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.
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
<table>
<thead>
<tr>
<th>Writing system</th>
<th>Scripts</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Alphabet</strong></td>
<td>Roman</td>
</tr>
<tr>
<td>Greek</td>
<td>Λατρεύω την αναγνώριση ομιλίας</td>
</tr>
<tr>
<td>Cyrillic</td>
<td>Би яриа таних дуртай</td>
</tr>
<tr>
<td>Korean</td>
<td>나는 음성 인식을 좋아해야요</td>
</tr>
<tr>
<td><strong>Semanto-Phonetic</strong></td>
<td>Chinese</td>
</tr>
<tr>
<td>Chinese</td>
<td>我喜欢语音识别</td>
</tr>
<tr>
<td><strong>Syllabic Alphabet</strong></td>
<td>Devanāgarî</td>
</tr>
<tr>
<td>Thai</td>
<td>ननखोबकरुजंकाबुध</td>
</tr>
<tr>
<td>Tamil</td>
<td>நான் பேபச்சு அங்கீகாரத்ைதவிரும்புகிநேறன்</td>
</tr>
<tr>
<td><strong>Abjad</strong></td>
<td>Arabic</td>
</tr>
<tr>
<td>Arabic</td>
<td>أنا أحب التعرف على الكلام</td>
</tr>
<tr>
<td>Hebrew</td>
<td></td>
</tr>
</tbody>
</table>

Languages can have different scripts

Adapted from Tan et. al, 2010
Languages can have **lexical tone**

The pitch of the word changes the meaning of the word

* wá ́ wọ̀

* wà̀ wò̀
Languages can have different dialects

<table>
<thead>
<tr>
<th>English</th>
<th>I don’t know what to do</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jordanian Arabic</td>
<td>مش عارف شو اعمل</td>
</tr>
<tr>
<td>Palestinian Arabic</td>
<td>شو بدي اعمل</td>
</tr>
<tr>
<td>Emirati Arabic</td>
<td>معرف شو اسوي</td>
</tr>
<tr>
<td>Modern Arabic</td>
<td>لا اعلم ماذا افعل</td>
</tr>
<tr>
<td>Egyptian Arabic</td>
<td>مش عارف اعمل ايه</td>
</tr>
<tr>
<td>Tunisian Arabic</td>
<td>منعرفش</td>
</tr>
<tr>
<td>Algerian Arabic</td>
<td>ما على بالي</td>
</tr>
<tr>
<td>Kuwaiti Arabic</td>
<td>ما ادري شو اسوي</td>
</tr>
</tbody>
</table>

Image from Bani-Hani et al., 2017
Languages can have codeswitching.
Languages can have little data available to train models.
Multilingual large pretrained speech models
Multilingual versions of English-only models: wav2vec 2.0 XLSR

- 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

- Latent multilingual speech representations are theorised

Image from Conneau et al., 2020
Multilingual versions of English-only models: wav2vec 2.0 XLS-R

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

Image from Babu et. al, 2021
Multilingual versions of English-only models: wav2vec 2.0 XLS-R

- Tested on ASR and AST (Automatic Speech Translation)

Image from Babu et. al, 2021
Multilingual versions of English-only models: mHuBERT

- 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

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

Image from Radford et. al, 2022
What is the data distribution?
How multilingual are these models?

wav2vec 2.0 XLSR

Languages in wav2vec 2.0 XLSR

- English: 81.1%
- Non-English: 18.9%
How multilingual are these models?

wav2vec 2.0
XLS-R

Languages in wav2vec 2.0 XLS-R

- English: 15.9%
- Non-English: 84.1%
How multilingual are these models?

mHuBERT

Languages in mHuBERT

- Spanish: 33.3%
- English: 33.3%
- French: 33.3%
How multilingual are these models?

languages in Whisper

- English: 22.2%
- Non-English: 37.5%
- Translation: 40.3%

Whisper
Open Source Multilingual Datasets: CommonVoice and Yodas
### 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**

<table>
<thead>
<tr>
<th>Language</th>
<th>Hours</th>
<th>Speakers</th>
<th>Validation Progress</th>
<th>Sentences</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>3485</td>
<td>93140</td>
<td>73%</td>
<td>1673433</td>
</tr>
<tr>
<td>Catalan</td>
<td>3838</td>
<td>35957</td>
<td>73%</td>
<td>1214229</td>
</tr>
<tr>
<td>Kinyarwanda</td>
<td>2388</td>
<td>1134</td>
<td>84%</td>
<td>1404853</td>
</tr>
<tr>
<td>Belarusian</td>
<td>1794</td>
<td>8363</td>
<td>97%</td>
<td>379505</td>
</tr>
<tr>
<td>Esperanto</td>
<td>1936</td>
<td>1758</td>
<td>75%</td>
<td>180562</td>
</tr>
<tr>
<td>German</td>
<td>1412</td>
<td>19151</td>
<td>84%</td>
<td>2056443</td>
</tr>
<tr>
<td>French</td>
<td>1145</td>
<td>19584</td>
<td>88%</td>
<td>1646292</td>
</tr>
<tr>
<td>Kabyle</td>
<td>699</td>
<td>1560</td>
<td>81%</td>
<td>182716</td>
</tr>
<tr>
<td>Spanish</td>
<td>2326</td>
<td>26107</td>
<td>20%</td>
<td>1080695</td>
</tr>
<tr>
<td>Luganda</td>
<td>583</td>
<td>660</td>
<td>76%</td>
<td>191407</td>
</tr>
<tr>
<td>Swahili</td>
<td>1081</td>
<td>1459</td>
<td>37%</td>
<td>134669</td>
</tr>
<tr>
<td>Persian</td>
<td>426</td>
<td>4453</td>
<td>87%</td>
<td>55667</td>
</tr>
</tbody>
</table>
Common Voice

Marathi Common Voice

Kinyarwanda Common Voice
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.

- 420,000 hours of transcribed data
Benchmarking large models multilingually: ML-SUPERB
### SUPERB

Speech processing Universal PERformance Benchmark

![Number of Parameters Diagram](image)

- **Word Error Rate** vs. **Number of Parameters**

```latex
\begin{itemize}
  \item admin\_baseline
  \item STaRHuBERT-S
  \item STaRHuBERT
  \item STaRHuBERT-XS
  \item STaRHuBERT-L
  \item IIITD
  \item ARMwavLM-S
  \item ARMHuBERT
  \item DPWavLM
  \item DPHuBERT
  \item data2vec-aqc Base
  \item CoBERT Base
  \item ccc-wav2vec 2.0 Base
  \item LightHuBERT Small
  \item data2vec Large
  \item data2vec base
  \item LightHuBERT Stage 1
  \item WavLM Large
  \item WavLM Base+
  \item WavLM Base
  \item FaST-VGS+
  \item DistilHuBERT
  \item DeCoAR 2.0
  \item HuBERT Large
  \item HuBERT Base
  \item wav2vec 2.0 Large
  \item wav2vec 2.0 Base
  \item vq-wav2vec
  \item wav2vec
  \item Mockingjay
  \item NPC
  \item VQ-APC
  \item PASE+
  \item APC
  \item modified CPC
  \item TERA
\end{itemize}
```
## ML-SUPERB

> Multilingual Speech processing Universal PERformance Benchmark

Automatic speech recognition and language identification for 143 languages

<table>
<thead>
<tr>
<th>Method</th>
<th>Mono-ASR ↓</th>
<th>Multi-ASR (Normal) ↓</th>
<th>Multi-ASR (Few-shot) ↓</th>
<th>LID ↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>HuBERT Base</td>
<td>35.3</td>
<td>31.4</td>
<td>42.7</td>
<td>86.1</td>
</tr>
<tr>
<td>HuBERT Large</td>
<td>32.2</td>
<td>37.7</td>
<td>43.5</td>
<td>64.1</td>
</tr>
<tr>
<td>Mandarin HuBER...</td>
<td>45.6</td>
<td>43.2</td>
<td>46.6</td>
<td>85.3</td>
</tr>
<tr>
<td>Mandarin HuBER...</td>
<td>33.7</td>
<td>39.6</td>
<td>45.1</td>
<td>57.3</td>
</tr>
<tr>
<td>Robust wav2vec 2...</td>
<td>35.7</td>
<td>31.1</td>
<td>42.2</td>
<td>72.1</td>
</tr>
</tbody>
</table>
Language-specific techniques
Languages can have different scripts

<table>
<thead>
<tr>
<th>Writing system</th>
<th>Scripts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alphabet</td>
<td>Roman</td>
</tr>
<tr>
<td></td>
<td>napenda utambuzi wa hotuba</td>
</tr>
<tr>
<td>Greek</td>
<td>Λατρεύω την αναγνώριση ομιλίας</td>
</tr>
<tr>
<td>Cyrillic</td>
<td>Βι яриа таних дуртай</td>
</tr>
<tr>
<td>Korean</td>
<td>나는 음성 인식을 좋아해요</td>
</tr>
<tr>
<td>Semanto-Phonetic</td>
<td>Chinese</td>
</tr>
<tr>
<td></td>
<td>我喜欢语音识别</td>
</tr>
<tr>
<td>Syllabic Alphabet</td>
<td>Devanāgarī</td>
</tr>
<tr>
<td></td>
<td>मलाई बोली पहिचान मन पछि</td>
</tr>
<tr>
<td>Thai</td>
<td>नान作物การรู้จำคำพูด</td>
</tr>
<tr>
<td>Tamil</td>
<td>கேலும் பெரு அமைக்கப் புது குறிப்பிட்டன்</td>
</tr>
<tr>
<td>Abjad</td>
<td>Arabic</td>
</tr>
<tr>
<td></td>
<td>أنا أحب التعرف على الكلام</td>
</tr>
<tr>
<td>Hebrew</td>
<td></td>
</tr>
</tbody>
</table>

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  wò
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.
Languages can have different dialects

<table>
<thead>
<tr>
<th>English</th>
<th>I don’t know what to do</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jordanian Arabic</td>
<td>مش عارف شو اعمل</td>
</tr>
<tr>
<td>Palestinian Arabic</td>
<td>شو بدي اعمل</td>
</tr>
<tr>
<td>Emirati Arabic</td>
<td>معرف شو اسوي</td>
</tr>
<tr>
<td>Modern Arabic</td>
<td>لا اعلم ماذا افعل</td>
</tr>
<tr>
<td>Egyptian Arabic</td>
<td>مش عارف اعمل ايه</td>
</tr>
<tr>
<td>Tunisian Arabic</td>
<td>منعرفش</td>
</tr>
<tr>
<td>Algerian Arabic</td>
<td>ما على بالي</td>
</tr>
<tr>
<td>Kuwaiti Arabic</td>
<td>ما ادري شو اسوي</td>
</tr>
</tbody>
</table>

Image from Bani-Hani et.al, 2017
Next week!
Languages can have little data available to train models.
Making datasets

ÌròyìnSpeech: A multi-purpose Yorùbá Speech Corpus

Ọgúnrèmí 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 Piech²
¹ GINCode
² Stanford University
moussa@gincode.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.
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

Ọgünrèmì et. al, 2023
## Data: South African Soap Opera Clips

<table>
<thead>
<tr>
<th>Lang par</th>
<th>Train</th>
<th>Dev</th>
<th>Test</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eng-Zul</td>
<td>4.81h</td>
<td>0.13h</td>
<td>0.51h</td>
<td>5.45h</td>
</tr>
<tr>
<td>Eng-Xho</td>
<td>2.68h</td>
<td>0.23h</td>
<td>0.23h</td>
<td>3.14h</td>
</tr>
<tr>
<td>Eng-Tsn</td>
<td>2.33h</td>
<td>0.23h</td>
<td>0.30h</td>
<td>2.86h</td>
</tr>
<tr>
<td>Eng-Sot</td>
<td>2.36h</td>
<td>0.21h</td>
<td>0.26h</td>
<td>2.83h</td>
</tr>
</tbody>
</table>
Model: wav2vec 2.0 XLSR

Ôgúnrèmí et. al, 2023
Does incorporating language information help?

what if etholwa amaphoyisa kuqala

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

WHAT IF etholwa amaphoyisa kuqala

Ögúnrèmí 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.

**Step 1**
- Wav2vec2 encoder
  - CTC head
- language + pair
- A: monolingual data
- B: the rest of the soap opera corpus

**Step 2**
- Wav2vec2 encoder
  - CTC head
- 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.

Ọgúnrèmí et. al, 2023
Language is varied

<table>
<thead>
<tr>
<th>Language</th>
<th>Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>I love speech recognition</td>
</tr>
<tr>
<td>Greek</td>
<td>Λατρεύω την αναγνώριση ομιλίας</td>
</tr>
<tr>
<td>Korean</td>
<td>나는 음성 인식을 좋아해요</td>
</tr>
<tr>
<td>Malay</td>
<td>ฉันชอบการรู้จักคําพูด</td>
</tr>
<tr>
<td>Thai</td>
<td>ฉันรักการตั้งคําถาม</td>
</tr>
<tr>
<td>Tamil</td>
<td>நான் இயற்கைய அங்கீகாரத்தை விரும்புகிறேன்</td>
</tr>
<tr>
<td>Arabic</td>
<td>أنا أحب التعرف على الكلام</td>
</tr>
</tbody>
</table>

Puedo hablar en español.
Or I can speak in English.
Or I can speak en los dos.
Surprisingly, all you need to do is chuck a bunch of data into a model.
Thank You