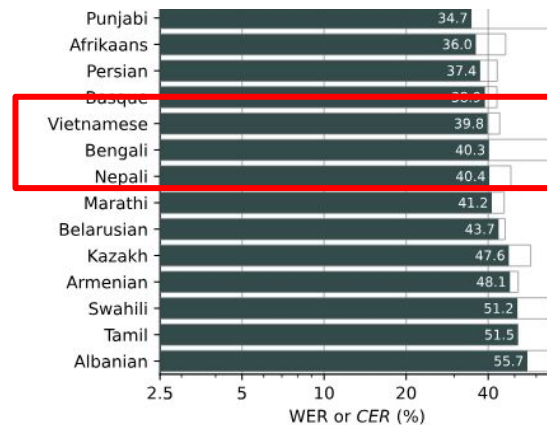
No speech

- 🔊 (background music playing)
- 🔊

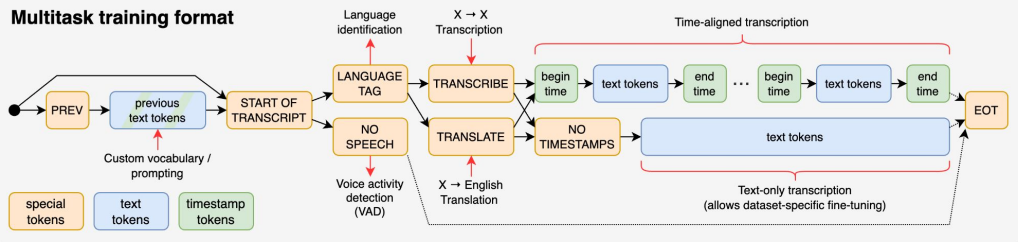
Sequence-to-sequence learning



WER de 40 % sur Common Voice
(clean speech), 10 % en Français






Multitask training format














Fine-Tuning









Hugging Face

 openai/whisper-small   like 106

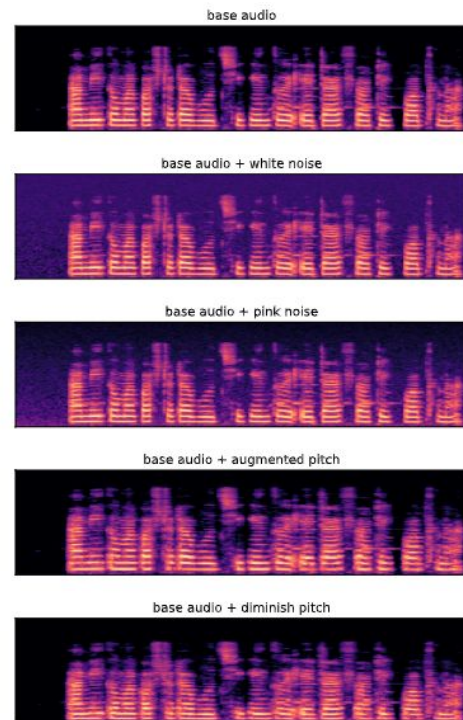
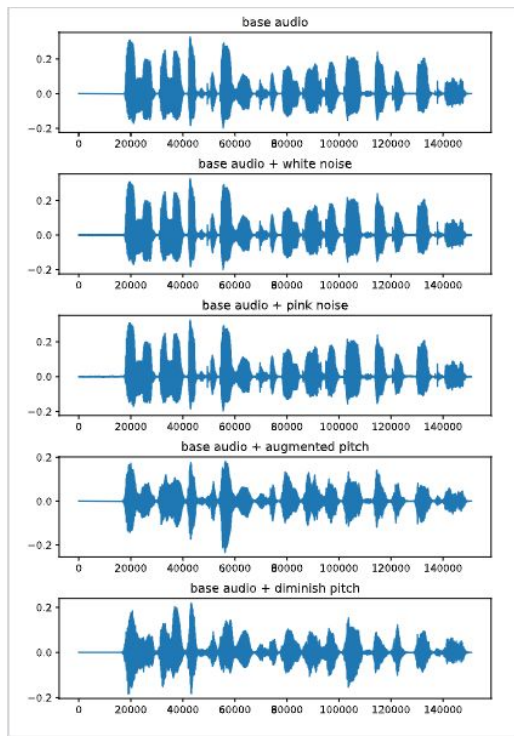
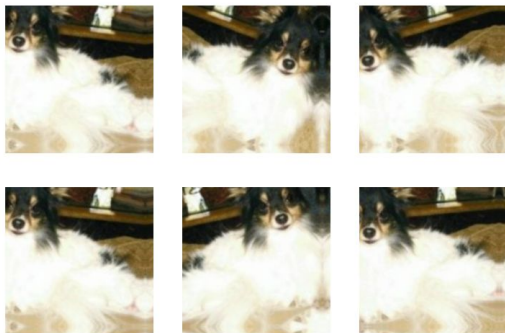
 Automatic Speech Recognition  Transformers  PyTorch  TensorFlow  JAX  Safetensors  99 languages whisper audio hf-asr-leaderboard  Eval Results  Inference Endpoints

 arxiv:2212.04356  License: apache-2.0

 bangla-speech-processing/BanglaASR   like 8

 Automatic Speech Recognition  Transformers  PyTorch  Safetensors whisper audio  Inference Endpoints  License: mit

Data augmentation



autres augmentations non-utilisées : roll , spec-augment : masquage fréquentiel (après analyse)

Spécificité de l'entraînement



gradient_accumulation_steps (gas)

Taille effective du batch = gas *
Taille batch

la notion de “steps” : nombre de
mise à jour des poids

warmup_steps : augmentation du
gradient en début d'apprentissage

```
{
  'loss': 2.3925, 'learning_rate': 0.0, 'epoch': 0.16}
{'loss': 2.4853, 'learning_rate': 1.0000000000000002e-06, 'epoch': 0.32}
{'eval_loss': 2.402637441689453, 'eval_wer': 135.52168815943728, 'eval_runtime': 46.7447, 'eval_samples_per_second': 2.139, 'eval_steps_per_second': 0.278, 'epoch': 0.4}
{'loss': 2.6314, 'learning_rate': 3e-06, 'epoch': 0.48}
{'loss': 2.2658, 'learning_rate': 5e-06, 'epoch': 0.64}
{'loss': 2.1455, 'learning_rate': 7e-06, 'epoch': 0.8}
{'eval_loss': 1.9902739924841309, 'eval_wer': 288.27667057444313, 'eval_runtime': 45.8182, 'eval_samples_per_second': 2.183, 'eval_steps_per_second': 0.284, 'epoch': 0.8}
{'loss': 1.948, 'learning_rate': 9e-06, 'epoch': 0.96}
{'loss': 1.8413, 'learning_rate': 9.888888888888889e-06, 'epoch': 1.12}
{'eval_loss': 1.67385646810913, 'eval_wer': 255.45134818288395, 'eval_runtime': 46.7391, 'eval_samples_per_second': 2.14, 'eval_steps_per_second': 0.278, 'epoch': 1.2}
{'loss': 1.6722, 'learning_rate': 9.666666666666667e-06, 'epoch': 1.28}
{'loss': 1.6741, 'learning_rate': 9.444444444444445e-06, 'epoch': 1.44}
{'loss': 1.5862, 'learning_rate': 9.222222222222224e-06, 'epoch': 1.6}
{'eval_loss': 1.54243959963684, 'eval_wer': 251.2309495896835, 'eval_runtime': 47.9619, 'eval_samples_per_second': 2.085, 'eval_steps_per_second': 0.271, 'epoch': 1.6}
{'loss': 1.5129, 'learning_rate': 9e-06, 'epoch': 1.76}
{'loss': 1.4348, 'learning_rate': 8.777777777777778e-06, 'epoch': 1.92}
{'eval_loss': 1.466407060623169, 'eval_wer': 171.9812426729191, 'eval_runtime': 48.4451, 'eval_samples_per_second': 2.064, 'eval_steps_per_second': 0.268, 'epoch': 2.0}
{'loss': 1.45, 'learning_rate': 8.555555555555556e-06, 'epoch': 2.08}
{'loss': 1.4206, 'learning_rate': 8.333333333333334e-06, 'epoch': 2.24}
{'loss': 1.3169, 'learning_rate': 8.111111111111112e-06, 'epoch': 2.4}
{'eval_loss': 1.403424382299778, 'eval_wer': 167.1746776084408, 'eval_runtime': 48.969, 'eval_samples_per_second': 2.042, 'eval_steps_per_second': 0.265, 'epoch': 2.4}
{'loss': 1.4174, 'learning_rate': 7.888888888888889e-06, 'epoch': 2.56}
```

```
{
  'loss': 0.858, 'learning_rate': 3e-06, 'epoch': 6.08}
{'loss': 0.8413, 'learning_rate': 2.777777777777778e-06, 'epoch': 6.24}
{'loss': 0.8709, 'learning_rate': 2.555555555555557e-06, 'epoch': 6.4}
{'eval_loss': 1.02837561097308, 'eval_wer': 112.5409640834584, 'eval_runtime': 40.1534, 'eval_samples_per_second': 2.49, 'eval_steps_per_second': 0.324, 'epoch': 6.4}
{'loss': 0.7937, 'learning_rate': 2.333333333333333e-06, 'epoch': 6.56}
{'loss': 0.7652, 'learning_rate': 2.111111111111111e-06, 'epoch': 6.72}
{'eval_loss': 1.013447315216064, 'eval_wer': 105.04103165298946, 'eval_runtime': 41.7542, 'eval_samples_per_second': 2.395, 'eval_steps_per_second': 0.311, 'epoch': 6.8}
{'loss': 0.9254, 'learning_rate': 1.888888888888889e-06, 'epoch': 6.88}
{'loss': 0.7739, 'learning_rate': 1.666666666666667e-06, 'epoch': 7.04}
{'loss': 0.7979, 'learning_rate': 1.444444444444445e-06, 'epoch': 7.2}
{'eval_loss': 1.008072638320923, 'eval_wer': 104.2203959520047, 'eval_runtime': 42.0506, 'eval_samples_per_second': 2.334, 'eval_steps_per_second': 0.303, 'epoch': 7.2}
{'loss': 0.8590, 'learning_rate': 1.0000000000000002e-06, 'epoch': 7.52}
{'eval_loss': 1.0035457611083984, 'eval_wer': 104.92379835873309, 'eval_runtime': 39.4886, 'eval_samples_per_second': 2.532, 'eval_steps_per_second': 0.329, 'epoch': 7.6}
{'loss': 0.7787, 'learning_rate': 7.777777777777779e-07, 'epoch': 7.68}
{'loss': 0.7718, 'learning_rate': 5.555555555555555e-07, 'epoch': 7.84}
{'loss': 0.8063, 'learning_rate': 3.333333333333335e-07, 'epoch': 8.0}
{'eval_loss': 1.0006736516952515, 'eval_wer': 105.86166471277842, 'eval_runtime': 40.6008, 'eval_samples_per_second': 2.463, 'eval_steps_per_second': 0.32, 'epoch': 8.0}
}

here were missing keys in the checkpoint model loaded: ['proj_out.weight'].
{'train_runtime': 973.5162, 'train_samples_per_second': 0.822, 'train_steps_per_second': 0.103, 'train_loss': 1.240164566604039, 'epoch': 8.0
100%
```


Résultats

baseline :

whisper-small : 400% WER

BenglaASR (whisper fine
tuned by MIT for Bengali)
74 %

whisper-small fine-tuned :

WER : 67 %

sur le jeu de test extrait de
celui d'entrainement...

```
,id,sentence,predicted
0,620e39f17d7f,পরে সিদ্ধ জুড়ে বিদ্যমান ছড়িয়ে পড়ে।,পরে সিদ্ধ জুড়ে বিদ্যমান ছড়িয়ে পড়ে
1,b7caab80349f,এতে সাড়া দিয়ে আজকেতো শিপারি বাংলাদেশে আসবেন বলে আগ্রহ প্রকাশ করেন।,এতে সাহ দিয়ে আজকেকে ভূঁই বিব বাংলাদেশে আজবেন বলে আগ্রহ প্রকাশ করেন।
2,2c45978cd589,জীব বাবা ছিলেন ব্রাহ্মণ বর্ণের একজন পুণি সুপার, ফলে তিনি নিজের সামাজিক ভাবমূর্তি রক্ষার জন্য জন্মের পরপরই তাকে ত্যাগ করেছিলেন।,জর বাবা ছিলেন ব্রাহ্মণবর্ণের
3,7f272ba654dc,নিশ্চয়ই চাইবেন না আমি এই অফিসারের কোনো ক্ষতি করা, শুক্লর জাহাযন না মির হিতে পুগো খুজি দেখ।
4,0d54108cef93,স্বাইজনের প্রস্তুতিতে পরিকল্পনাবদ্ধ এবং ভূমিকম্প সংশ্লিষ্ট বাড়ি বানানোর ওপর জোর দেওয়া হচ্ছে।,স্বাইজনের প্রস্তুতিতে পরিবেশ বান্দব এবং ভূমিকম্প সংশ্লিষ্ট বাড়ি বানানোর
5,969af10bea22,তোরা কখনোই চাকা দিস না।,তোরা কখনোই চাকা দিস না।
6,9d5194ff408f,নাড়া, আমার কাছে উপস্থল করণ আছে।,নাড়া, আমার কাছে উপস্থল করণ আছে।
7,c5dd1523c625,নিজ প্রদেশের পক্ষে তিনি বেশ সফল হয়েছিলেন।,নিজ প্রদেশের পক্ষে তিনি বেশ সফল হয়েছিলেন।
8,ba05a2b07d41,জর সংগৃহীত দশ খাজার পাখির চামড়া তিনি প্রাকৃতিক ইতিহাস জাদুঘরে দান করে দেন।,জর সংগৃহীত দশখাজার পাখির চামড়া তিনি প্রাকৃতিক ইতিহাস জাদুঘরে দান করে দেন।
9,c776519c541a,কই হয় না তোর?,কই হয় না তোর?
10,d7e658cd818,তুমি তোমার শক্তি দিয়ে মানুষের হাত ভাঙতে পারো।,তুমি তোমার শক্তি দিয়ে মানুষের হাত ভাঙতে পারো।
11,46e23d34f1c5,অতএব দাওয়ায় বসিয়া বেদনান আবাব ছুটি চাচিতে লাগিলেন।,অতএব দাওয়ায় বসিয়া বেদনান আবাব ছুটি চাচিতে লাগিলেন।
12,4ead06503681,কুমুদ ভাবিল, হ, এবার মানিব্যাক তোমার চুরি যাবে!.,কুমুদ ভাবিল, হ, এবার মানিব্যাক, তোমার চুরি যাবে।
13,736a94833db5,সেগুলো হচ্ছে.,সেগুলো হচ্ছে।
14,e9bf93ba5238,এই সমস্তই তরাই নামেও পরিচিত।,এই সমস্তই তরাই নামেও পরিচিত।
15,7ba628fa088d,আমি খারাপ বলছি কি প্রিয়জন অগ্রিয় হবে?,আমি খারাপ বলছি কি প্রিয়জন অগ্রিয় হবে?
16,1830f580804a,দাদাকান্দি উপজেলার উত্তর-পশ্চিমাংশে দাদাকান্দি উত্তর ইউনিয়নের অবস্থান।,দাদাকান্দি উপজেলার উত্তর-পশ্চিমাংশে দাদাকান্দি উত্তর ইউনিয়নের অবস্থান।
17,d6f24eaf038a,তিনি ব্রী শালোটিকে নিয়ে বিবেকানন্দের সাথে ভারতে ভ্রমণ করেন।,তিনি ব্রী শালোটিকে নিয়ে বিবেকানন্দের সাথে ভারতে ভ্রমণ করেন।
18,35cfff205602,তিনি ডিজনরি 'আলাদিন'-এ জেসমিন চরিত্রের জন্য অভিনয় দিয়েছিলেন।,তিনি ডিজনরি 'আলাদিন'-এ জেসমিন চরিত্রের জন্য অভিনয় দিয়েছিলেন।
19,4c3441de3d52,মেয়েদের সে বাগাই নেই।,মেয়েদের সে বাগাই নেই।
20,11b1608395df,একই বছর তিনি আরও কয়েকটি ছোট চিত্রকর্ম অমা দিয়ে তৈরি করেছিলেন।,একই বছর তিনি আরও কয়েকটি ছোট চিত্রকর্ম অমা দিয়ে তৈরি করেছিলেন।
21,f10a1a30264f,তাকে প্রভুত করতেন।,তাকে প্রভুত করতেন।
22,73a6f992d017,এটি ভারত উপমহাদেশীয় লোকসংস্কৃতির একটি গুরুত্বপূর্ণ উপকরণ এবং এটি বিভিন্ন অঞ্চলে বিভিন্ন নামে পরিচিত।,এটি ভারত উপমহাদেশীয় লোকসংস্কৃতির একটি গুরুত্বপূর্ণ উপকরণ এবং এটি বিভিন্ন অঞ্চলে বিভিন্ন নামে পরিচিত।
23,f033c5396c99,ইংরেজি, ইংরেজি।,ইংরেজি, ইংরেজি।
```

```
,id,sentence,predicted
0,620e39f17d7f,পরে সিদ্ধ জুড়ে বিদ্যমান ছড়িয়ে পড়ে।,পরে সিদ্ধ জুড়ে বিদ্যমান ছড়িয়ে পড়ে
1,b7caab80349f,এতে সাড়া দিয়ে আজকেতো শিপারি বাংলাদেশে আসবেন বলে আগ্রহ প্রকাশ করেন।,এতে সাহ দিয়ে আজকেকে ভূঁই বিব বাংলাদেশে আজবেন বলে আগ্রহ প্রকাশ করেন।
2,2c45978cd589,জীব বাবা ছিলেন ব্রাহ্মণ বর্ণের একজন পুণি সুপার, ফলে তিনি নিজের সামাজিক ভাবমূর্তি রক্ষার জন্য জন্মের পরপরই তাকে ত্যাগ করেছিলেন।,জর বাবা ছিলেন ব্রাহ্মণবর্ণের
3,7f272ba654dc,নিশ্চয়ই চাইবেন না আমি এই অফিসারের কোনো ক্ষতি করা, শুক্লর জাহাযন না মির হিতে পুগো খুজি দেখ।
4,0d54108cef93,স্বাইজনের প্রস্তুতিতে পরিকল্পনাবদ্ধ এবং ভূমিকম্প সংশ্লিষ্ট বাড়ি বানানোর ওপর জোর দেওয়া হচ্ছে।,স্বাইজনের প্রস্তুতিতে পরিবেশ বান্দব এবং ভূমিকম্প সংশ্লিষ্ট বাড়ি বানানোর
5,969af10bea22,তোরা কখনোই চাকা দিস না।,তোরা কখনোই চাকা দিস না।
6,9d5194ff408f,নাড়া, আমার কাছে উপস্থল করণ আছে।,নাড়া, আমার কাছে উপস্থল করণ আছে।
7,c5dd1523c625,নিজ প্রদেশের পক্ষে তিনি বেশ সফল হয়েছিলেন।,নিজ প্রদেশের পক্ষে তিনি বেশ সফল হয়েছিলেন।
8,ba05a2b07d41,জর সংগৃহীত দশ খাজার পাখির চামড়া তিনি প্রাকৃতিক ইতিহাস জাদুঘরে দান করে দেন।,জর সংগৃহীত দশখাজার পাখির চামড়া তিনি প্রাকৃতিক ইতিহাস জাদুঘরে দান করে দেন।
9,c776519c541a,কই হয় না তোর?,কই হয় না তোর?
10,d7e658cd818,তুমি তোমার শক্তি দিয়ে মানুষের হাত ভাঙতে পারো।,তুমি তোমার শক্তি দিয়ে মানুষের হাত ভাঙতে পারো।
11,46e23d34f1c5,অতএব দাওয়ায় বসিয়া বেদনান আবাব ছুটি চাচিতে লাগিলেন।,অতএব দাওয়ায় বসিয়া বেদনান আবাব ছুটি চাচিতে লাগিলেন।
12,4ead06503681,কুমুদ ভাবিল, হ, এবার মানিব্যাক তোমার চুরি যাবে!.,কুমুদ ভাবিল, হ, এবার মানিব্যাক, তোমার চুরি যাবে।
13,736a94833db5,সেগুলো হচ্ছে.,সেগুলো হচ্ছে।
14,e9bf93ba5238,এই সমস্তই তরাই নামেও পরিচিত।,এই সমস্তই তরাই নামেও পরিচিত।
15,7ba628fa088d,আমি খারাপ বলছি কি প্রিয়জন অগ্রিয় হবে?,আমি খারাপ বলছি কি প্রিয়জন অগ্রিয় হবে?
16,1830f580804a,দাদাকান্দি উপজেলার উত্তর-পশ্চিমাংশে দাদাকান্দি উত্তর ইউনিয়নের অবস্থান।,দাদাকান্দি উপজেলার উত্তর-পশ্চিমাংশে দাদাকান্দি উত্তর ইউনিয়নের অবস্থান।
17,d6f24eaf038a,তিনি ব্রী শালোটিকে নিয়ে বিবেকানন্দের সাথে ভারতে ভ্রমণ করেন।,তিনি ব্রী শালোটিকে নিয়ে বিবেকানন্দের সাথে ভারতে ভ্রমণ করেন।
18,35cfff205602,তিনি ডিজনরি 'আলাদিন'-এ জেসমিন চরিত্রের জন্য অভিনয় দিয়েছিলেন।,তিনি ডিজনরি 'আলাদিন'-এ জেসমিন চরিত্রের জন্য অভিনয় দিয়েছিলেন।
19,4c3441de3d52,মেয়েদের সে বাগাই নেই।,মেয়েদের সে বাগাই নেই।
20,11b1608395df,একই বছর তিনি আরও কয়েকটি ছোট চিত্রকর্ম অমা দিয়ে তৈরি করেছিলেন।,একই বছর তিনি আরও কয়েকটি ছোট চিত্রকর্ম অমা দিয়ে তৈরি করেছিলেন।
21,f10a1a30264f,তাকে প্রভুত করতেন।,তাকে প্রভুত করতেন।
22,73a6f992d017,এটি ভারত উপমহাদেশীয় লোকসংস্কৃতির একটি গুরুত্বপূর্ণ উপকরণ এবং এটি বিভিন্ন অঞ্চলে বিভিন্ন নামে পরিচিত।,এটি ভারত উপমহাদেশীয় লোকসংস্কৃতির একটি গুরুত্বপূর্ণ উপকরণ এবং এটি বিভিন্ন অঞ্চলে বিভিন্ন নামে পরিচিত।
23,f033c5396c99,ইংরেজি, ইংরেজি।,ইংরেজি, ইংরেজি।
```



Conclusion

Source :

<https://people.irisa.fr/Gwenole.Lecorve/lectures/ASR.pdf>

<https://lorenlugosch.github.io/posts/2020/11/transducer/>

https://maelfabien.github.io/machinelearning/speech_reco/#statistical-historical-approach-to-asr

Annexe : custom dataset et dataloader

```
class CustomAudioDataset(Dataset):
    def __init__(self, folder_path, vocabulary_path, dataframe):
        self.folder_path = folder_path
        self.dataframe = dataframe
        self.sampling_rate = FS

        with open(vocabulary_path) as vocabulary_file:
            char_voc = vocabulary_file.read()
        self.vocabulary_to_id = json.loads(char_voc) # convert to dictionary
```

```
    def __len__(self):
        return self.dataframe.shape[0]
```

```
    def __getitem__(self, idx):
        # this function load a single instance of the dataset
        # print(f"idx : {idx}")
        # Audio
        audio_id = self.dataframe["id"][idx]
        file_path = os.path.join(self.folder_path, f"{audio_id}.wav")
        pcm, sample_rate = torchaudio.load(file_path)

        if self.sampling_rate != sample_rate :
            pcm16k = torchaudio.functional.resample(pcm, sample_rate, self.sampling_rate)

        # Label
        sentence = self.dataframe["sentence"][idx]

        # tokenize bengali characters
        sentence_split = [START_TOKEN] + [ char for char in sentence.replace(' ', ' ').split() if len(char) > 0 ]
        sentence_tokenized = [ self.vocabulary_to_id[char] for char in sentence_split ]

        return pcm16k, sentence_tokenized
```

```
    def collate_fn(batch):
        batch_size = len(batch)
        T = [ pcm16k.shape[1] for (pcm16k, _) in batch]
        U = [ len(sentence_tokenized) - 1 for (_, sentence_tokenized) in batch]
        max_t = max(T)
        max_u = max(U)
        spectrograms = []
        sentence_tokenized_pad_s = []

        mfcc = torchaudio.transforms.MFCC(sample_rate=FS,
                                         n_mfcc=MFCC_N, # Number of MFCC coefficients
                                         # 25ms FFT window 10ms frame shift
                                         melkwargs={'n_fft': MFCC_N_FFT,
                                                    "n_mels": N_MELS,
                                                    'hop_length': MFCC_HOP_LENGTH})

        for index in range(batch_size) :
            pcm16k, sentence_tokenized = batch[index]
            # pad audio and mfcc
            pad_right = int( max_t - pcm16k.shape[1])
            assert pad_right >= 0, "error , one audio file superior to max_input_length 16k"
            pcm16kpad = torch.nn.functional.pad(pcm16k, (0, pad_right) , "constant", 0)
            spectrogram = mfcc(pcm16kpad)
            # pad tokens
            pad_right = int(max_u - len(sentence_tokenized))
            sentence_tokenized_pad = torch.nn.functional.pad(torch.tensor(sentence_tokenized), (0, pad_right) , "constant", 0)

            spectrograms.append(spectrogram)
            sentence_tokenized_pad_s.append(sentence_tokenized_pad)

        spectrograms = torch.stack(spectrograms)
        sentence_tokenized_pad_s = torch.stack(sentence_tokenized_pad_s)
        T_f = [int(t / MFCC_HOP_LENGTH) + 1 for t in T] # because of mfcc padding
        return (spectrograms, sentence_tokenized_pad_s, torch.tensor(T_f), torch.tensor(U))
```