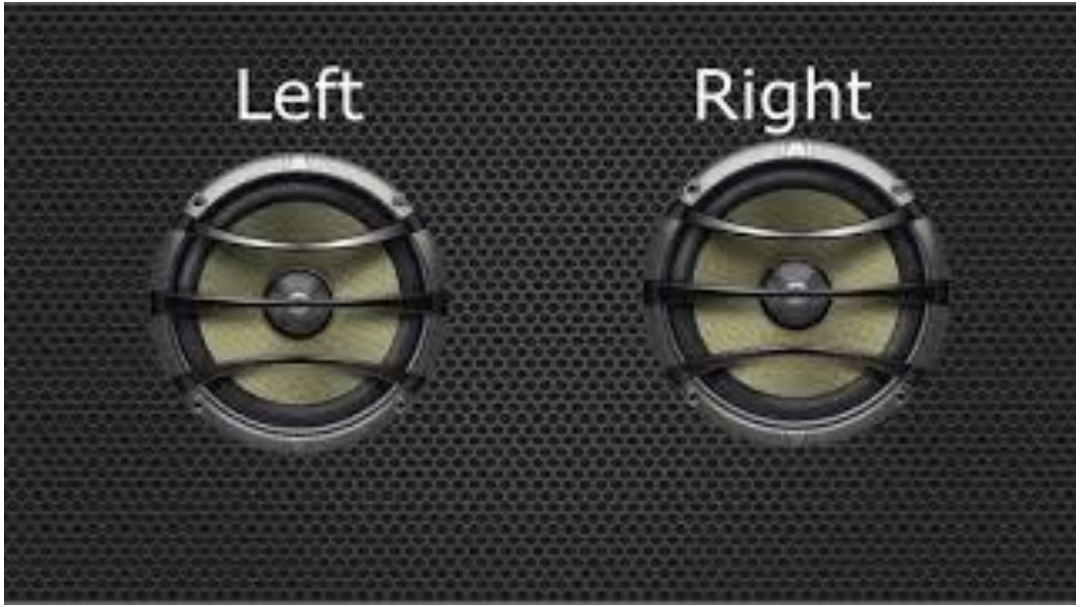
CS 224S / Linguist 285
Spoken Language Processing

Tolúlọpẹ́ Ògúnrẹ̀mí | Stanford University | Spring 2024

Lecture 12: Speech Recognition Beyond English

Sound check



Project Check-Ins

Outline

- How languages can differ from English
- Multilingual large pre-trained models
- Datasets and Benchmarks
- Language-specific ASR techniques

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

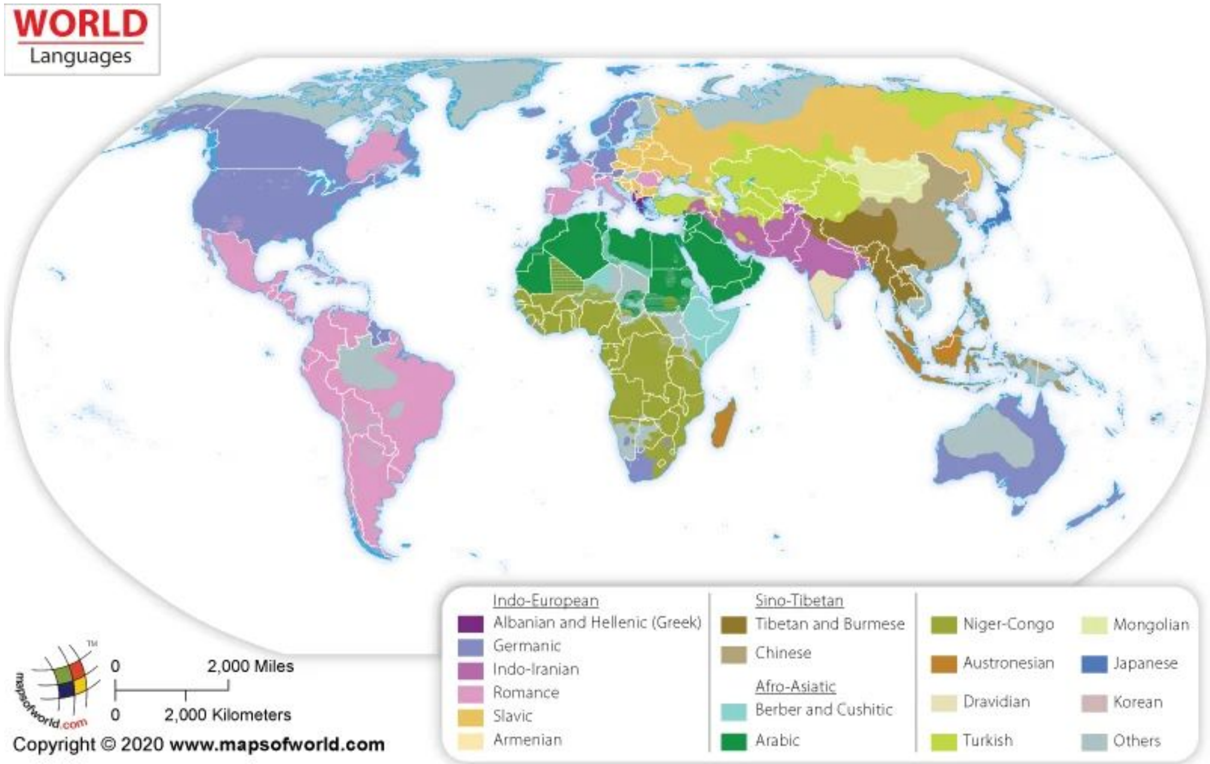
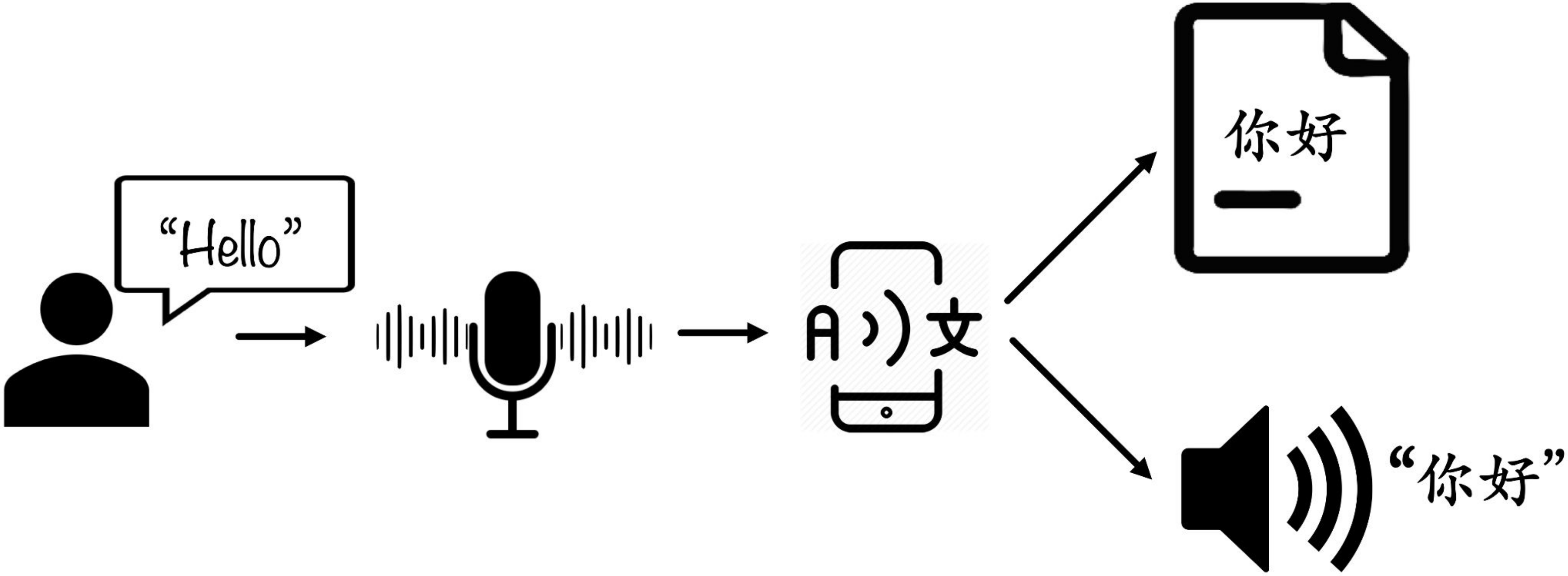


Image from mapsoftheworld.com

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

wọ

wà

wɔ̀

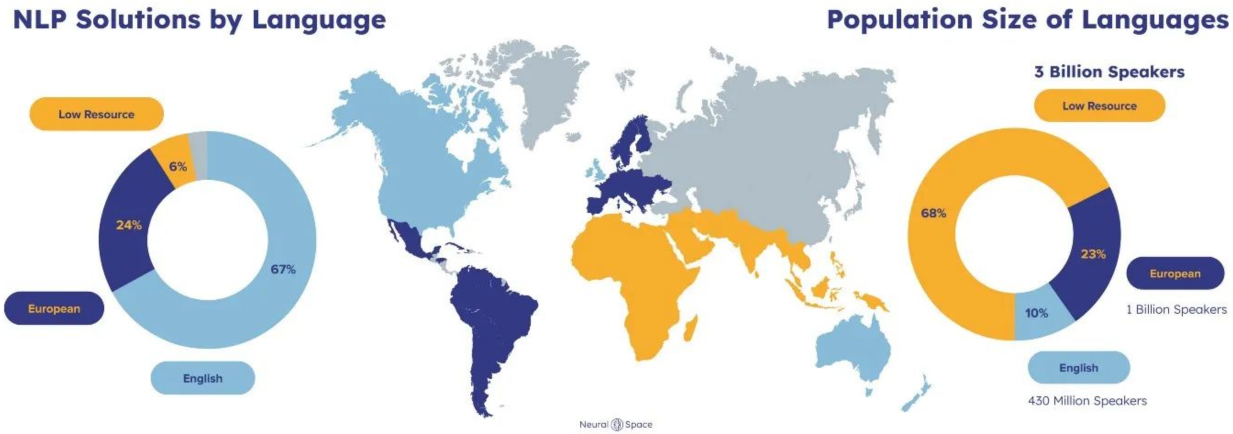
Languages can have different dialects

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

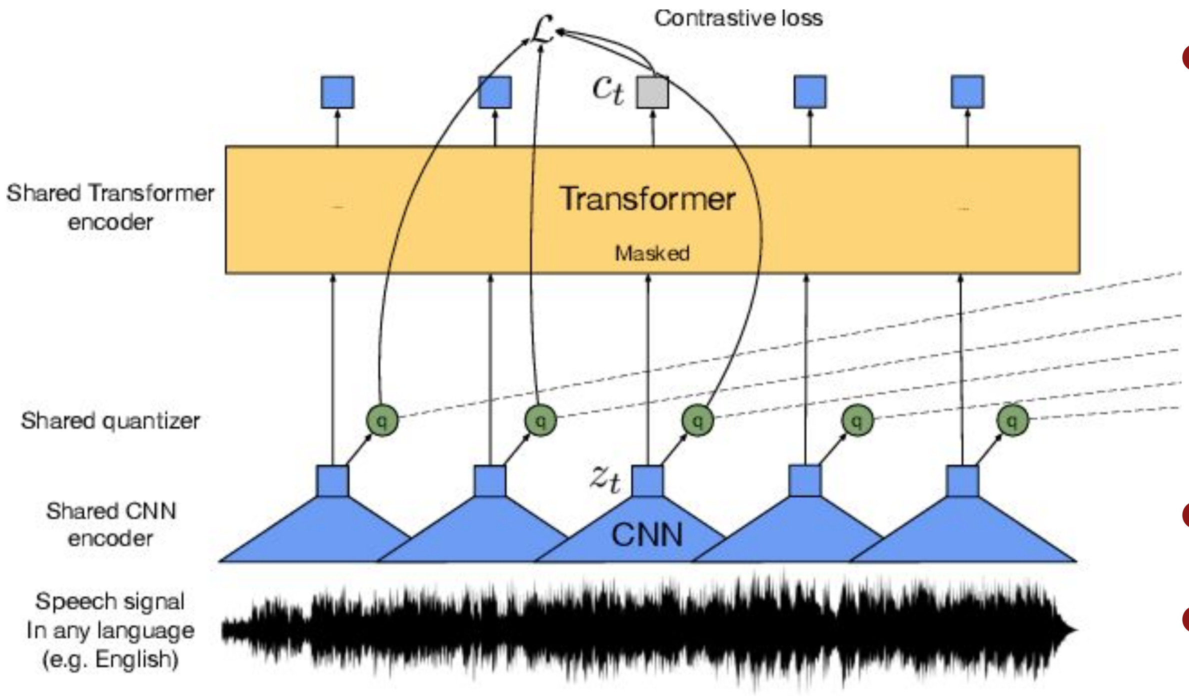


Image from Conneau et.al, 2020

- 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

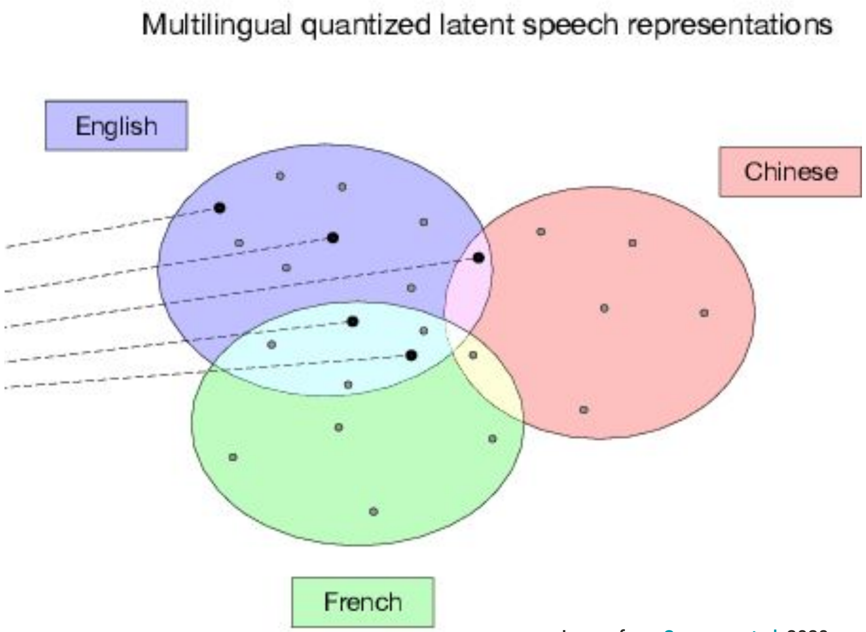


Image from [Conneau et.al](), 2020

- Latent multilingual speech representations are theorised

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

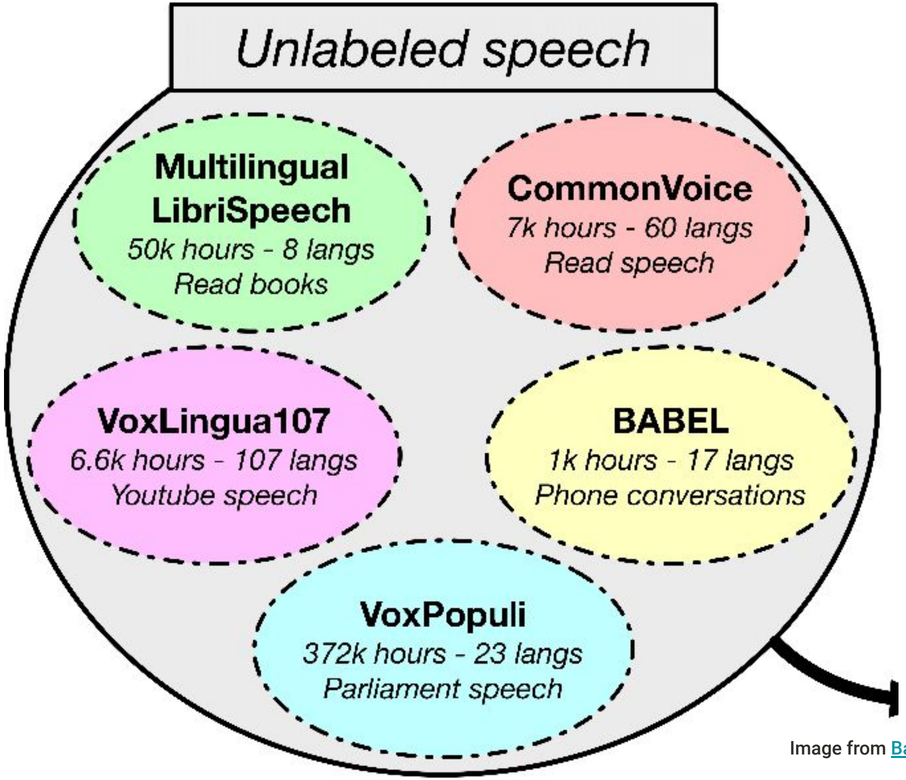


Image from Babu et. al, 2021

- 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

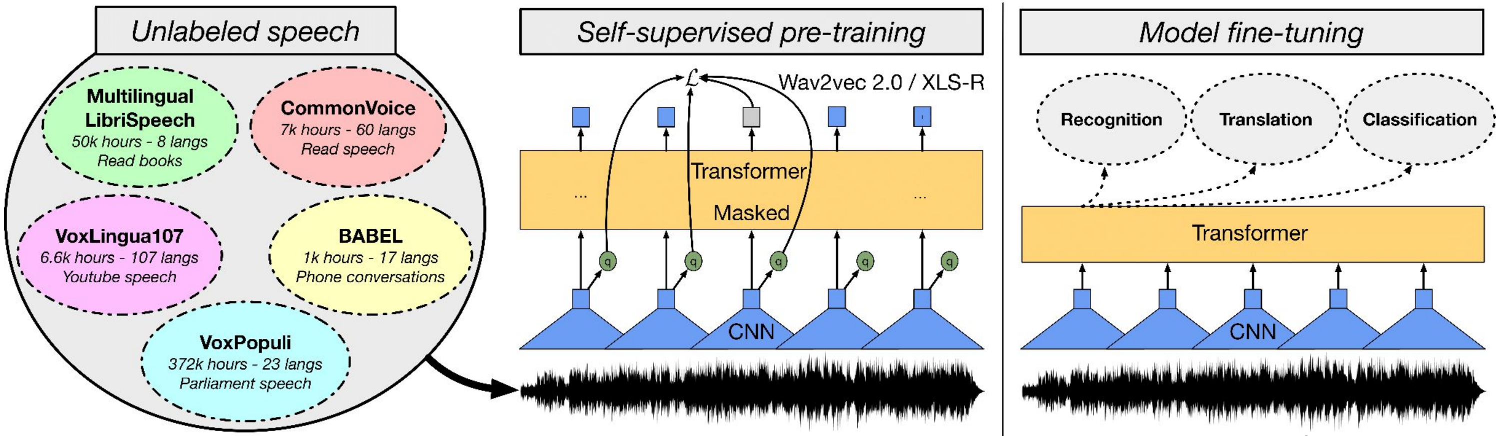


Image from Babu et. al, 2021

- Tested on ASR and AST (Automatic Speech Translation)

Multilingual versions of English-only models: mHuBERT

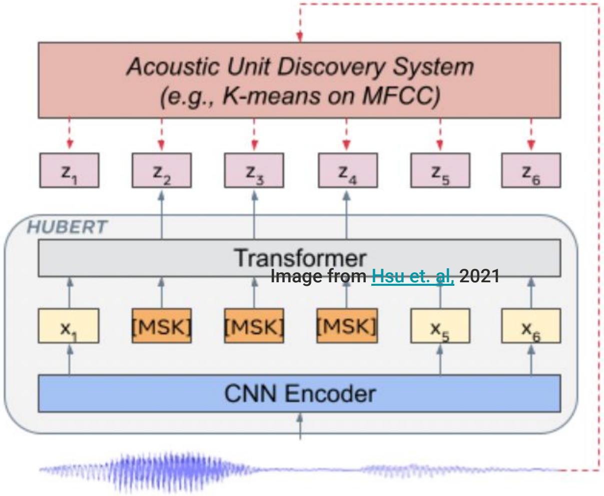


Image from Hsu et. al, 2021

- Trained specifically for speech translation in "Textless Speech-to-Speech Translation on Real Data" (Lee et. al, 2022)
- Trained with the 100,000 hour subset of Vox Populi

Multilingual from the start: Whisper

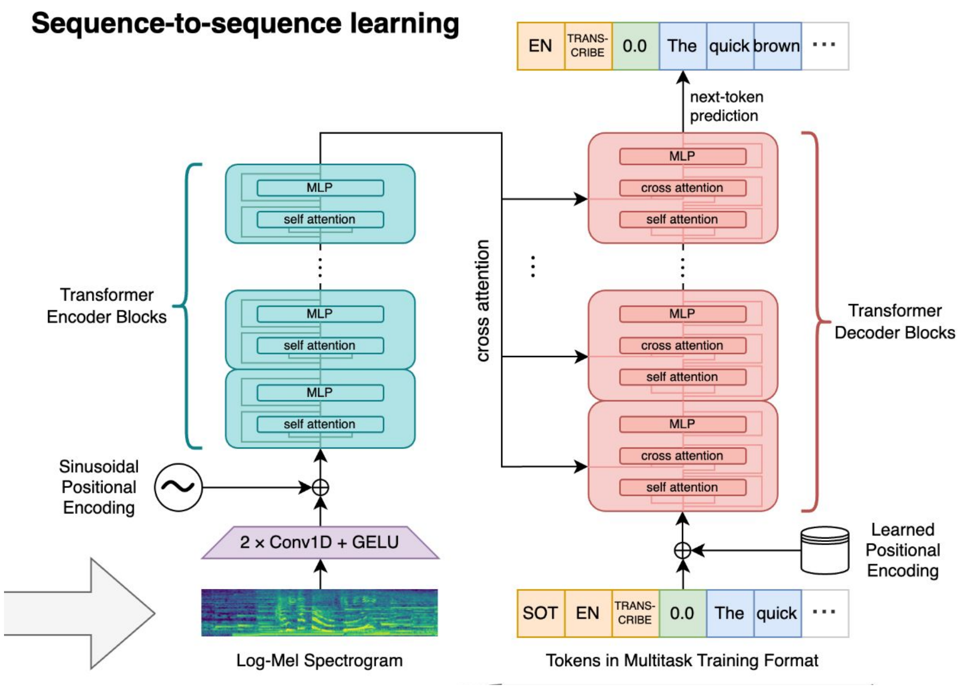


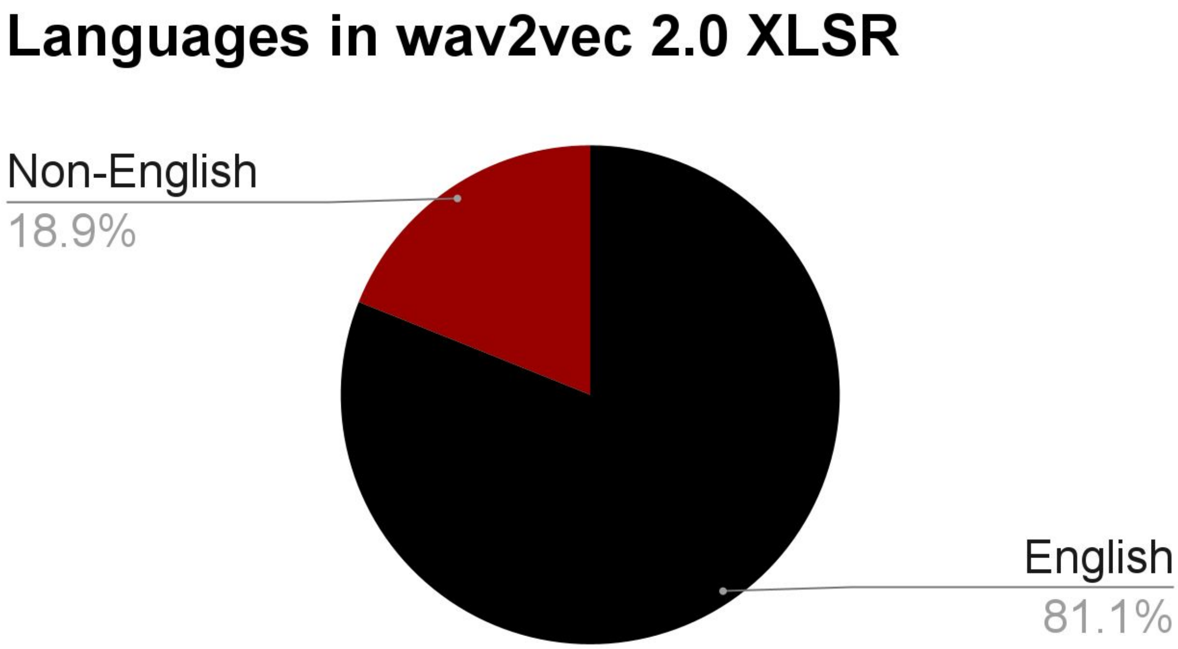
Image from [Radford et. al](), 2022

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

What is the data distribution?

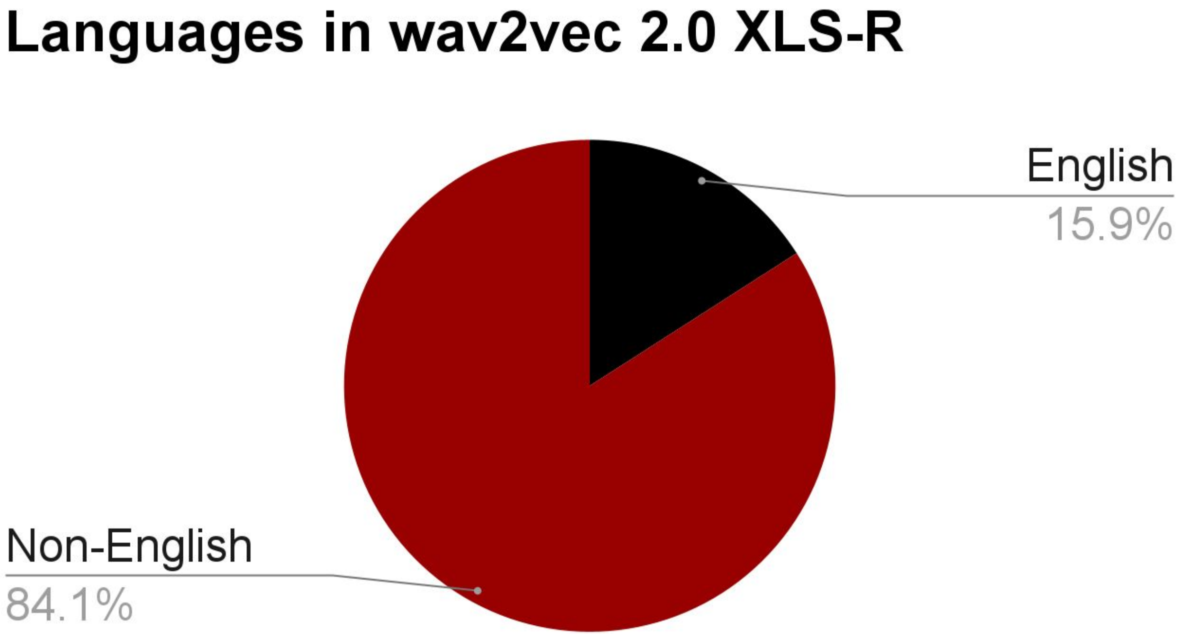
How multilingual are these models?

wav2vec 2.0 XLSR



How multilingual are theses models?

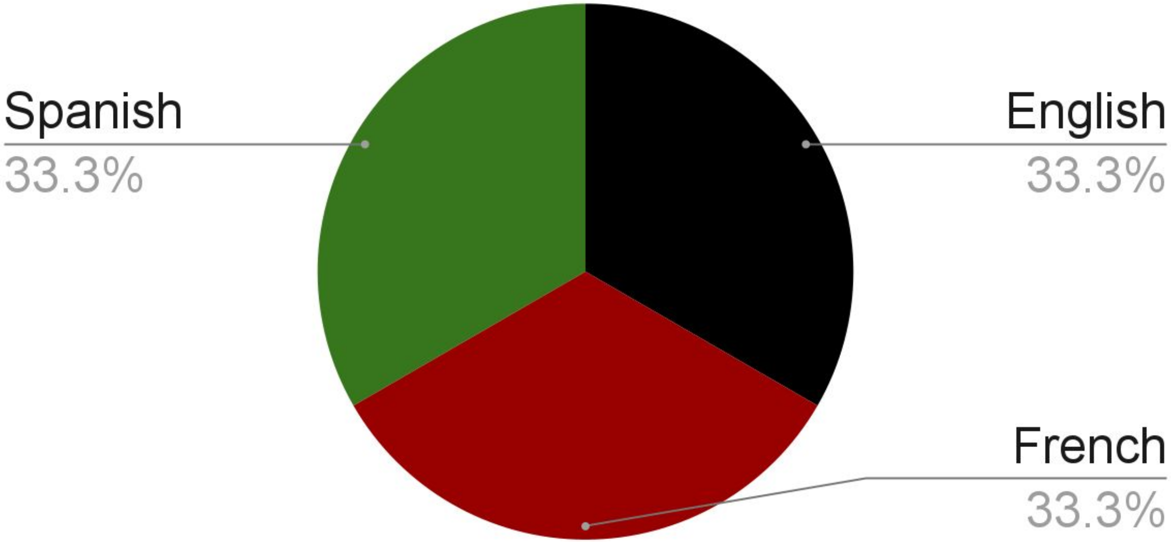
wav2vec 2.0
XLS-R



How multilingual are these models?

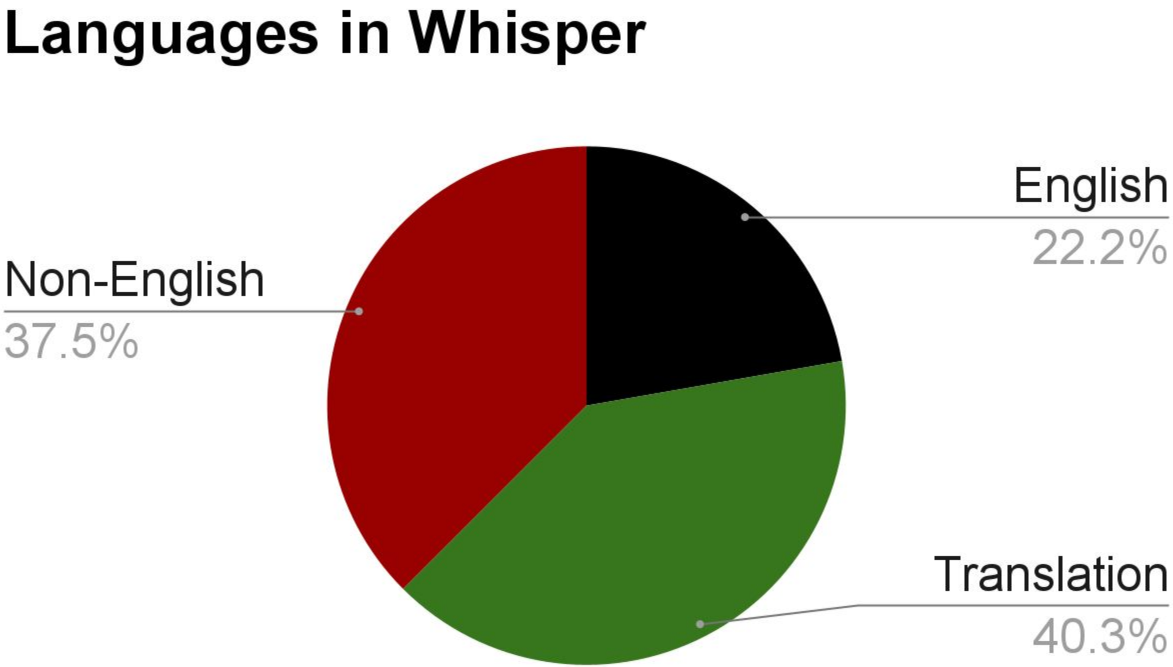
mHuBERT

Languages in mHuBERT



How multilingual are these models?

Whisper



Open Source Multilingual Datasets: CommonVoice and Yodas

Common Voice

Language	Hours	Speakers	Validation Progress	Sentences	
English	3485	93140	73%	1673433	CONTRIBUTE
Catalan	3838	35957	73%	1214229	COL·LABOREU-HI
Kinyarwanda	2388	1134	84%	1404853	FASHA, TANGA UMUSANZU
Belarusian	1794	8363	97%	379505	ЗРАБІЦЬ УНЁСАК
Esperanto	1936	1758	75%	180562	KONTRIBUI
German	1412	19151	94%	2056443	MITMACHEN
French	1145	19584	88%	1646292	CONTRIBUER
Kabyle	699	1560	81%	182716	TTEKKI
Spanish	2326	26107	20%	1080695	COLABORAR
Luganda	583	660	76%	191407	YAMBAKO
Swahili	1081	1459	37%	134669	CHANGIA
Persian	426	4453	87%	55667	مشارکت

- 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

Common Voice

Marathi Common Voice

Kinyarwanda Common Voice

Attempting to open source datasets: Yodas

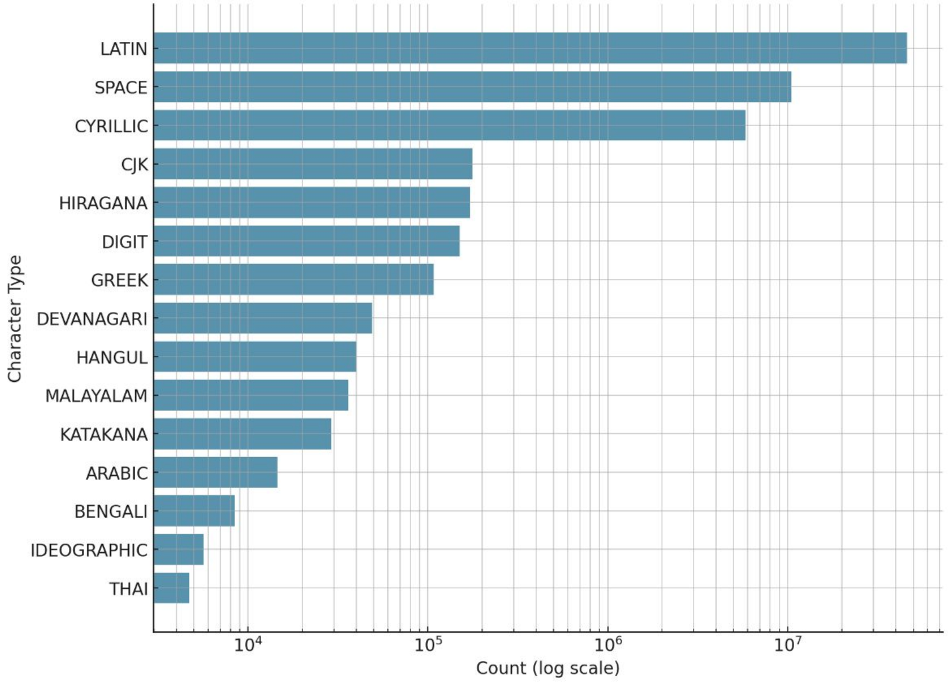


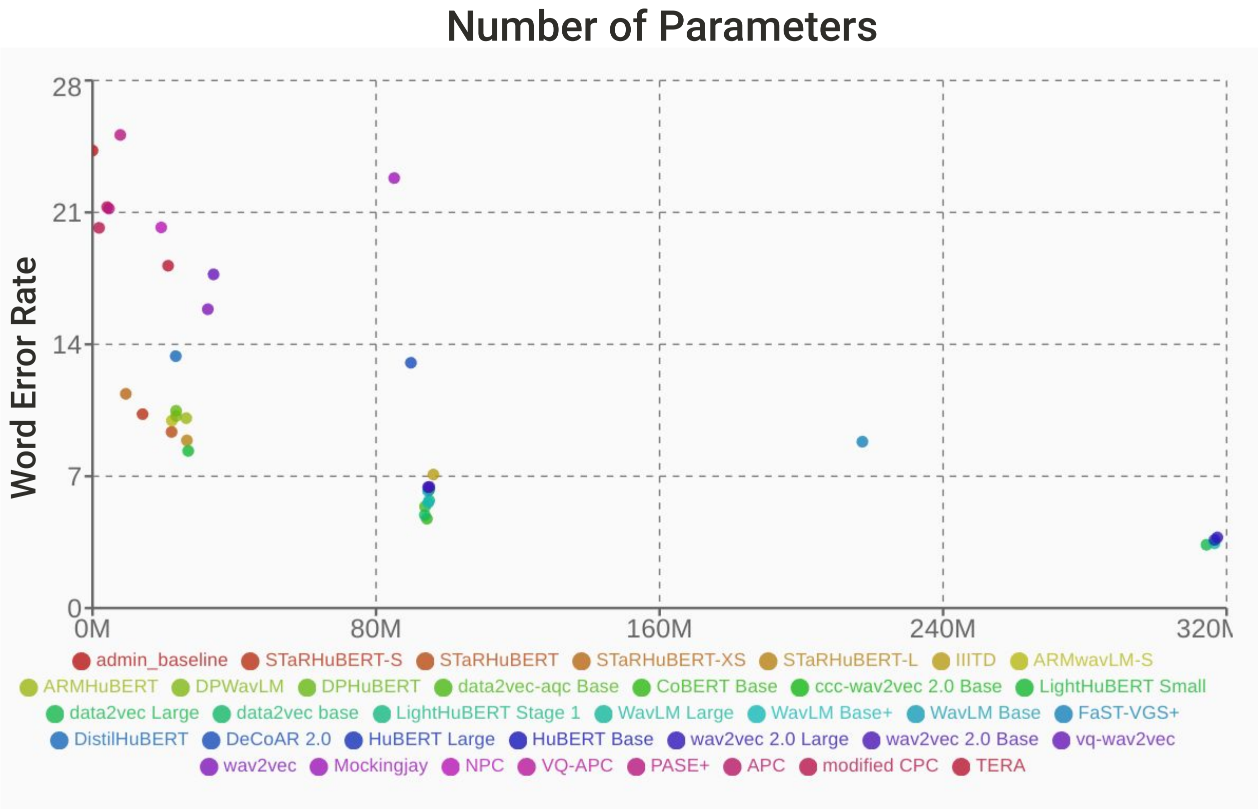
Image from [WAVLab post]

- [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

Benchmarking large models multilingually: ML-SUPERB

SUPERB

Speech processing Universal PERformance Benchmark



ML-SUPERB

Multilingual Speech processing Universal PERformance Benchmark

Automatic speech recognition and language identification for 143 languages

Method	Mono-ASR ↓	Multi-ASR (Normal) ↓	Multi-ASR (Few-shot) ↓	LID ↑
HuBERT Base	35.3	31.4	42.7	86.1
HuBERT Large	32.2	37.7	43.5	64.1
Mandarin HuBER...	45.6	43.2	46.6	85.3
Mandarin HuBER...	33.7	39.6	45.1	57.3
Robust wav2vec 2...	35.7	31.1	42.2	72.1

Language-specific techniques

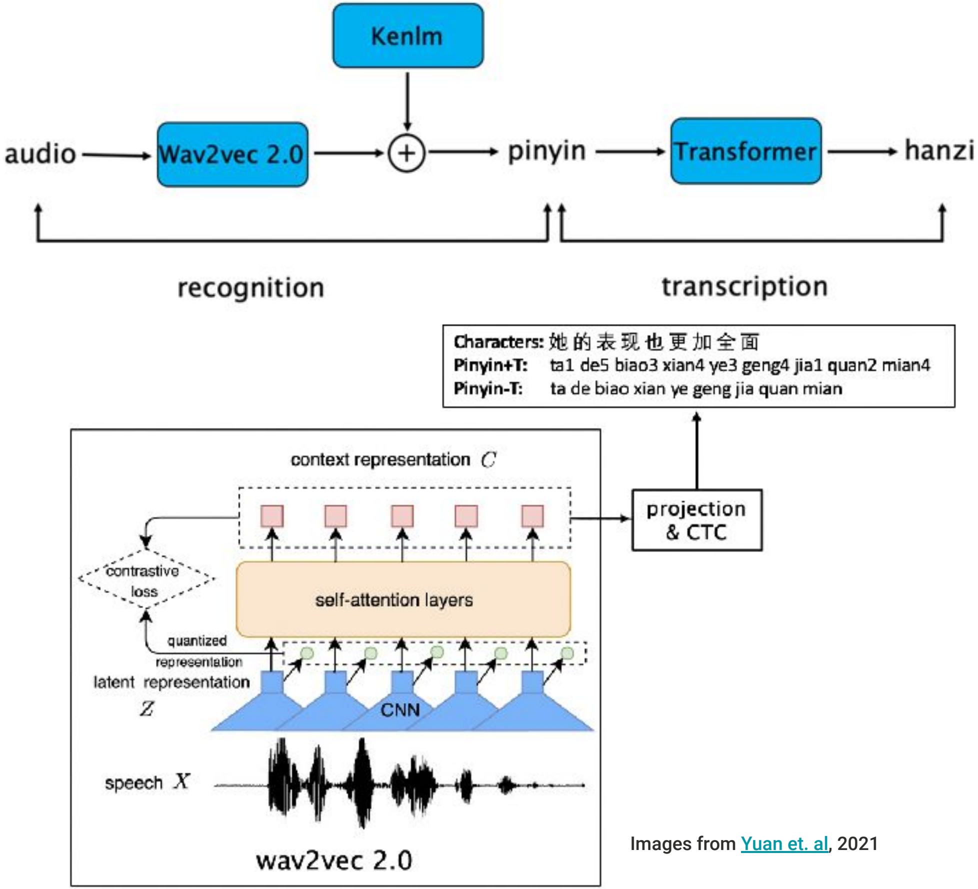
Languages can have different scripts

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



Images from [Yuan et. al](), 2021

Languages can have lexical tone

The pitch of the word changes the meaning of the word

wá

wọ

wà

wɔ̀

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, but they do not follow the same developmental trajectory.

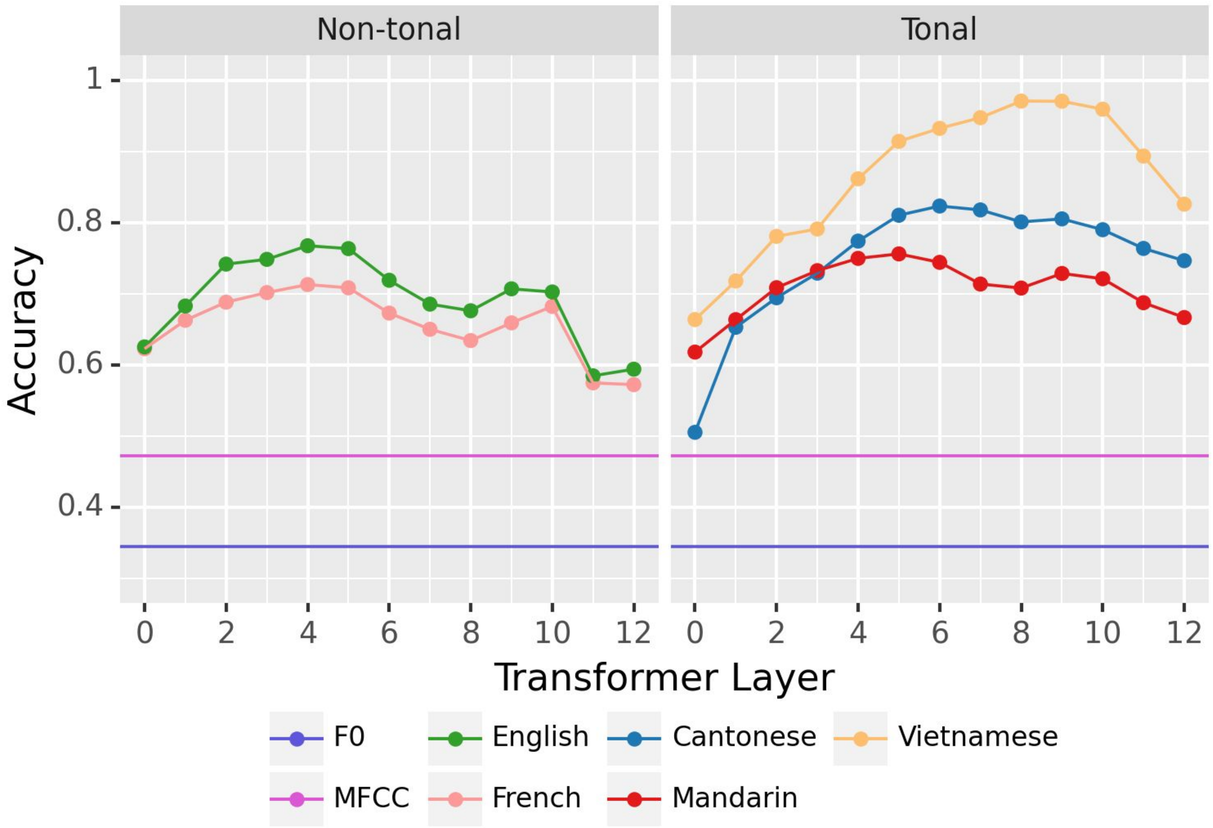


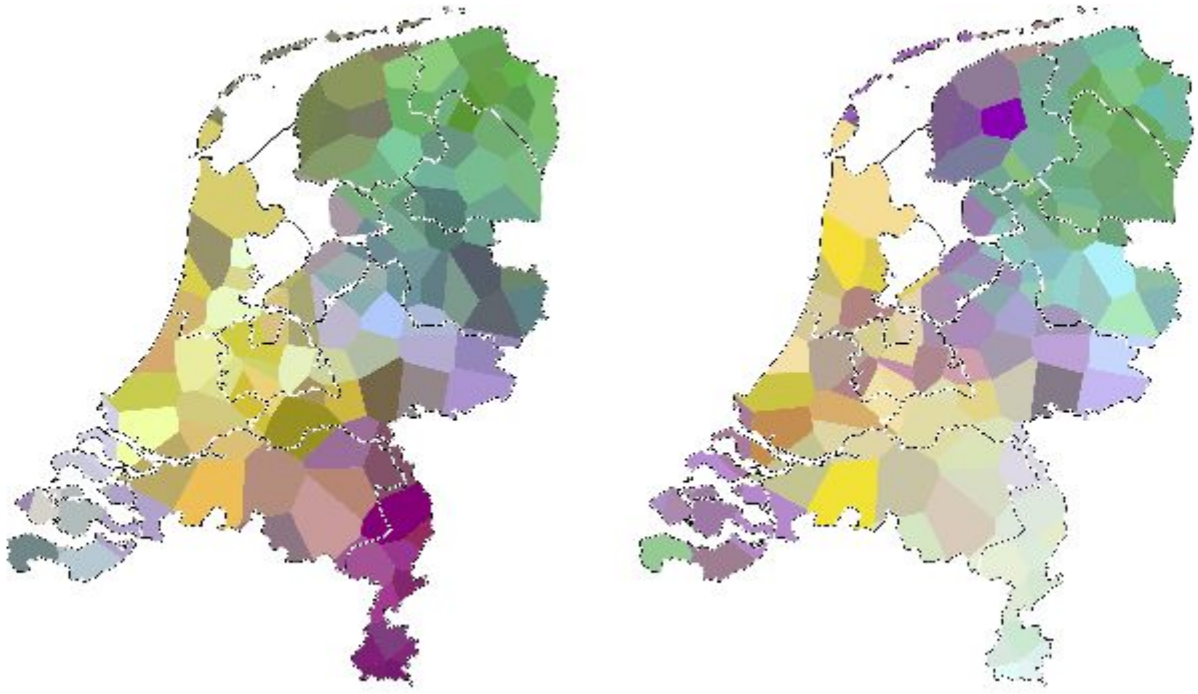
Image from Shen et. al, 2024

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

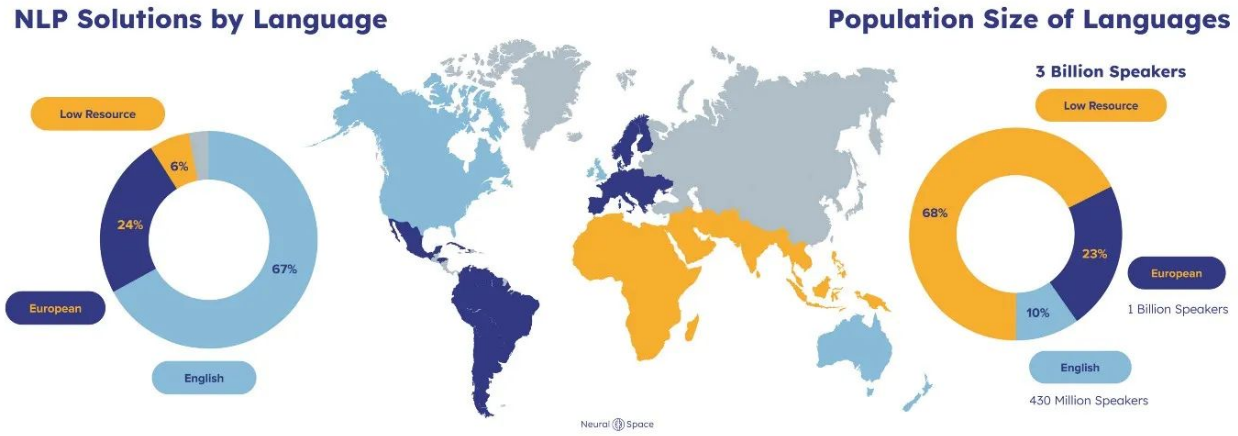
Next week!



(a) XLSR-nl layer 15 (b) LD

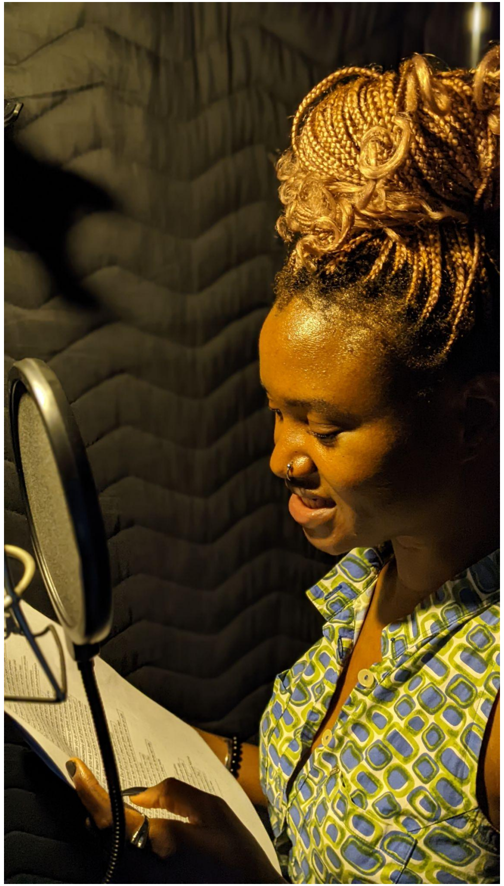
Image from Bartelds & Wieling, 2022

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 monolingual versions of large pretrained speech models

HUBERT-TR: REVIVING TURKISH AUTOMATIC SPEECH RECOGNITION WITH SELF-SUPERVISED SPEECH REPRESENTATION LEARNING

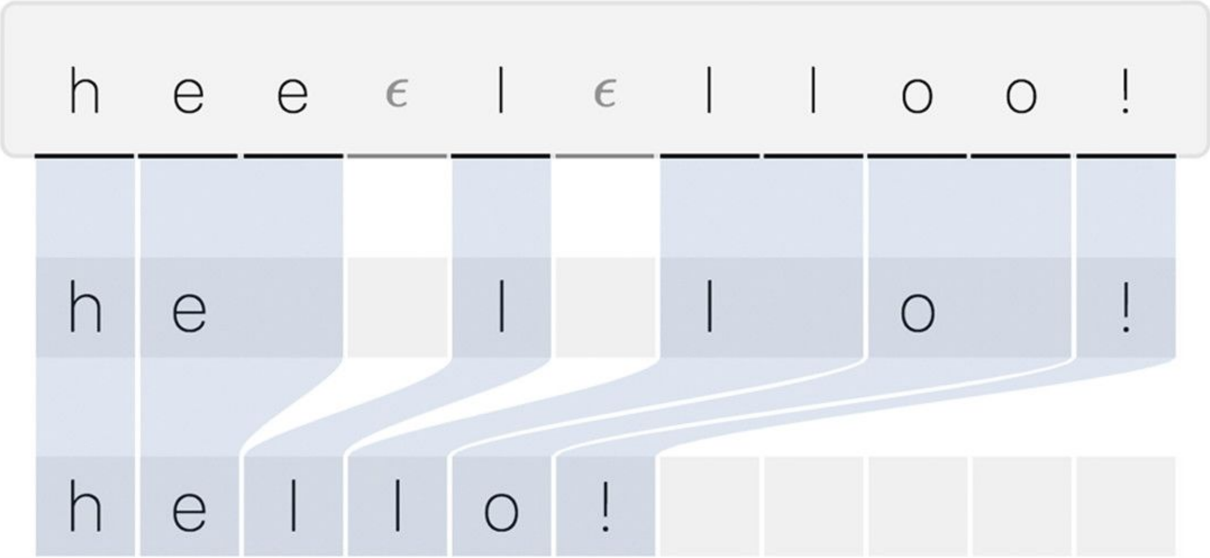
Ali Safaya *, *Engin Erzin*

KUIS AI Center
Computer Engineering Department
Koç University

Using Radio Archives for Low-Resource Speech Recognition: Towards an Intelligent Virtual Assistant for Illiterate Users

Moussa Doumbouya,[1] Lisa Einstein,[1,2] Chris Piech[2]
[1] GNCode
[2] Stanford University
moussa@gncode.org, lisae@stanford.edu, piech@cs.stanford.edu

Different scripts leverage CTC - no need for huge language model



Languages can have codeswitching

What is code-switching?

The mixing of words, phrases and sentences from two distinct grammatical (sub) systems across sentence boundaries within the same speech event.
(Bokomba, 1988)

I'll tell you exactly when I have to leave, at ten o'clock. Y son las nueve y cuarto.

Off-the-shelf multilingual models don't work well in this scenario.

Ògúnrèmí et. al, 2023

How can we improve the performance of large multilingual models on code-switched data?

Ògúnrèmí et. al, 2023

Data: South African Soap Opera Clips

Four South-African languages code-switched with English

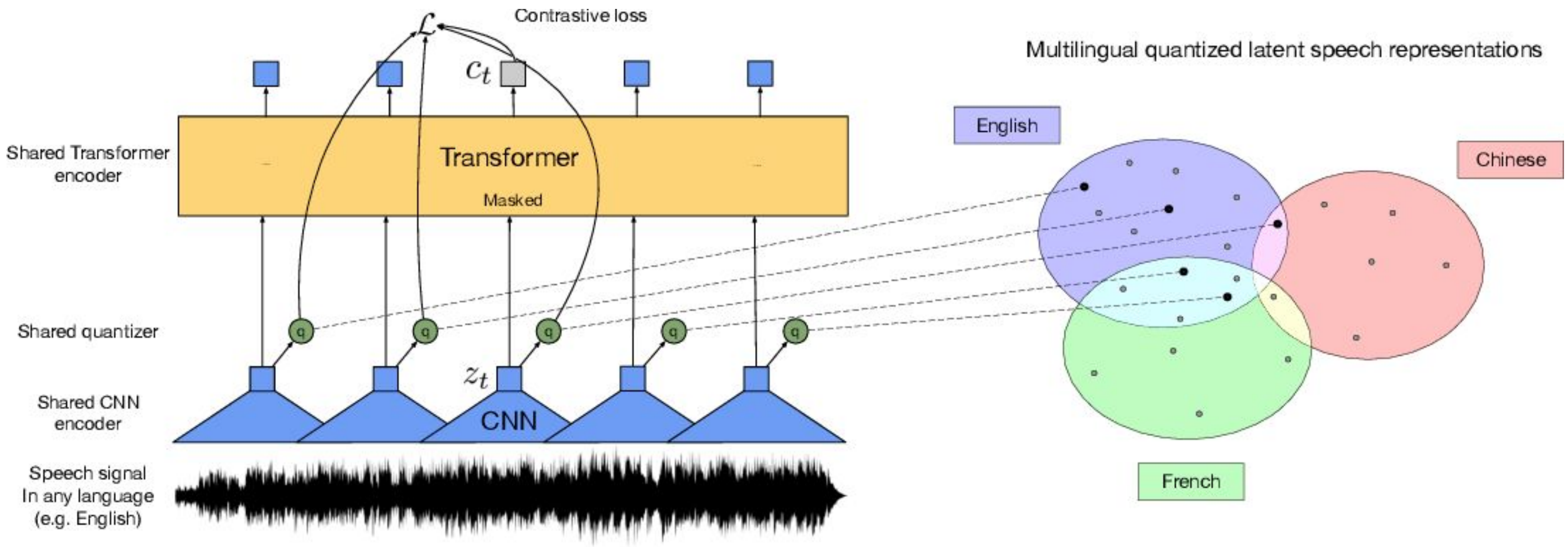
3 - 6 hours per language

Ògúnrèmí et. al, 2023

Data: South African Soap Opera Clips

Lang par	Train	Dev	Test	Total
Eng-Zul	4.81h	0.13h	0.51h	**5.45h**
Eng-Xho	2.68h	0.23h	0.23h	**3.14h**
Eng-Tsn	2.33h	0.23h	0.30h	**2.86h**
Eng-Sot	2.36h	0.21h	0.26h	**2.83h**

Model: wav2vec 2.0 XLSR



Ògúnrèmí et. al, 2023

Does incorporating language information help?

what if etholwa amaphoyisa kuqala

<eng> what if </eng> <zul> etholwa amaphoyisa kuqala </zul>

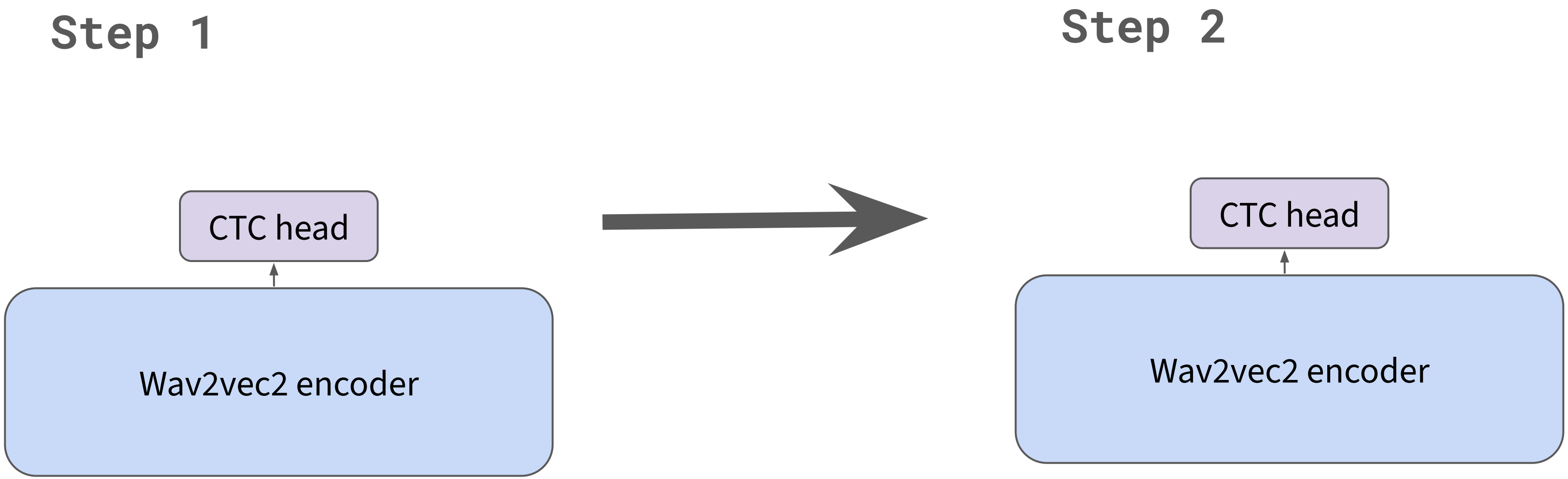
TAGS

WHAT IF etholwa amaphoyisa kuqala

CASING

Ògúnrèmí et. al, 2023

We fine-tune (with a CTC head) first on the language pair along with additional data, then on the language pair itself



language + pair

A: monolingual data
B: the rest of the soap opera corpus

language pair

We find that finetuning with utterances in the same domain (soap opera data) but different, neighbouring languages improve performance over finetuning a single language pair.

Ògúnrèmí et. al, 2023

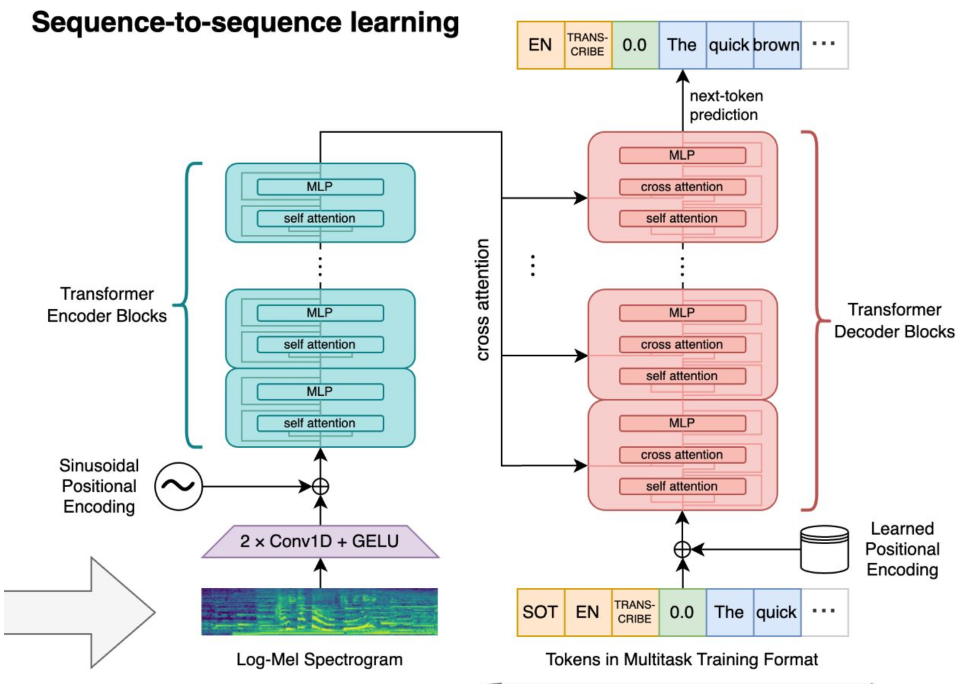
Language is varied

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.



Thank You