CS 224S / Linguist 285
Spoken Language Processing

Tolúlọpẹ́ Ògúnrẹ̀mí | Stanford University | Spring 2025

Lecture 14: Speech Recognition Beyond English

Outline

- How languages can differ from English
- Multilingual large pretrained models
- Datasets
- Language-specific ASR techniques

There are over 7,000 known languages in the world.

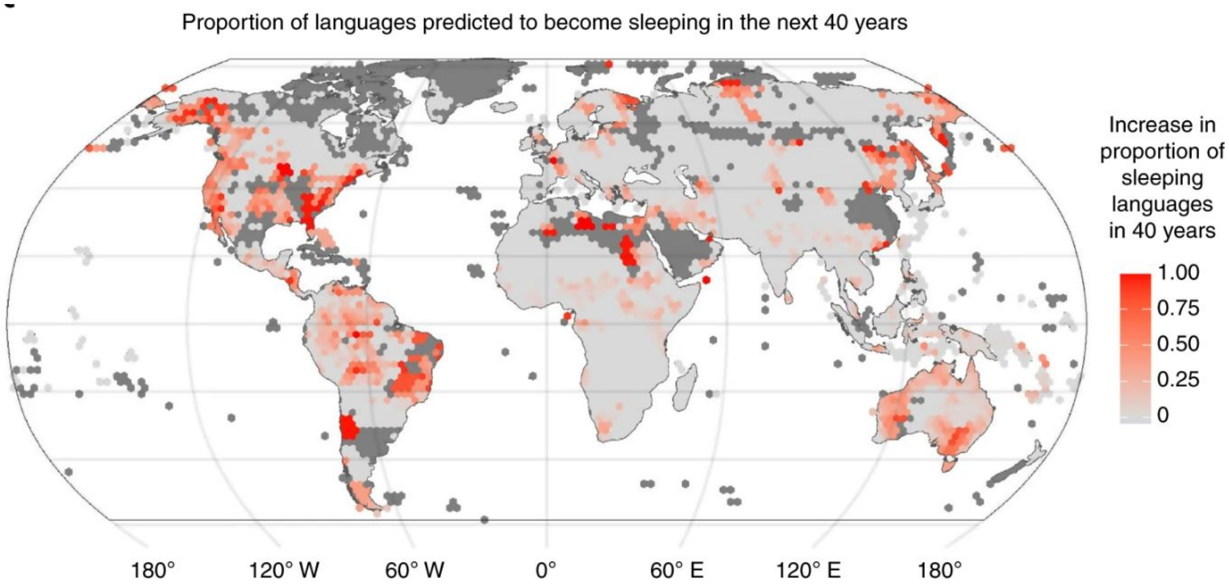
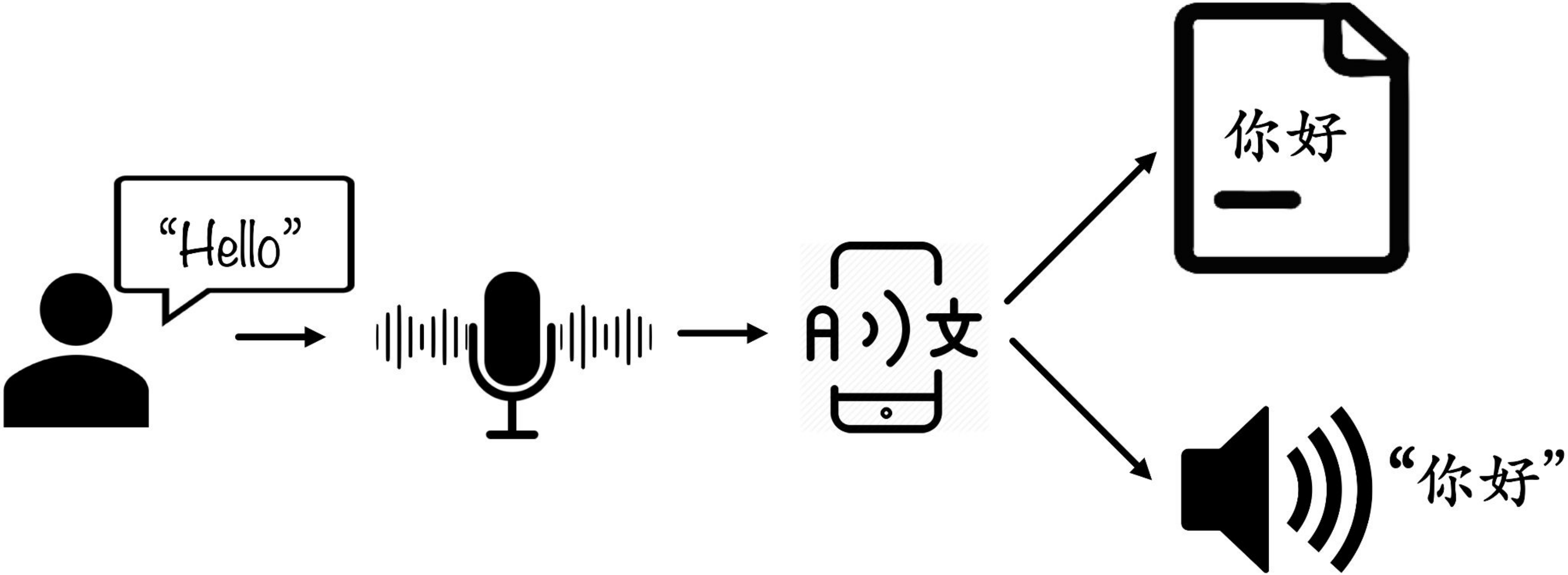

Proportion of languages predicted to become sleeping in the next 40 years
Increase in proportion of sleeping languages in 40 years
1.00
0.75
0.50
0.25
0
180°
120° W
60° W
0°
60° E
120° E
180°

Image from Bromham et. al, 2022

We need to process (as many of) the languages of the world (as we can).

Example: Speech Translation



Most of the models we have seen in this class have been trained with only English data.

Languages vary

Languages can have different scripts

Writing system	Scripts	
Alphabet	Roman	napenda utambuzi wa hotuba
	Greek	Λατρεύω την αναγνώριση ομιλίας
	Cyrillic	Би яриа таних дуртай
	Korean	나는 음성 인식을 좋아해요
Semanto-Phonetic	Chinese	我喜欢语音识别
Syllabic Alphabet	Devanāgarī	मलाई बोली पहिचान मन पर्छ
	Thai	ฉันชอบการรู้จำคำพูด
	Tamil	நான் பேச்சு அங்கீகாரத்தை விரும்புகிறேன்
Abjad	Arabic	أنا أحب التعرف على الكلام
	Hebrew	אני אוהב זיהוי דיבור

Adapted from Tan et. al, 2010

Languages can have lexical tone

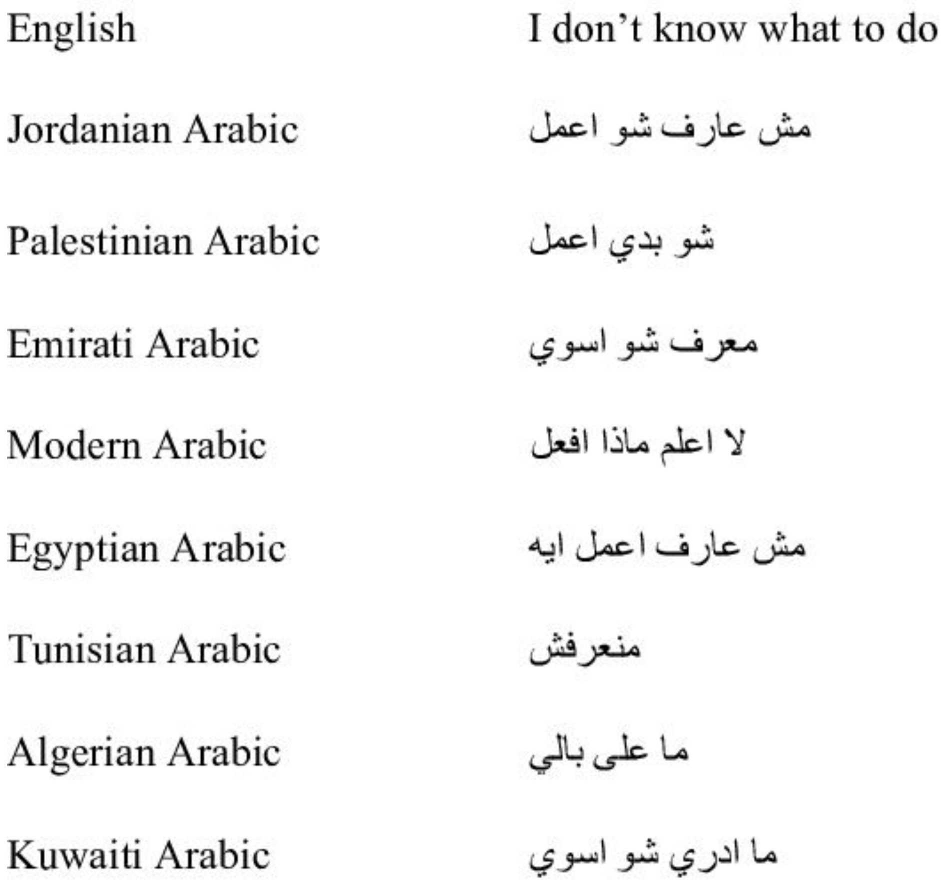
The pitch of the word changes the meaning of the word

wá 🔊

wà 🔊

- Yorùbá provides an orthographic representation of tone, with accents on vowels representing low (grave) and high (acute) tones. No accents is for mid tone.
- Underdots differentiate vowels (o vs ọ) or postalveolar articulation (s vs ṣ).

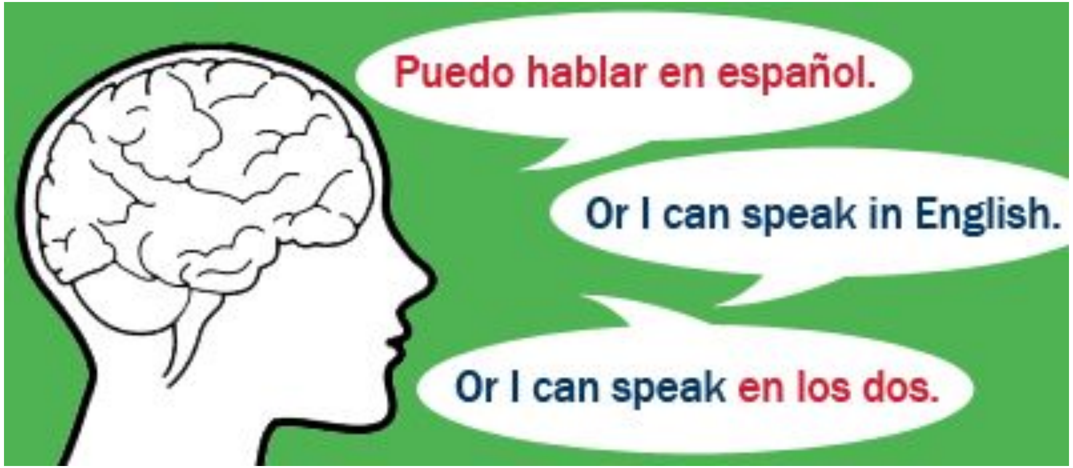
Languages can have different dialects



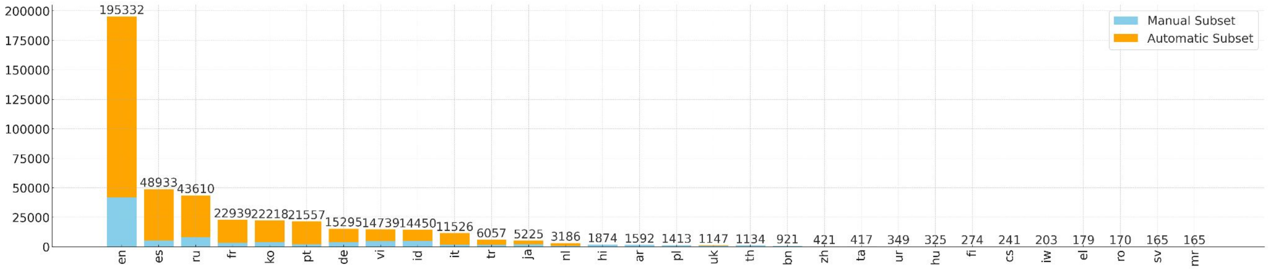
English	I don’t know what to do
Jordanian Arabic	مش عارف شو اعمل
Palestinian Arabic	شو بدي اعمل
Emirati Arabic	معرف شو اسوي
Modern Arabic	لا اعلم ماذا افعل
Egyptian Arabic	مش عارف اعمل ايه
Tunisian Arabic	منعرفش
Algerian Arabic	ما على بالي
Kuwaiti Arabic	ما ادري شو اسوي

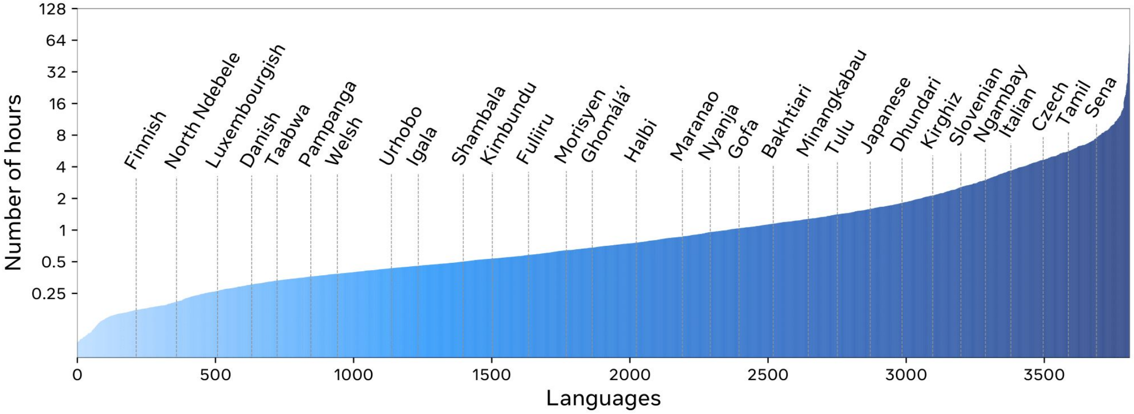
Image from Bani-Hani et.al, 2017

Languages can have codeswitching



Languages can have can have little data available to train models





Multilingual large pretrained speech models

Multilingual versions of English-only models: wav2vec 2.0 XLSR

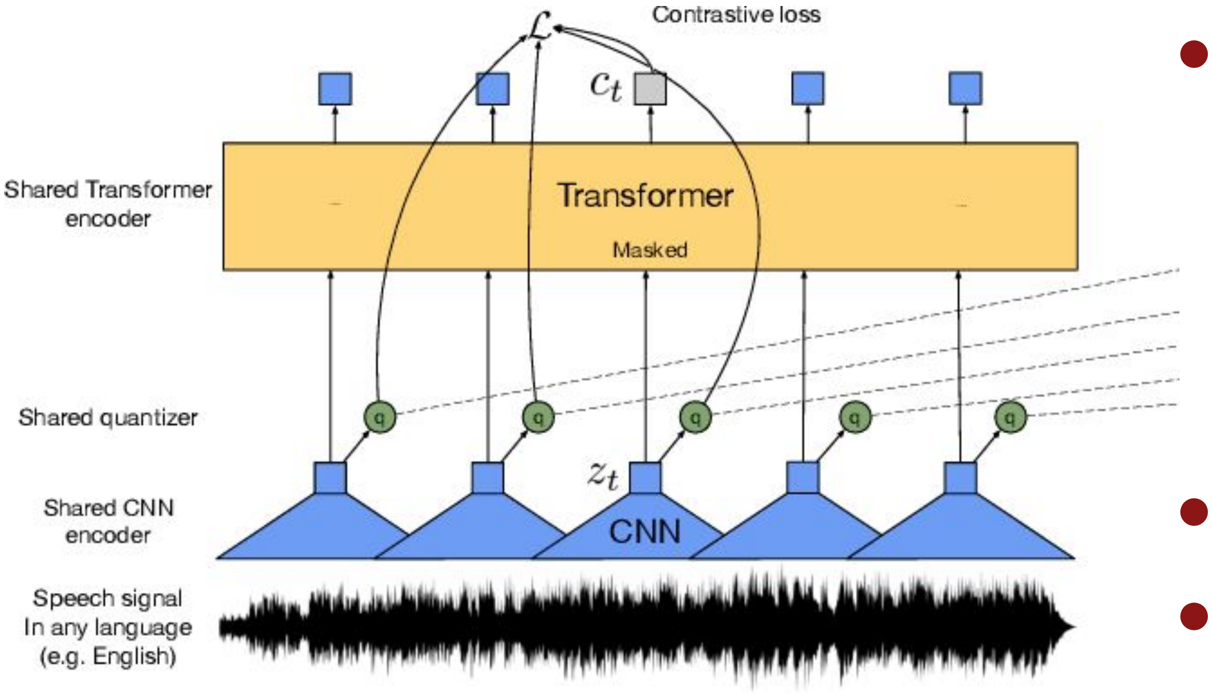


Image from Conneau et.al, 2020

- Trained on Multilingual LibriSpeech, Common Voice and BABEL
- (56,000 hours)
- 53 languages: XLSR-53

Multilingual versions of English-only models: wav2vec 2.0 XLSR

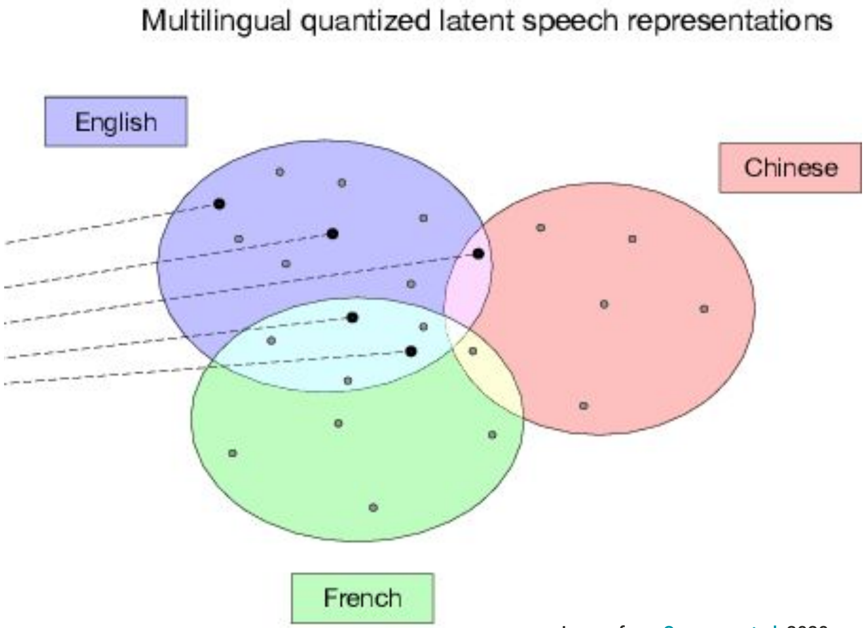


Image from Conneau et.al, 2020

- Latent multilingual speech representations are theorised

Multilingual versions of English-only models: wav2vec 2.0 XLS-R

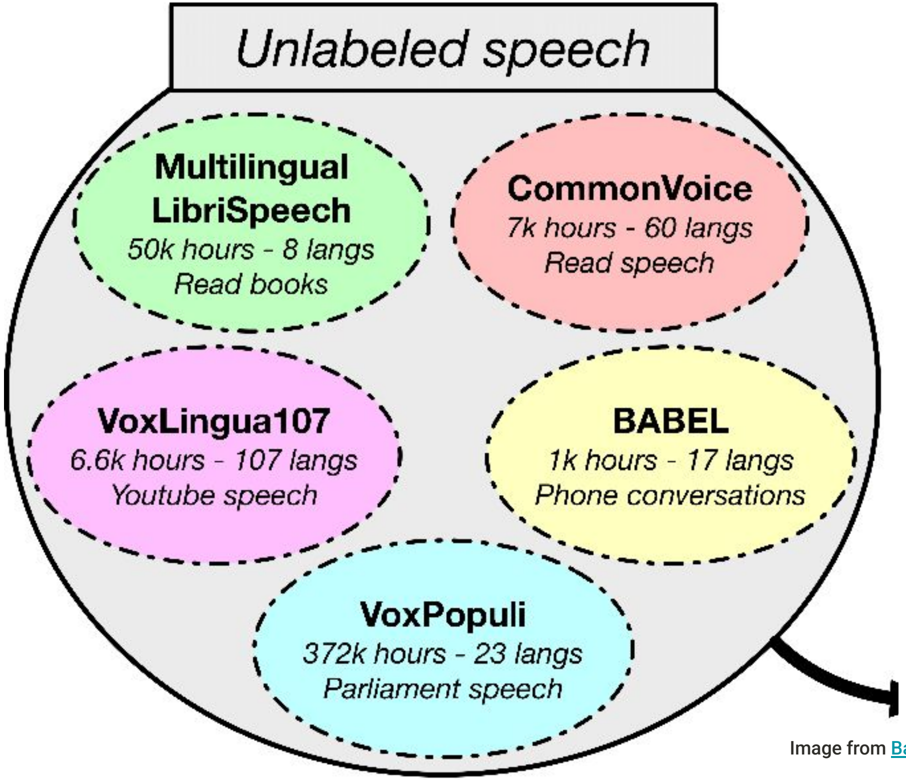


Image from Babu et. al, 2021

- Trained on XLSR datasets and Vox Lingua 107 and Vox Populi, totalling 436,000 hours

Multilingual versions of English-only models: wav2vec 2.0 XLS-R

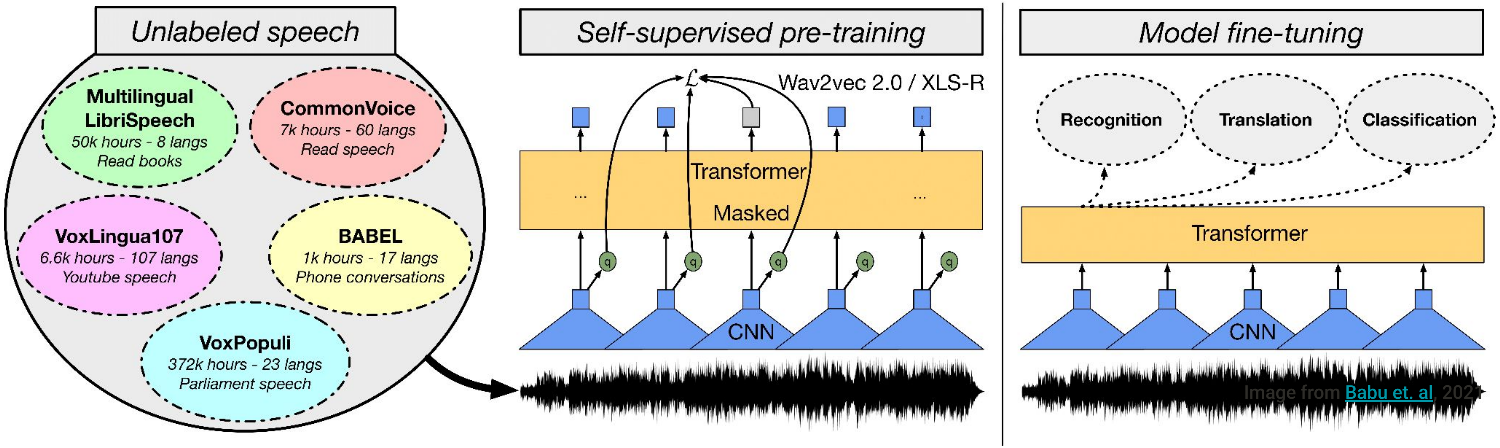


Image from Babu et. al, 2021

- Tested on ASR and AST (Automatic Speech Translation)

Multilingual from the start: Whisper

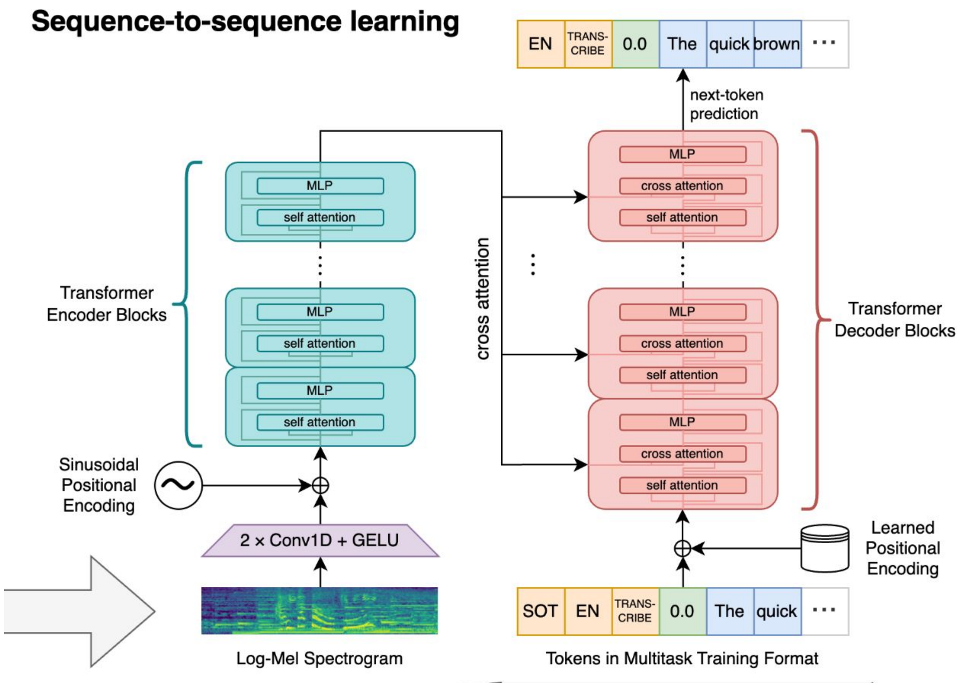

Sequence-to-sequence learning
EN
TRANS-CRIBE
0.0
The
quick
brown
···
next-token prediction
MLP
cross attention
self attention
Transformer Encoder Blocks
Transformer Decoder Blocks
cross attention
Sinusoidal Positional Encoding
2 × Conv1D + GELU
Learned Positional Encoding
SOT
Log-Mel Spectrogram
Tokens in Multitask Training Format

Image from Radford et. al, 2022

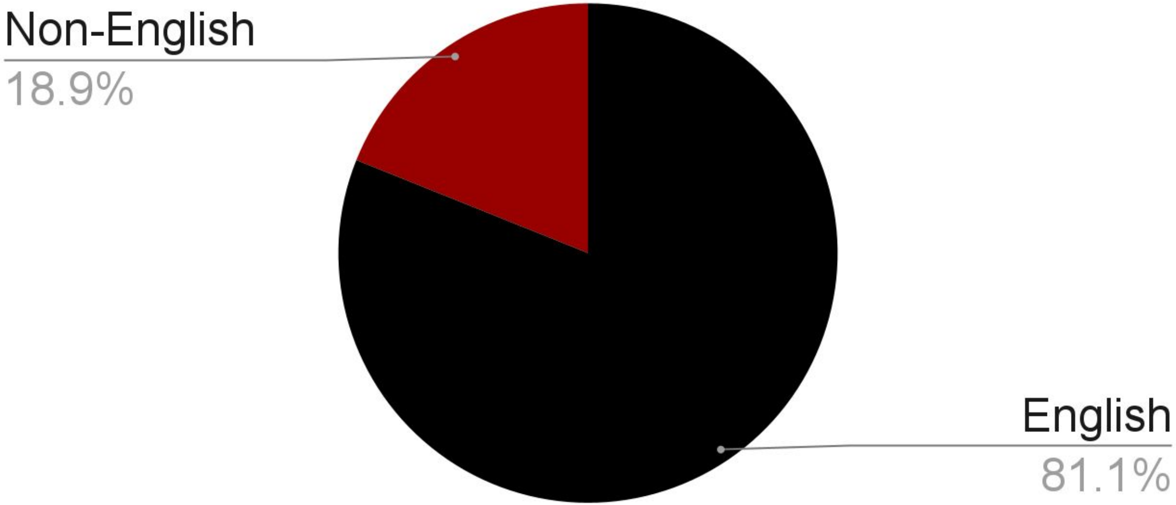
- “Multilingual and multitask”
- Trained with 680,000 hours of data
- Training data is not publicly available.

What is the data distribution?

How multilingual are these models?

wav2vec 2.0 XLSR

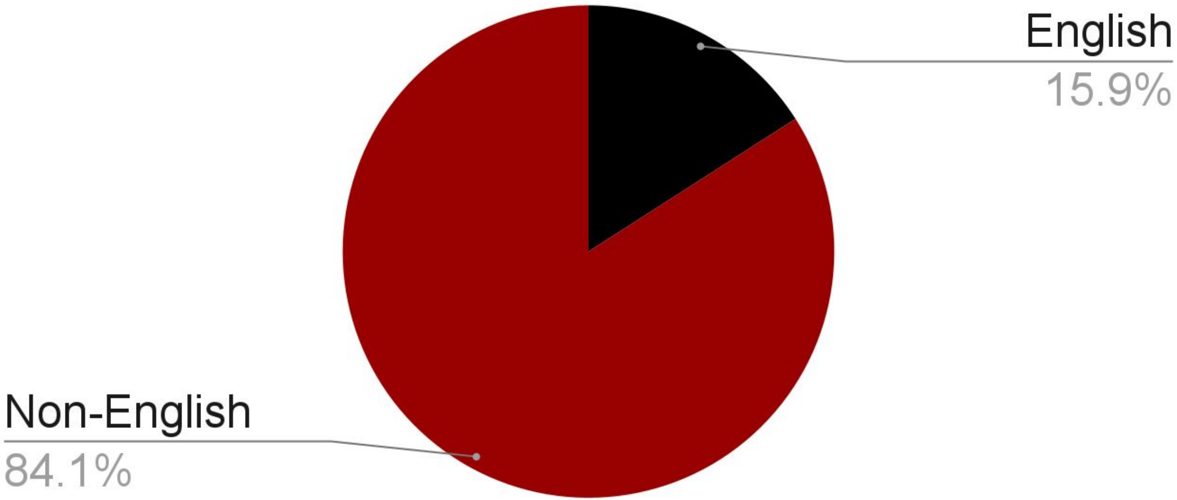
Languages in wav2vec 2.0 XLSR



How multilingual are theses models?

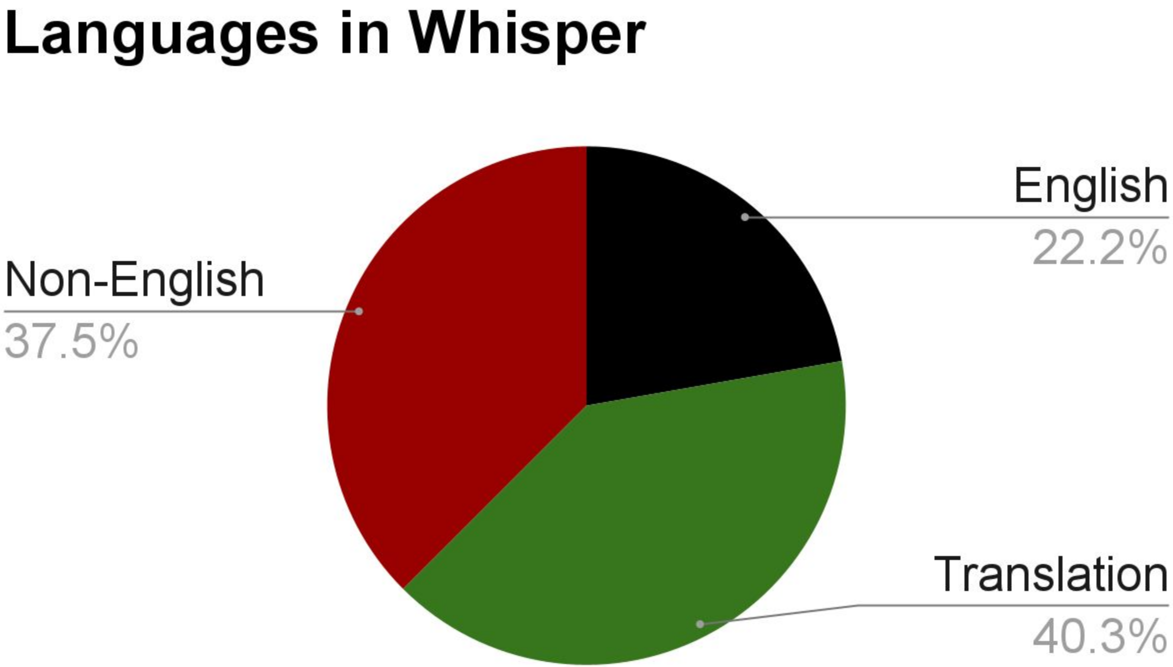
wav2vec 2.0
XLS-R

Languages in wav2vec 2.0 XLS-R



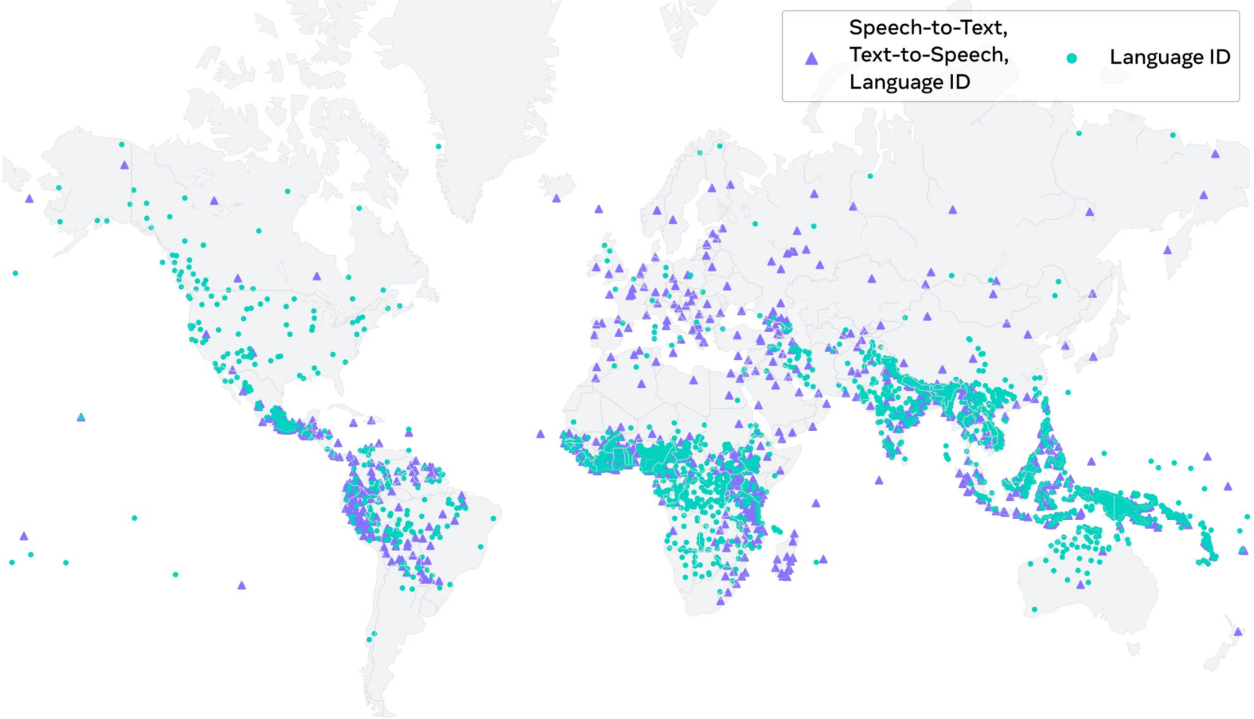
How multilingual are these models?

Whisper



Scaling up number of languages: massively multilingual speech models

Including more of the world’s languages: MMS



- Pre-trained wav2vec 2.0 models covering 1,406 languages
- A single multilingual automatic speech recognition model for 1,107 languages
- Language identification model for 4,017 languages

Pratap et. al, 2023

Language specific adapter weights: MMS

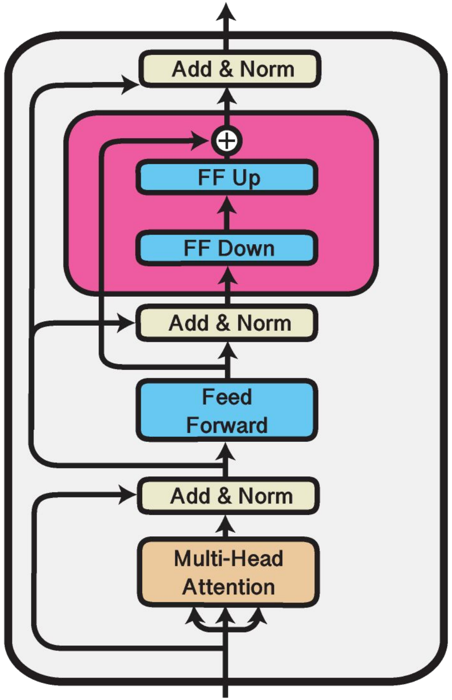
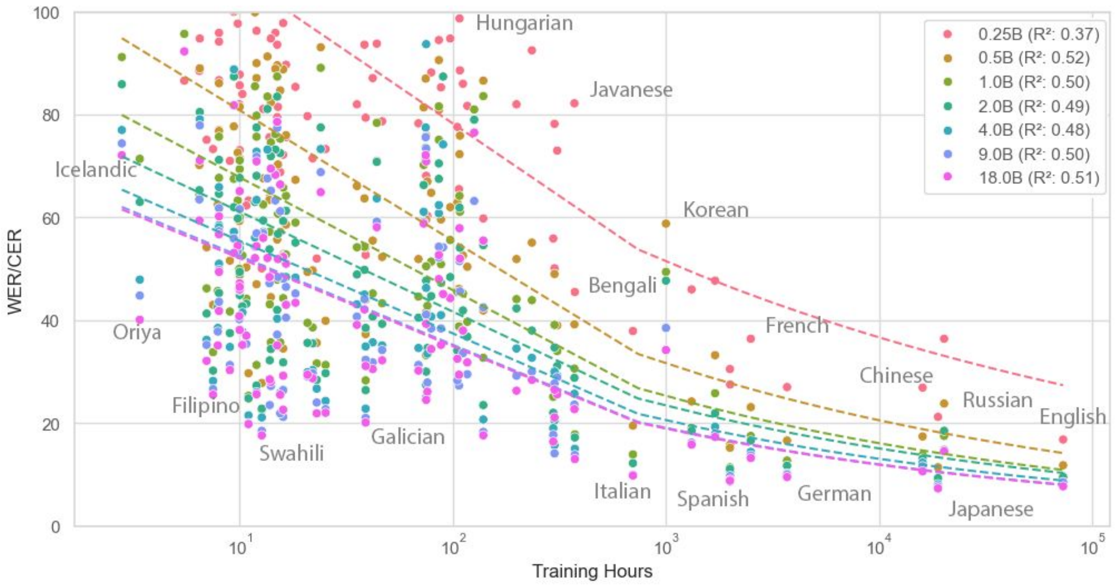


Image from AdapterHub

- MMS models (300 M, small and 1B, large) are trained with every language
- Authors train Houlsby Adapters (Houlsby et. al, 2019) for each language.
- There is an adapter in each transformer block, after the last feed-forward block.
- Each adapter constitutes an extra 2 million parameters.
- Authors also train a linear layer with CTC loss for each language vocabulary.

Pratap et. al, 2023

An open source reproduction of Whisper: OWLS

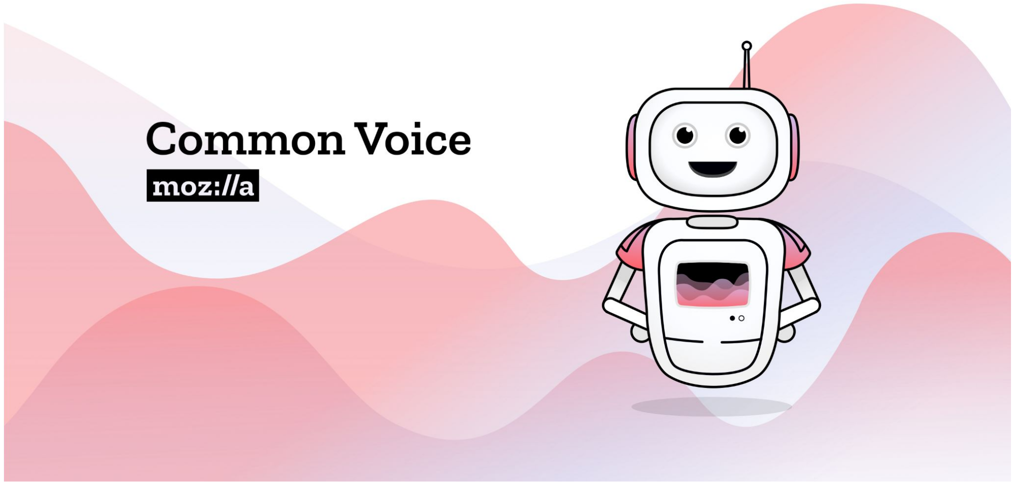


- Chen et. al investigate scaling laws for Whisper-style models
- OWLS is a collection of 13 AST/AST models trained with up to 360 000 hours of publicly available data.
- The largest model is an 18B parameter model.
- Training data is 180 000 hours publicly available labelled data and 180 000 of cleaned Yodas data.

Chen et. al, 2025

Open Source Multilingual Datasets

Common Voice



- Multilingual living dataset
- 30,000 recorded hours covering 124 languages
- Anyone can set up a Common Voice page for their language
- Anyone can record utterances for the dataset
- Dataset is noisier than LibriSpeech due to less controlled recording environments

Attempting to open source datasets: Yodas

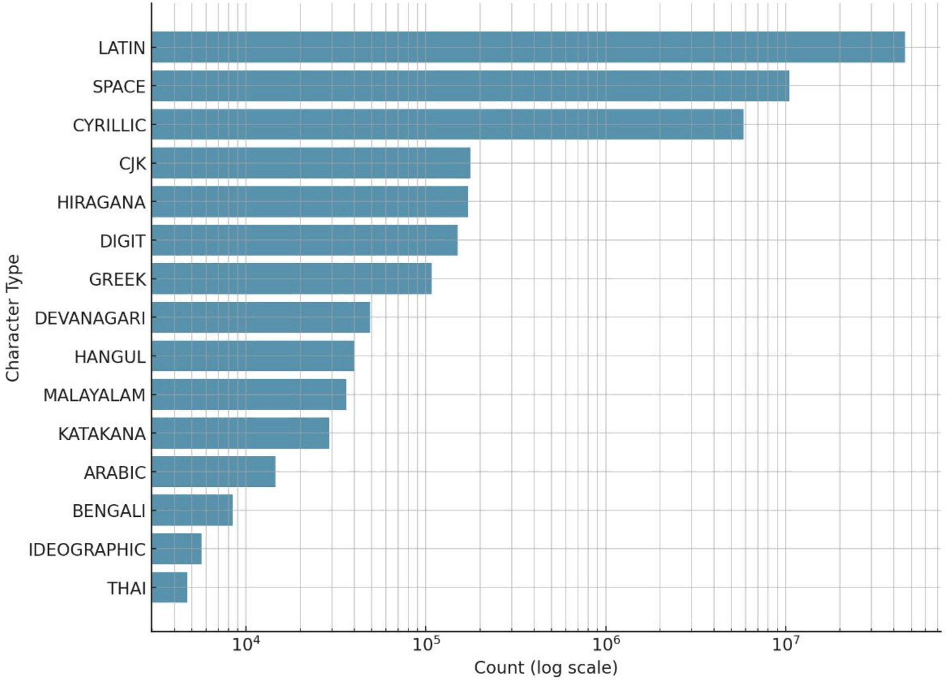
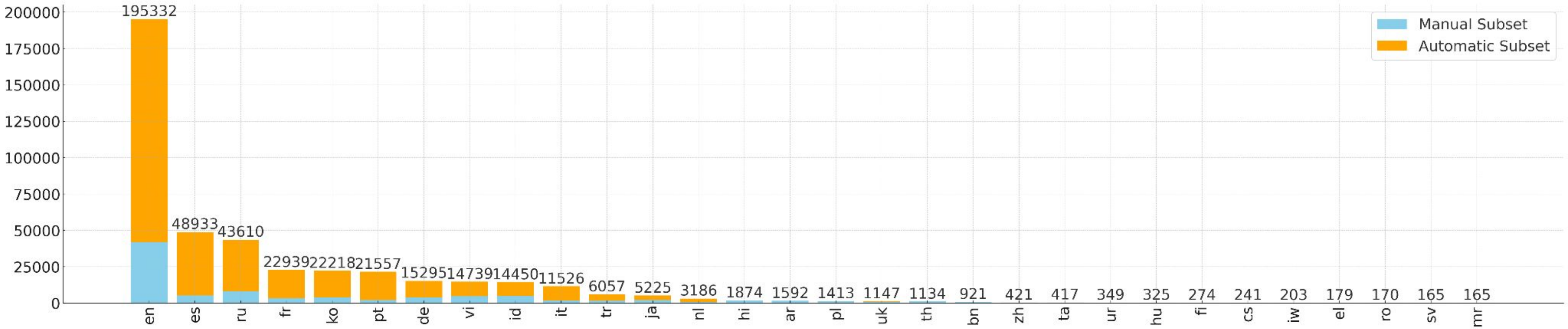


Image from WAVLab post

- Youtube-Oriented Dataset for Audio and Speech
- Result of a 6-month crawl of YouTube followed by alignment of transcript to audio.
- 500,000 hours of data across 140 languages.
- 420,000 hours of transcribed data.

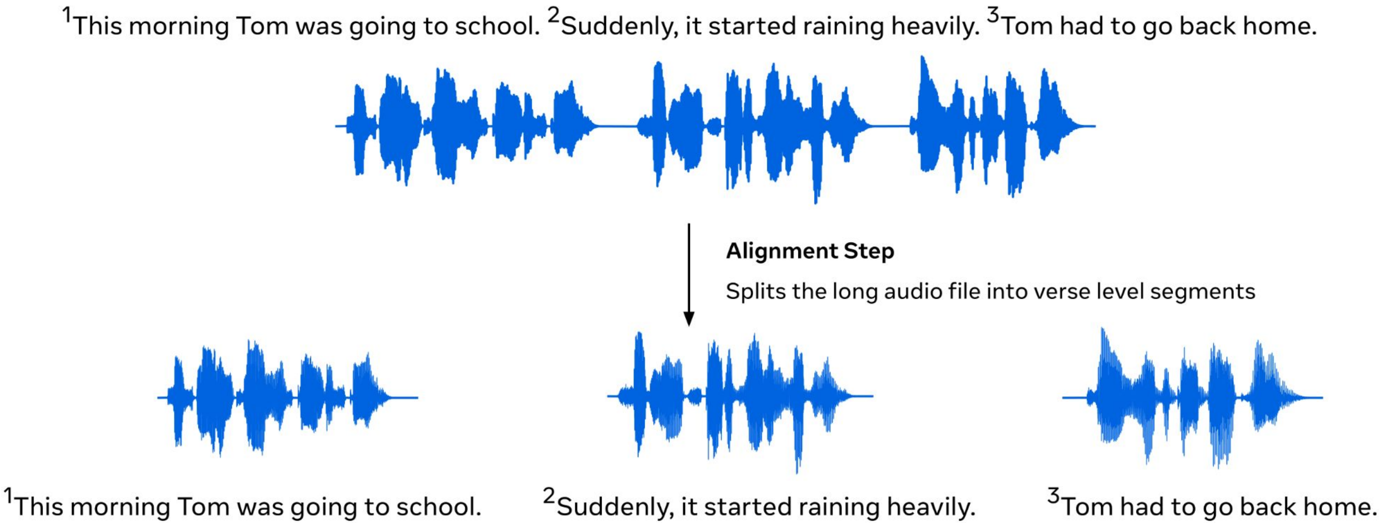
Attempting to open source datasets: Yodas



- Youtube-Oriented Dataset for Audio and Speech
- Result of a 6-month crawl of YouTube followed by alignment of transcript to audio.
- 500,000 hours of data across 140 languages.
- Most of the data is in English.

Image from paper.

MMS-lab: New Testament in 1107 languages



[1]This morning Tom was going to school. [2]Suddenly, it started raining heavily. [3]Tom had to go back home.

Alignment Step

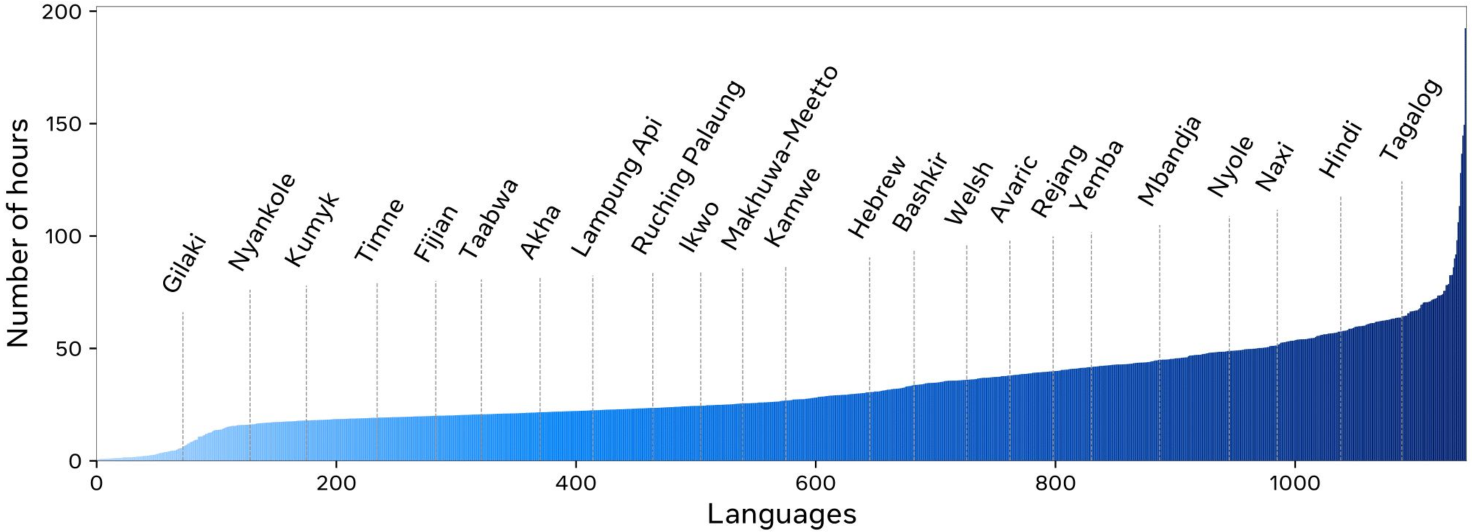
Splits the long audio file into verse level segments

[1]This morning Tom was going to school. [2]Suddenly, it started raining heavily. [3]Tom had to go back home.

- Chapters are aligned with forced alignment.
- Background music is removed.

Pratap et. al, 2023

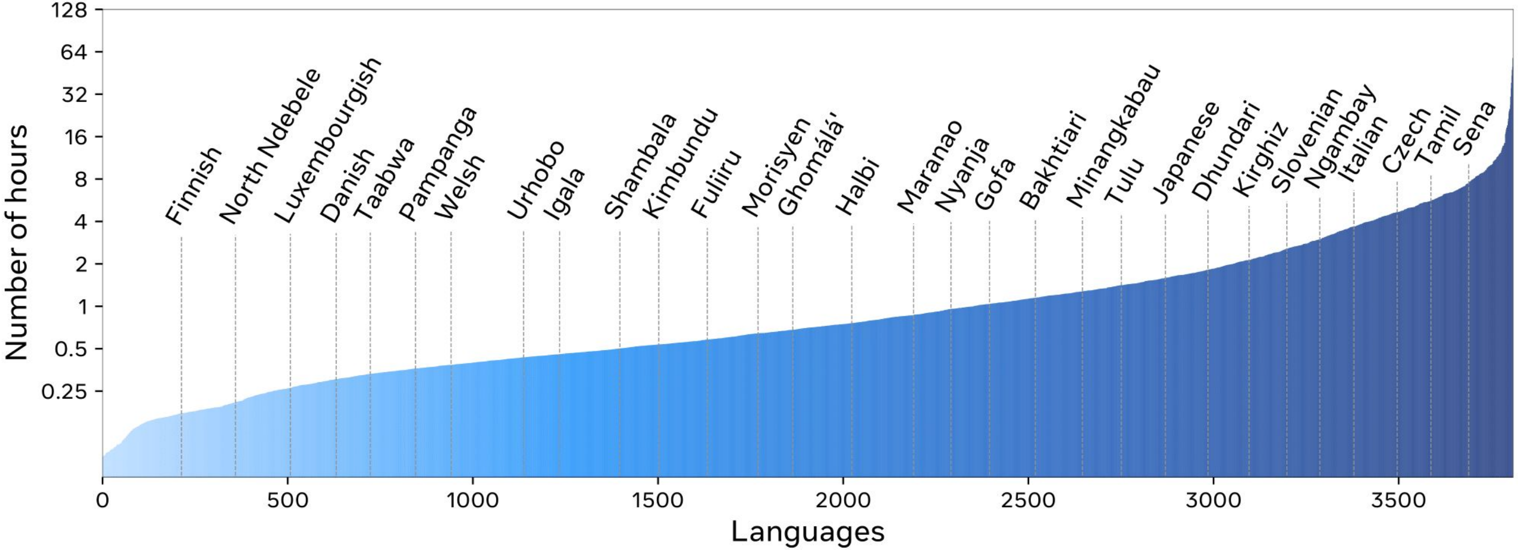
MMS-lab: New Testament in 1107 languages



- Dataset distribution across languages.
- Low quality samples are filtered out.

Pratap et. al, 2023

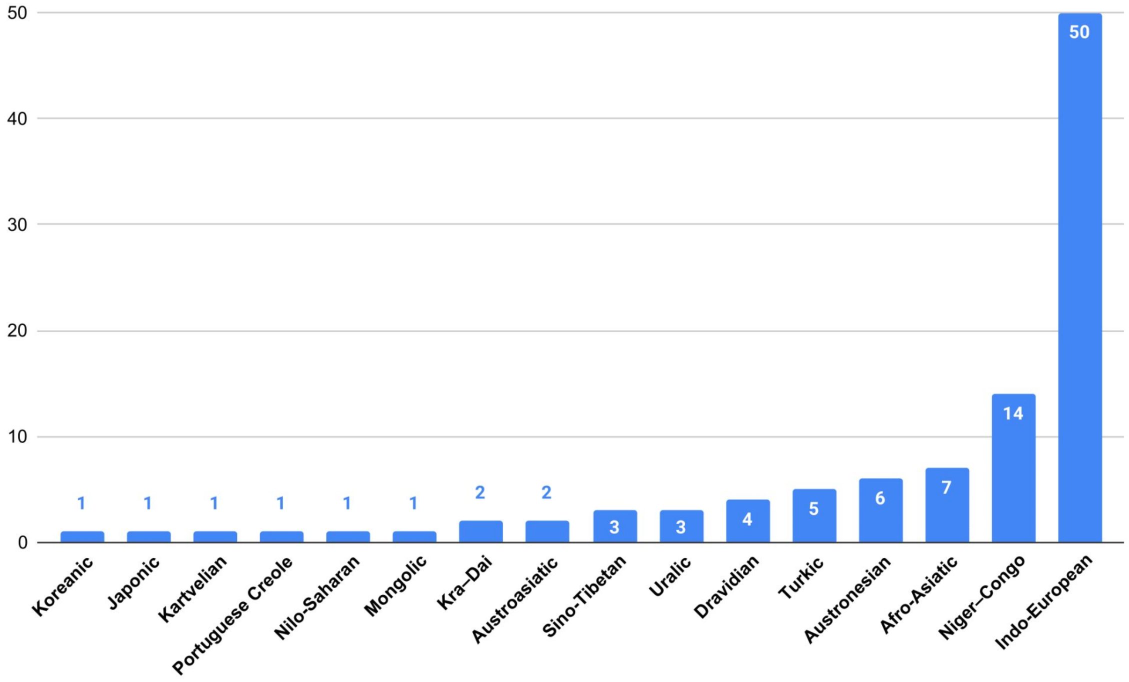
MMS-unlab



- Data from the Global Recordings Network, “recordings of Bible stories, evangelistic messages, scripture readings, and songs in more than 6,255 languages and dialects”.

Pratap et. al, 2023

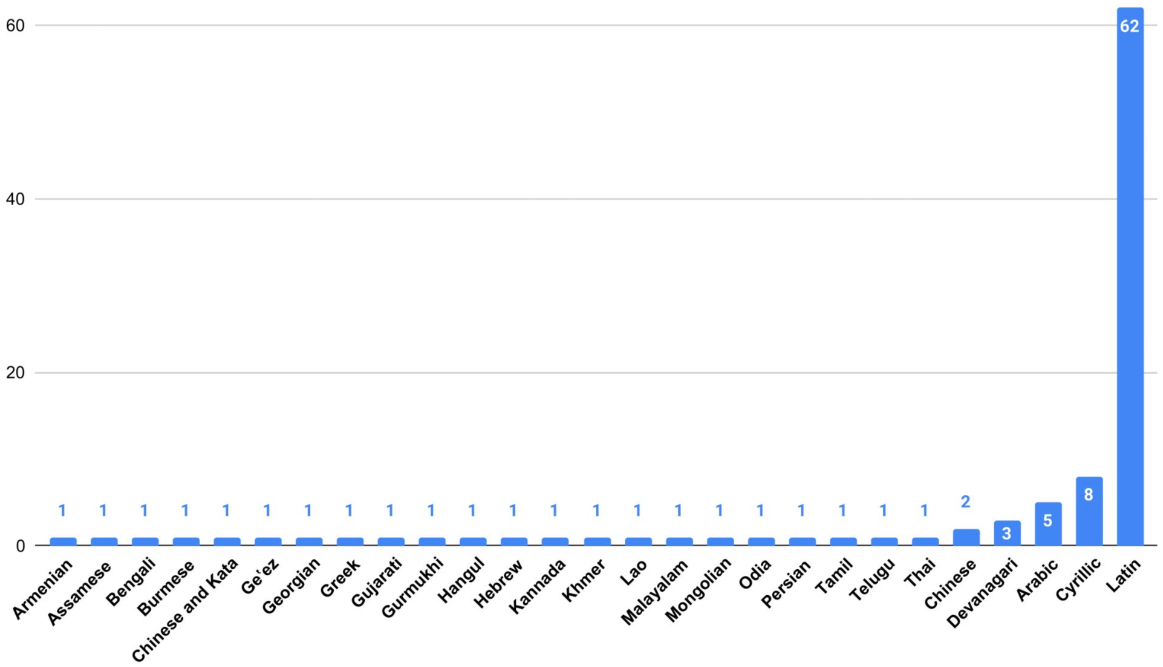
FLEURS: Parallel speech and text in 100+ languages



- Speech version of FloRes-101 benchmark
- Commonly used to evaluate automatic speech recognition and automatic speech translation in many languages across new model contributions.
- Covers a variety of language families

Conneau et. al, 2022

FLEURS: Parallel speech and text in 100+ languages



- Speech version of FloRes-101 benchmark
- Commonly used to evaluate automatic speech recognition and automatic speech translation in many languages across new model contributions.
- Covers a variety of writing systems.

Conneau et. al, 2022

Language-specific techniques

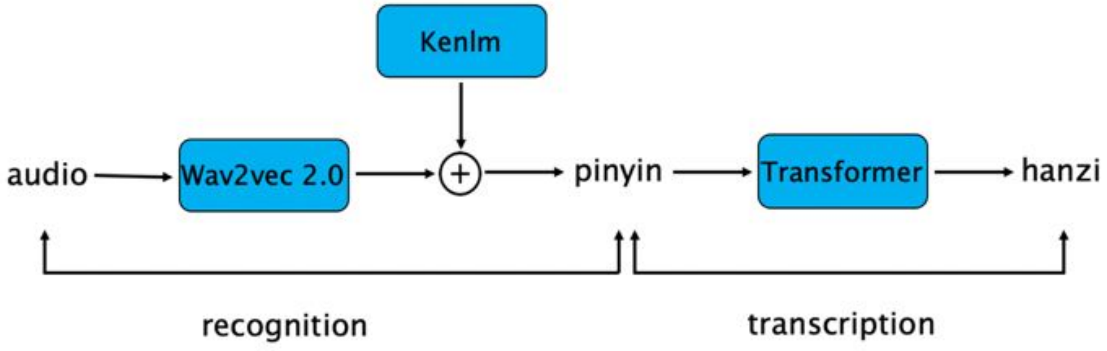
Languages can have different scripts

Writing system	Scripts	
Alphabet	Roman	napenda utambuzi wa hotuba
	Greek	Λατρεύω την αναγνώριση ομιλίας
	Cyrillic	Би яриа таних дуртай
	Korean	나는 음성 인식을 좋아해요
Semanto-Phonetic	Chinese	我喜欢语音识别
Syllabic Alphabet	Devanāgarī	मलाई बोली पहिचान मन पर्छ
	Thai	ฉันชอบการรู้จำคำพูด
	Tamil	நான் பேச்சு அங்கீகாரத்தை விரும்புகிறேன்
Abjad	Arabic	أنا أحب التعرف على الكلام
	Hebrew	אני אוהב זיהוי דיבור

Adapted from Tan et. al, 2010

Using different representations

Incorporating Pinyin for Mandarin Chinese - intermediary phonetic representation



Pinyin	nǐ hǎo
Hanzi	你好
English	Hello!

Images from Yuan et. al, 2021

Creating micro languages when you have multiple scripts per language

MMS authors treat different scripts as different languages when possible.

Serbian (Cyrillic)	Хвала
Serbian (Latin)	Hvala
English	Thank you

Languages can have lexical tone

The pitch of the word changes the meaning of the word

wá 🔊

wà 🔊

Tonal languages: can we find tones in the representations?

Shen et. al find that models behave similarly to native and non-native human participants in tone and consonant perception studies.

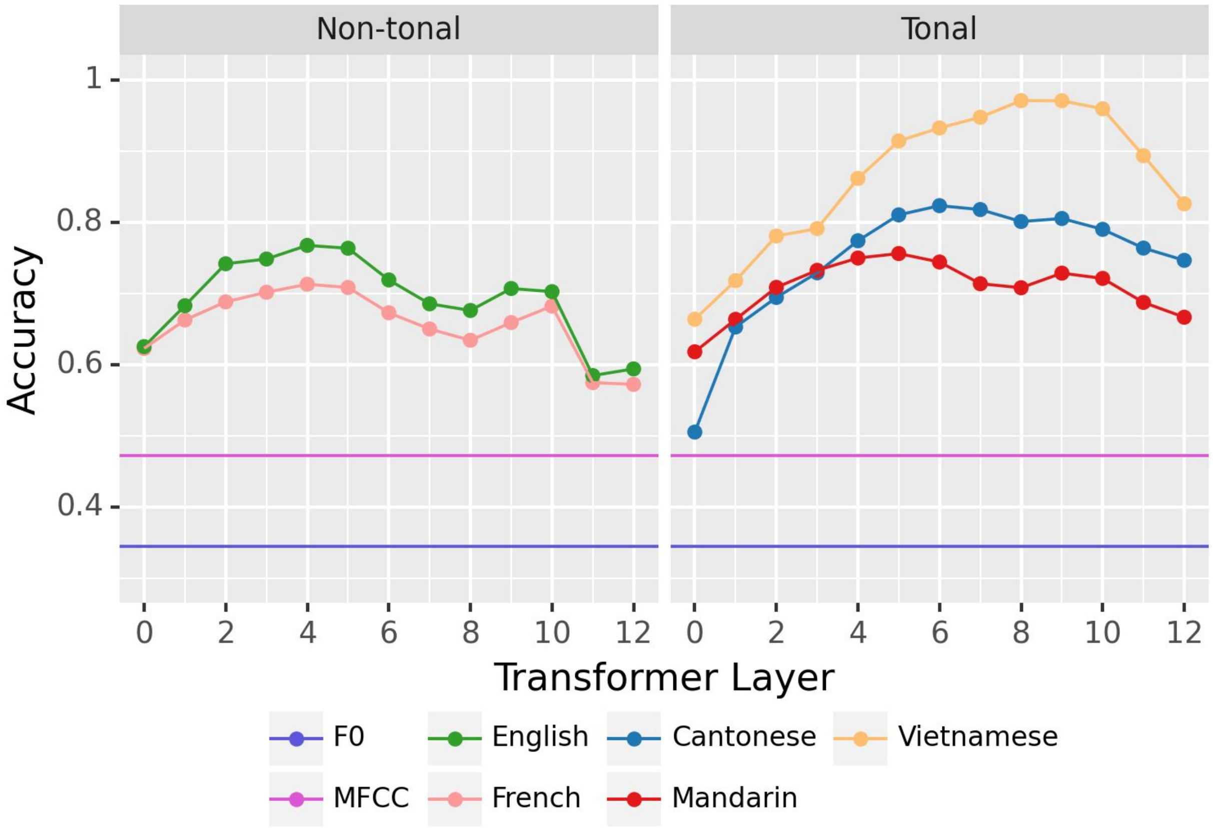
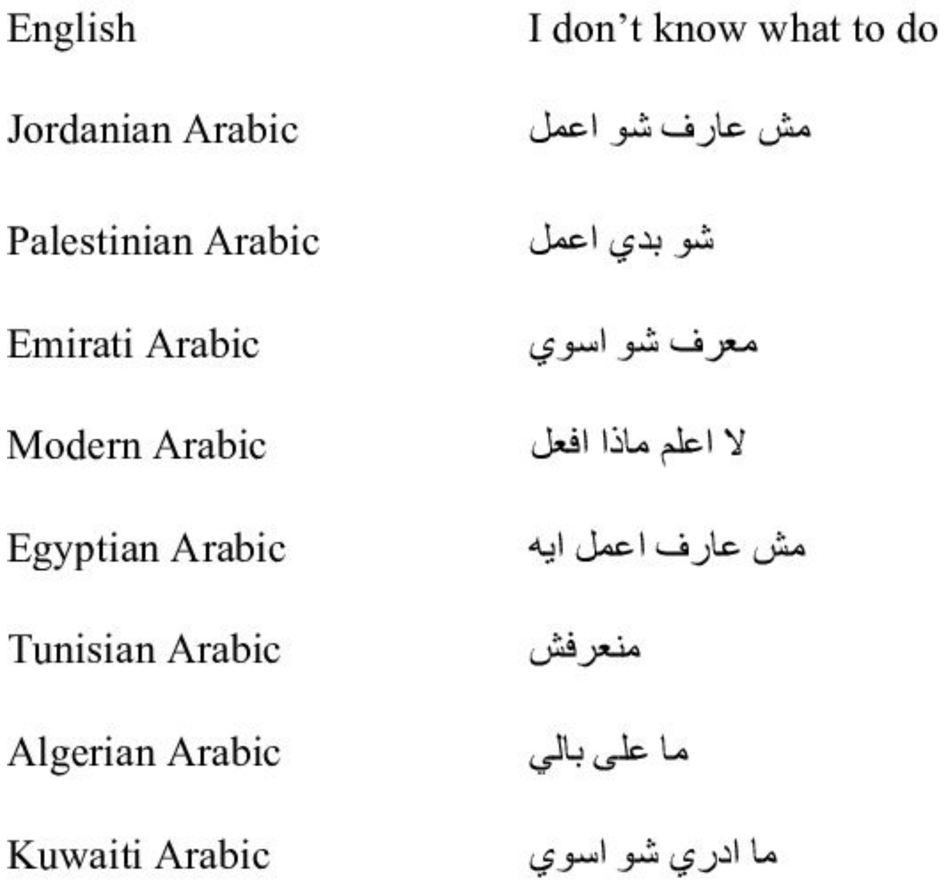


Image from Shen et. al, 2024

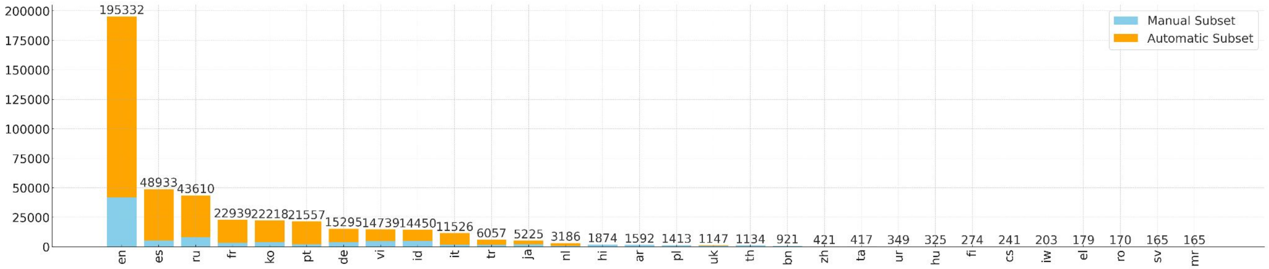
Languages can have different dialects

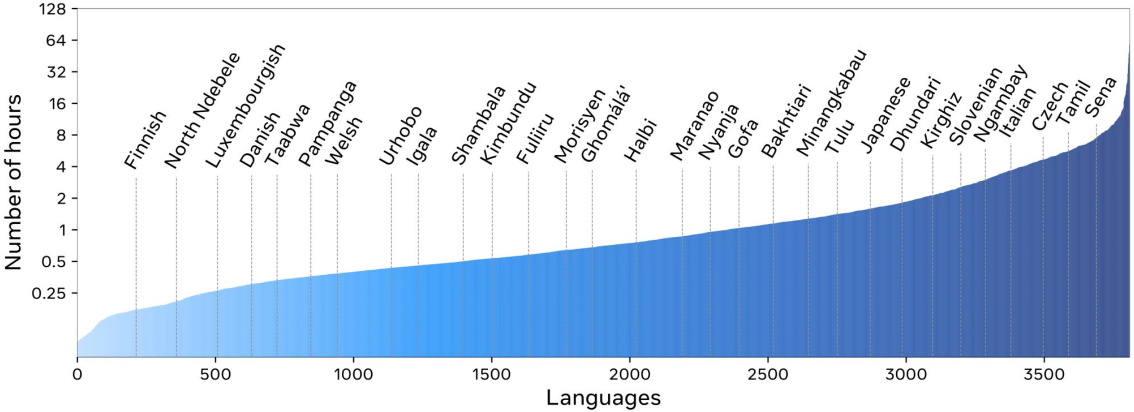


English	I don’t know what to do
Jordanian Arabic	مش عارف شو اعمل
Palestinian Arabic	شو بدي اعمل
Emirati Arabic	معرف شو اسوي
Modern Arabic	لا اعلم ماذا افعل
Egyptian Arabic	مش عارف اعمل ايه
Tunisian Arabic	منعرفش
Algerian Arabic	ما على بالي
Kuwaiti Arabic	ما ادري شو اسوي

Image from Bani-Hani et.al, 2017

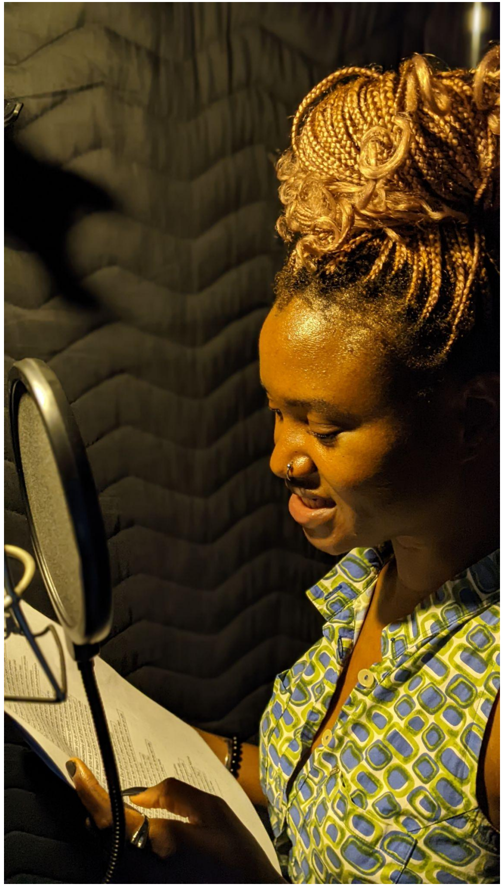
Languages can have can have little data available to train models





Making datasets

ÌròyìnSpeech: A multi-purpose Yorùbá Speech Corpus



Ògúnrèmí et. al, 2024

Making datasets

ÌròyìnSpeech: A multi-purpose Yorùbá Speech Corpus

- Multipurpose high quality speech dataset: TTS and ASR
- 42 hours of data in total
- 80 volunteers recorded utterances in a custom-made booth in Lagos, Nigeria
- One male and one female record 5 hours of TTS data

Dataset partition	Hours of data	No. of utterances
In-house ASR	26 hours	20 000
Common Voice ASR	6 hours	5 000
In-house TTS	10 hours 11 minutes	9 000

Ògúnrèmí et. al, 2024

Making datasets

ÌròyìnSpeech: A multi-purpose Yorùbá Speech Corpus

TTS:

- We train VITS models from scratch
- We also do continued pretraining of the Bible TTS model (also a VITS model) trained on only male speech

🔊 VITS from scratch: Male Voice

🔊 Bible TTS continued: Male Voice

🔊 VITS from scratch: Female Voice

🔊 Bible TTS continued: Female Voice

Ògúnrèmí et. al, 2024

Making datasets

ÌròyìnSpeech: A multi-purpose Yorùbá Speech Corpus

ASR:

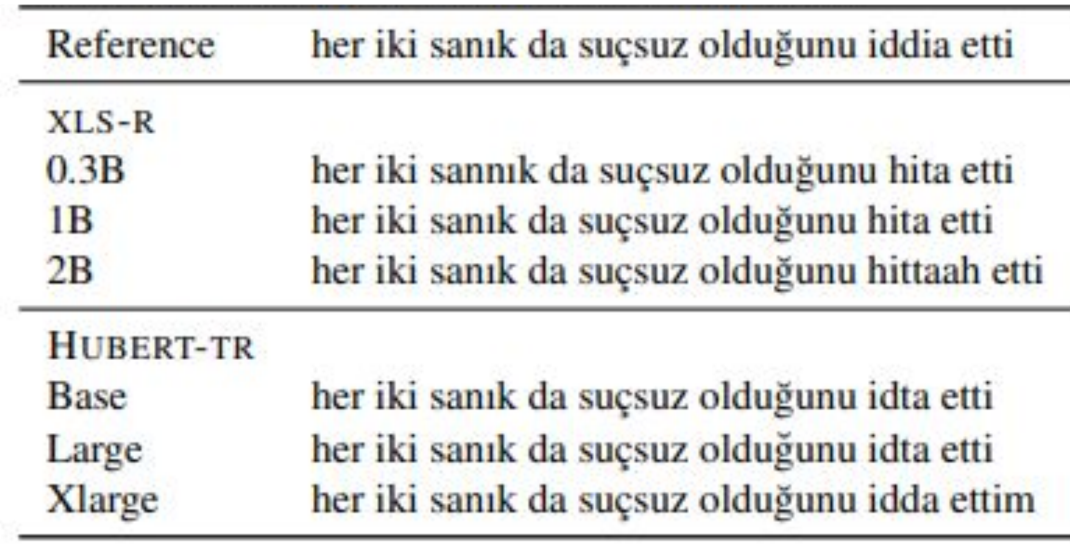
- We train a Conformer + RNN LM with ESPNet
- We also finetune wav2vec 2.0 and add n-gram language models

Model	WER
Conformer + RNN LM	69.7
wav2vec 2.0 finetuned	40.6
+bigram model	27.6
+trigram model	23.8

Ògúnrèmí et. al, 2024

Making monolingual versions of large pretrained speech models

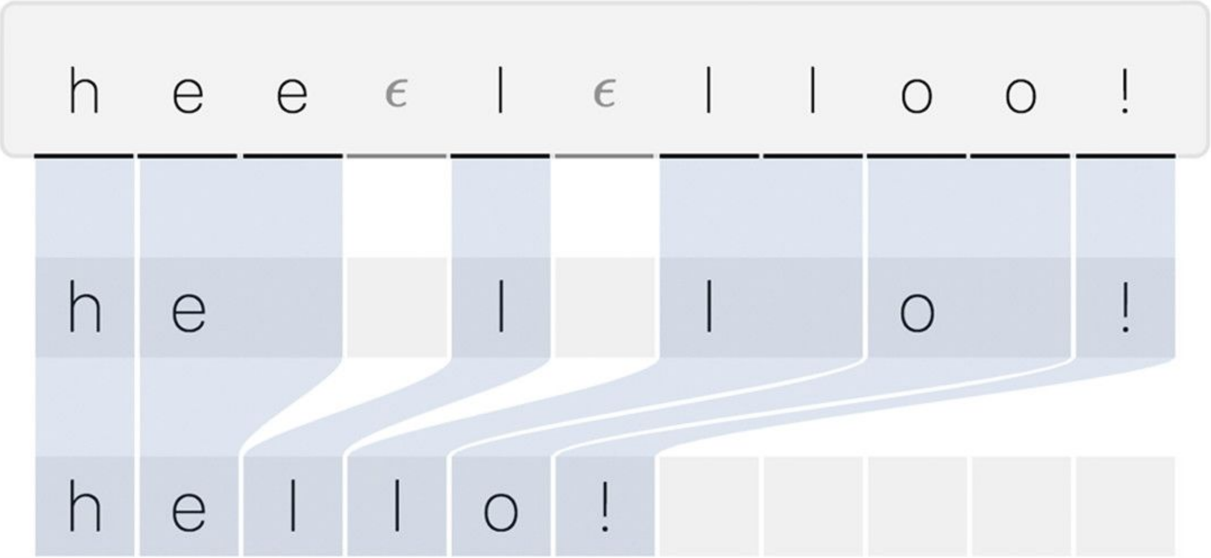
"HuBERT-TR: Reviving Turkish Automatic Speech Recognition with Self-supervised Speech Representation Learning"



Reference	her iki sanık da suçsuz olduğunu iddia etti
XLS-R	
0.3B	her iki sannık da suçsuz olduğunu hita etti
1B	her iki sanık da suçsuz olduğunu hita etti
2B	her iki sanık da suçsuz olduğunu hittaah etti
HUBERT-TR	
Base	her iki sanık da suçsuz olduğunu idta etti
Large	her iki sanık da suçsuz olduğunu idta etti
Xlarge	her iki sanık da suçsuz olduğunu idda ettim

Safaya and Erzin, 2022

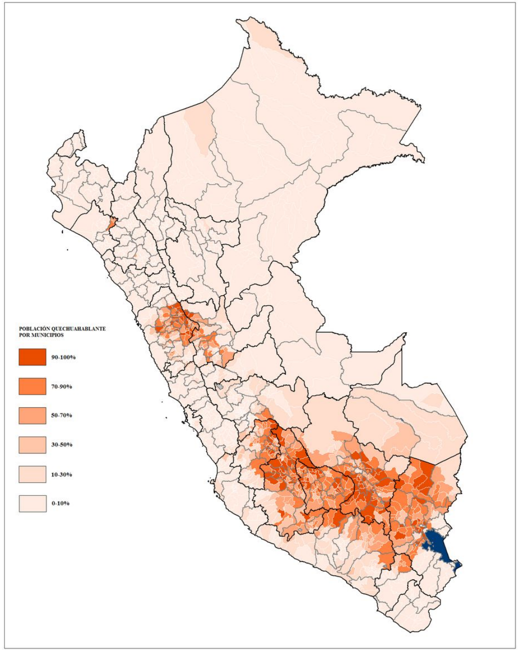
Different scripts leverage CTC - no need for huge language model



Low resource language example: Quechua

Quechua

EL QUECHUA COMO LENGUA MATERNA
(CENSO NACIONAL 2007)



- Language spoken in Peru, Ecuador, Bolivia, Argentina, Colombia and Chile
- Roughly 10 million speakers
- There are many language varieties/dialects, some more popular than others.
- Quechua is an agglutinating language
- Written using the Roman alphabet
- Spanish words are borrowed in Quechua

Improving Quechua ASR

Running example utterance:

Ground truth: Ima ninantaq awkaypata

English translation: What does this mean for evil?

Source: Bible verse

Current performance across models

	WER (%)↓
Whisper OOTB* (no forced decoder IDs)	326.34
Whisper OOTB* (with forced decoder IDs)	117.36
Whisper Fine-tuned (Spanish)	19.77
Whisper Fine-tuned (Japanese)	**17.79**
MMS OOTB* (Best quechua adapter)	49.24
MMS OOTB* (Worst quechua adapter)	84.43
MMS Fine-tuned (Reinitializing all adapter layers)	**31.9**

***OOTB = out-of-the-box. This will be getting the model to transcribe without any intervention.**

Quechua performance before finetuning: Whisper

	Transcription	Explanation / Translation	Severity
Ground truth	ima ninantaq awkaypata	*What does this mean for evil?*	-
Whisper OOTB (no forced decoder IDs)	imaninantag, hau keipata.	*What does this mean…* (rest is unintelligible)	*Very high*
Whisper OOTB (with forced decoder IDs - ja) (ouput is Japanese Katakana which is transliterated to Quechua)	いまになんたく、アウカイパタン (imaninantaku aukaypatan)	*What does this mean for his/her evil?* -n suffix: applies a third person singular possessive to the noun "evil" Ku: Added because japanese cannot end in consonants except -n	*Medium*

***OOTB = out-of-the-box. This will be getting the model to transcribe without any intervention.**

Quechua performance before finetuning: MMS

	Transcription	Explanation / Translation	Severity
Ground truth	ima ninantaq awkaypata	*What does this mean for evil?*	-
MMS OOTB * (Best quechua adapter - quy)	imaninantaq hawkaypata	*What does this mean for traditions?* Awkay vs hawkay are similar phonetically but very different semantically	*High*
MMS OOTB * (Worst quechua adapter - qvo)	ima ninanta* jaucaipata	*What this means for traditions* *absence of -q suffix used for questions, the sentence becomes declarative Jaucaipata sounds like hawkaypata in SPanish	*High*

***OOTB = out-of-the-box. This will be getting the model to transcribe without any intervention.**

Quechua performance after finetuning

	Transcription	Explanation / Translation	Severity
Ground truth	ima ninantaq awkaypata	*What does this mean for evil?*	-
Whisper Fine-tuned (Spanish)	ima ninantaq awkaypata	*What does this mean for evil?*	*None*
Whisper Fine-tuned (Japanese)	ima ninantaq awkaypata	*What does this mean for evil?*	*None*
MMS Fine-tuned	ima ninantaq awccaypata	*What does this mean for evil?*	*Low* *Alternative regional pronunciation

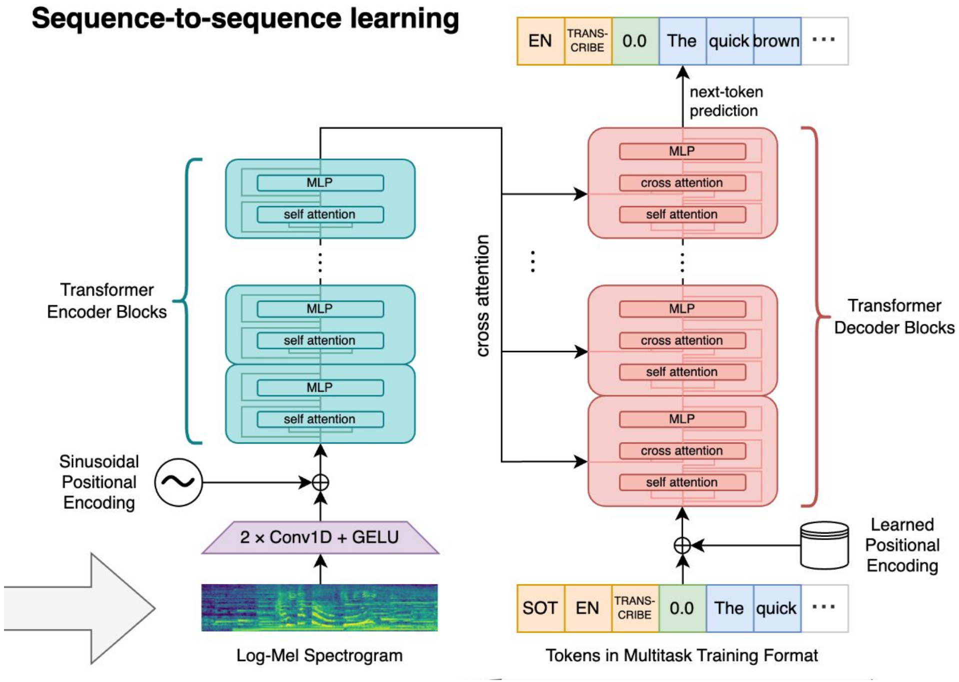
Language is varied

wá

wà

napenda utambuzi wa hotuba
Λατρεύω την αναγνώριση ομιλίας
나는 음성 인식을 좋아해요
मलाई बोली पहिचान मन पर्छ
ฉันชอบการรู้จำคำพูด
நான் பேச்சு அங்கீகாரத்தை விரும்புகிறேன்
أنا أحب التعرف على الكلام

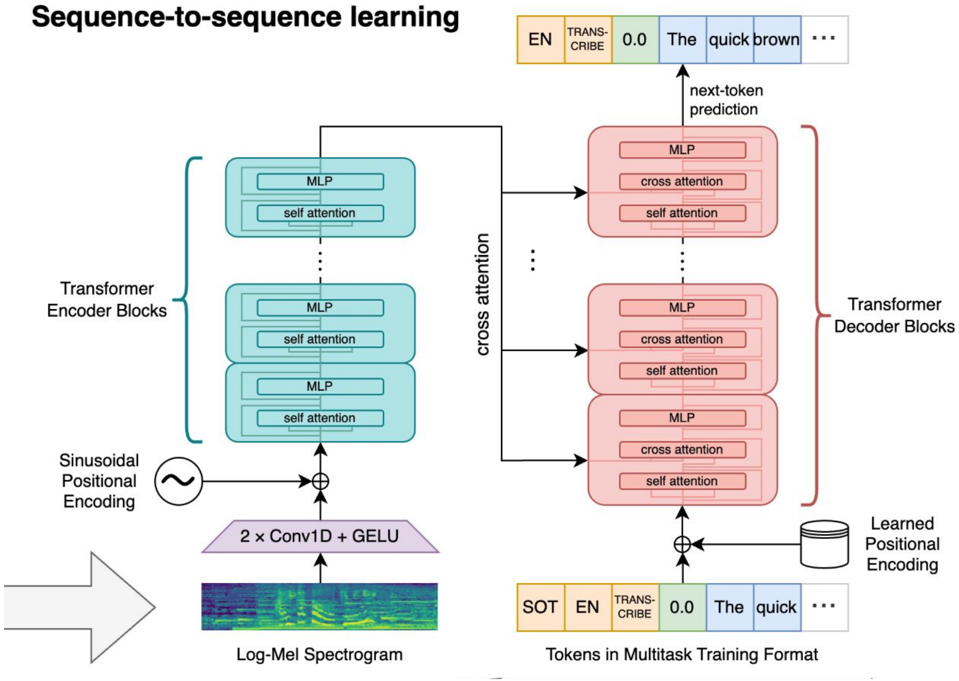
Surprisingly, all you need to do is chuck a bunch of data into a model.



Sequence-to-sequence learning
EN
TRANS-CRIBE
0.0
The
quick
brown
···
next-token prediction
MLP
cross attention
self attention
Transformer Encoder Blocks
Transformer Decoder Blocks
cross attention
Sinusoidal Positional Encoding
2 × Conv1D + GELU
Log-Mel Spectrogram
Learned Positional Encoding
SOT
EN
TRANS-CRIBE
0.0
The
quick
···
Tokens in Multitask Training Format

Surprisingly, all you need to do is chuck a bunch of data into a model.

And finetune it with reliable labelled data.


Sequence-to-sequence learning
EN
TRANS-CRIBE
0.0
The
quick
brown
···
next-token prediction
MLP
cross attention
self attention
Transformer Encoder Blocks
Transformer Decoder Blocks
cross attention
Sinusoidal Positional Encoding
2 × Conv1D + GELU
Log-Mel Spectrogram
Learned Positional Encoding
SOT
EN
TRANS-CRIBE
0.0
The
quick
···
Tokens in Multitask Training Format

How to train an ASR model for a low-resource language

How to train an ASR model for a new language

Things to consider:

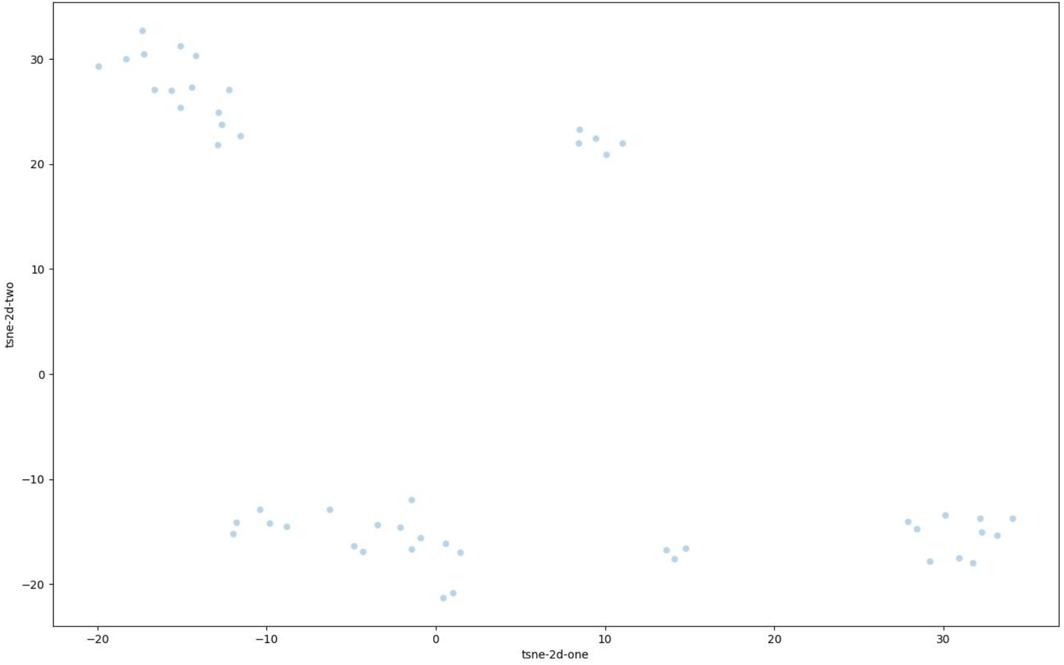
- Amount of supervised training data
- Amount of unsupervised training data
- Compute budget
- Language coverage in existing multilingual speech models

	Low compute (can't finetune all weights)	Medium compute (can finetune all weights)	High compute (can pretrain model from scratch)
Limited supervised data (less than 100 hours)	Train an MMS adapter	Try finetuning any large speech model	Try doing continued pretraining of a large speech model
Modest supervised data (1000+ hours)		Finetune a large multilingual speech model	Do continued pretraining of a large speech model
Large amounts of unsupervised data (3,000+ hours) and any amount of supervised data.		Try finetuning or continued pretraining of selected layers	Try training your favourite architecture from scratch

Homework 4

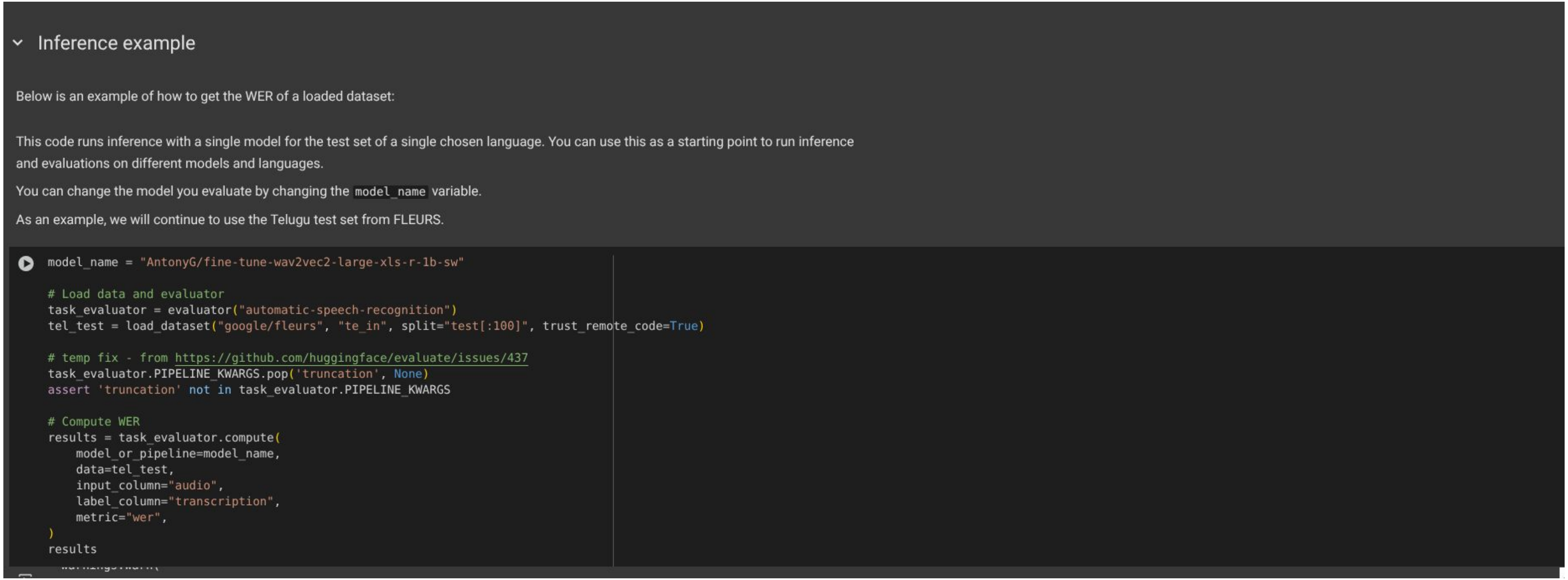
Part 1: Visualising representations of large pretrained self-supervised speech models

Part 1: Visualising representations of large pretrained self-supervised speech models



Part 2: Non-English ASR

Part 2: Non-English ASR



Inference example

Below is an example of how to get the WER of a loaded dataset:

This code runs inference with a single model for the test set of a single chosen language. You can use this as a starting point to run inference and evaluations on different models and languages.

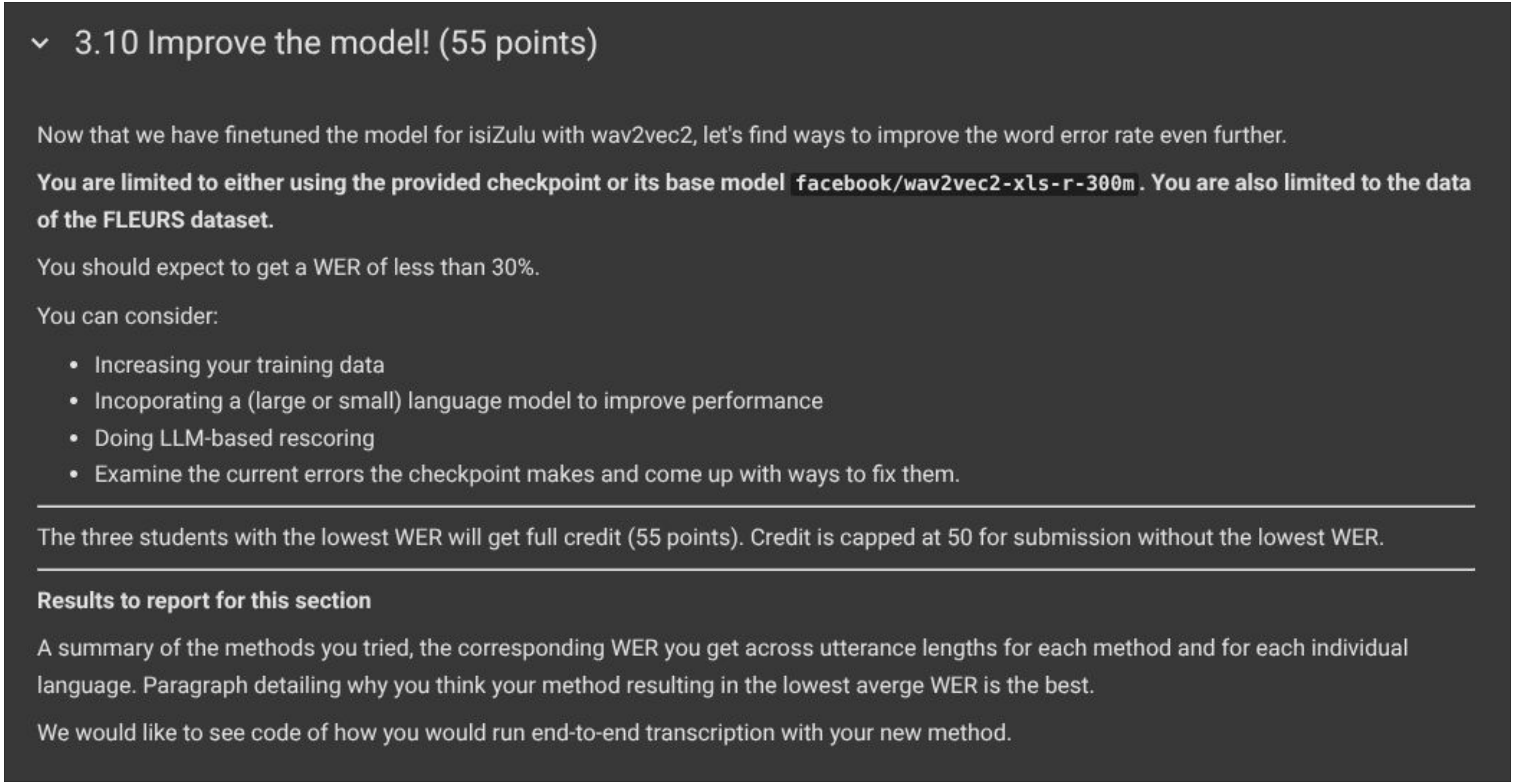
You can change the model you evaluate by changing the `model_name` variable.

As an example, we will continue to use the Telugu test set from FLEURS.

```
model_name = "AntonyG/fine-tune-wav2vec2-large-xls-r-1b-sw"

# Load data and evaluator
task_evaluator = evaluator("automatic-speech-recognition")
tel_test = load_dataset("google/fleurs", "te_in", split="test[:100]", trust_remote_code=True)

# temp fix - from https://github.com/huggingface/evaluate/issues/437
task_evaluator.PIPELINE_KWARGS.pop('truncation', None)
assert 'truncation' not in task_evaluator.PIPELINE_KWARGS

# Compute WER
results = task_evaluator.compute(
    model_or_pipeline=model_name,
    data=tel_test,
    input_column="audio",
    label_column="transcription",
    metric="wer",
)
results
```

Part 3: Fine tuning wav2vec2 with isiZulu FLEURS data

Part 3: Fine tuning wav2vec2 with isiZulu FLEURS data



3.10 Improve the model! (55 points)

Now that we have finetuned the model for isiZulu with wav2vec2, let's find ways to improve the word error rate even further.

You are limited to either using the provided checkpoint or its base model `facebook/wav2vec2-xls-r-300m`. You are also limited to the data of the FLEURS dataset.

You should expect to get a WER of less than 30%.

You can consider:

- Increasing your training data
- Incoporating a (large or small) language model to improve performance
- Doing LLM-based rescoring
- Examine the current errors the checkpoint makes and come up with ways to fix them.

The three students with the lowest WER will get full credit (55 points). Credit is capped at 50 for submission without the lowest WER.

Results to report for this section

A summary of the methods you tried, the corresponding WER you get across utterance lengths for each method and for each individual language. Paragraph detailing why you think your method resulting in the lowest averge WER is the best.

We would like to see code of how you would run end-to-end transcription with your new method.

Thank You