Benchmarking Code-switching Abilities of Speech Foundation Models

Presenter: R12921119 Chih-Kai Yang

Outline

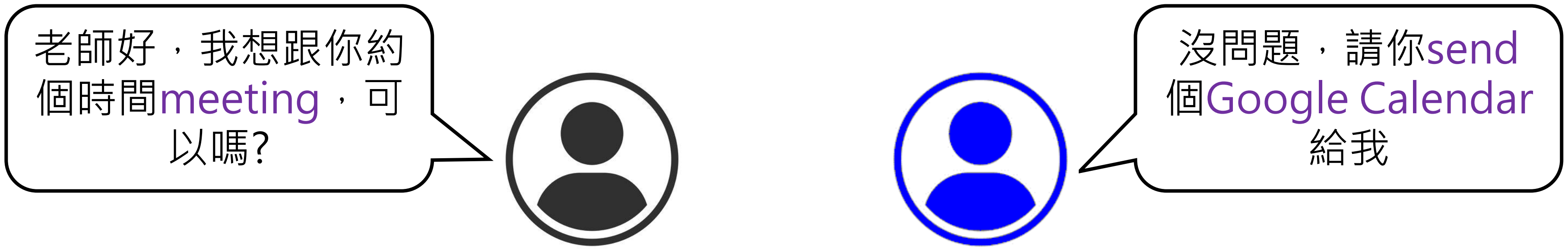
- Code-switching
- Speech foundation models
- Benchmarking code-switching understanding
  - Task
  - Dataset construction
  - Evaluated models
  - Results and findings
- Takeaways

Code-Switching

- Alternates between two or more languages

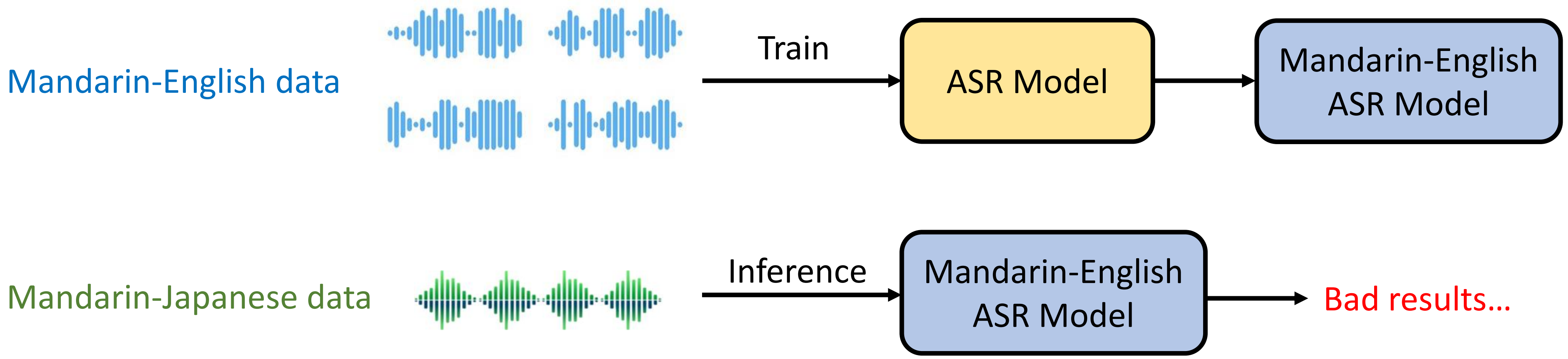
Code-Switching

- Why code-switching speech processing **important**?
  - Common in our daily life
  - Natural way of communication for some countries, e.g. India, Singapore…



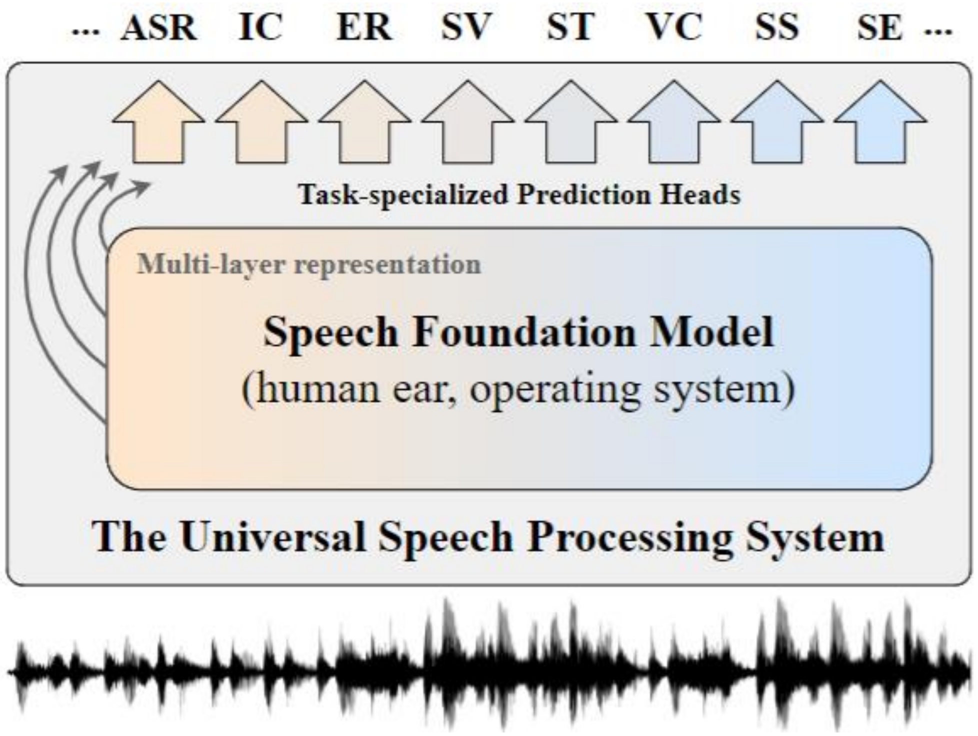
Code-Switching

- Why code-switching speech processing **challenging**?
  - Data scarcity: Rare high-quality labelled data for training
  - Hard to generalize: Typically relying on task-specific design
- May the **speech foundation models** help?



Speech Foundation Models

- Models that can solve many tasks with faithful speech representations

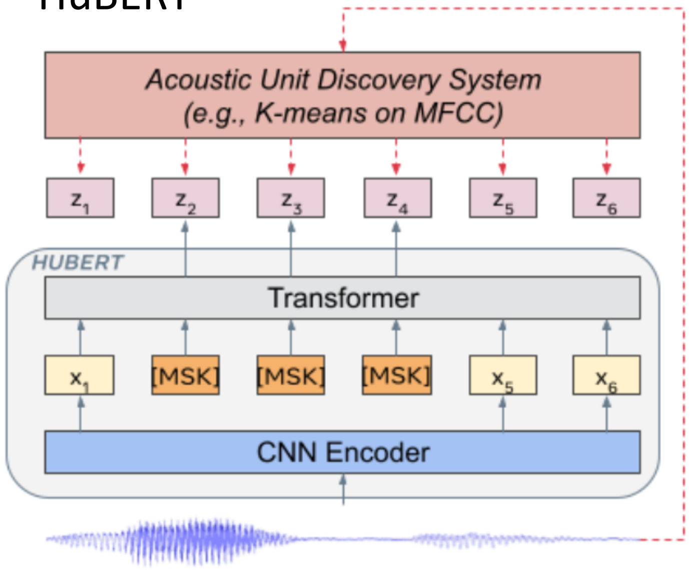


https://arxiv.org/abs/2404.09385

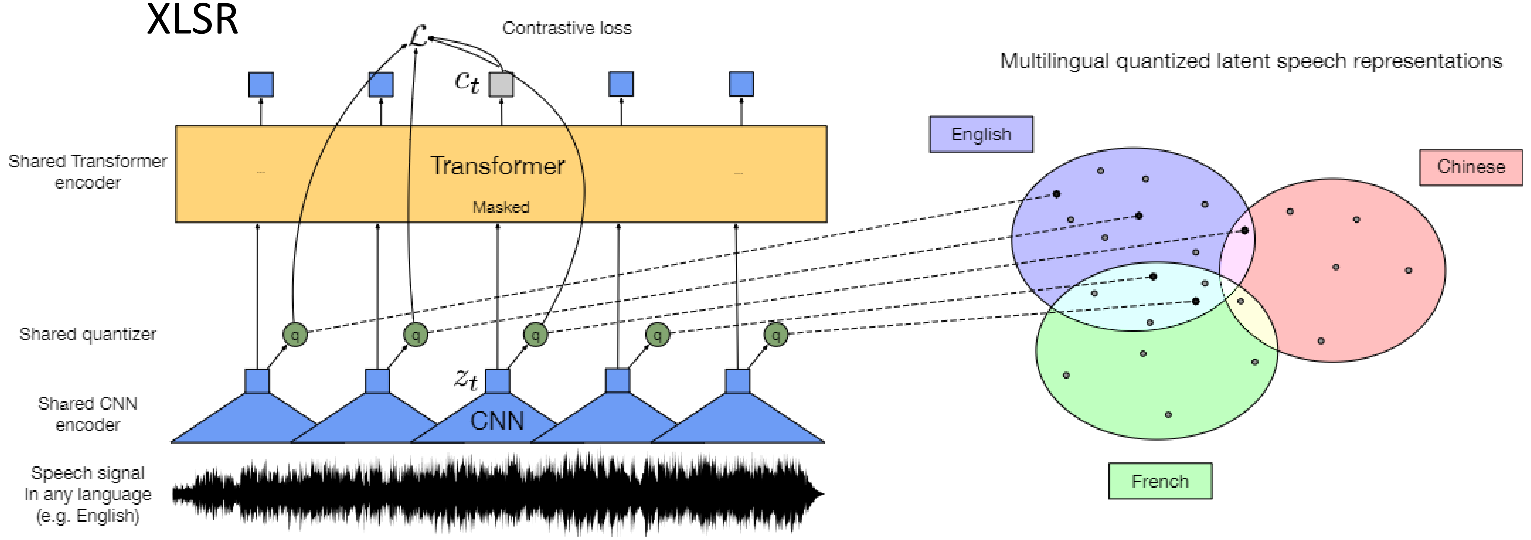
Speech Foundation Models

- There are many foundation models nowadays

HuBERT



XLSR



https://arxiv.org/abs/2106.07447

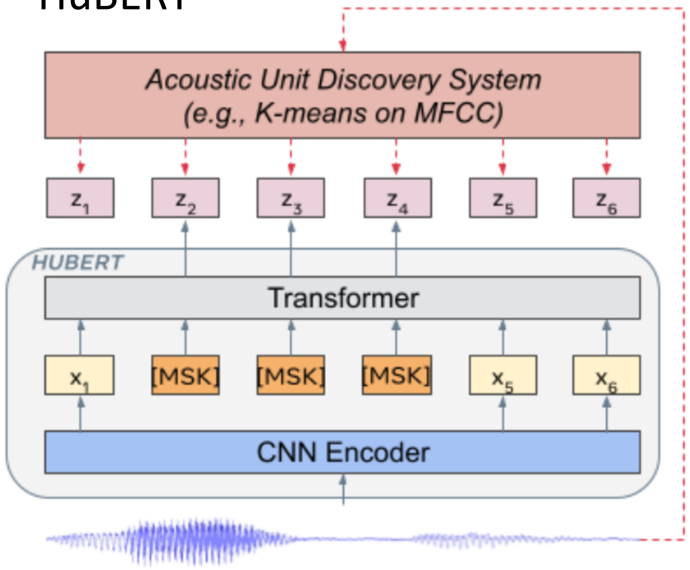
https://arxiv.org/abs/2006.13979

Speech Foundation Models

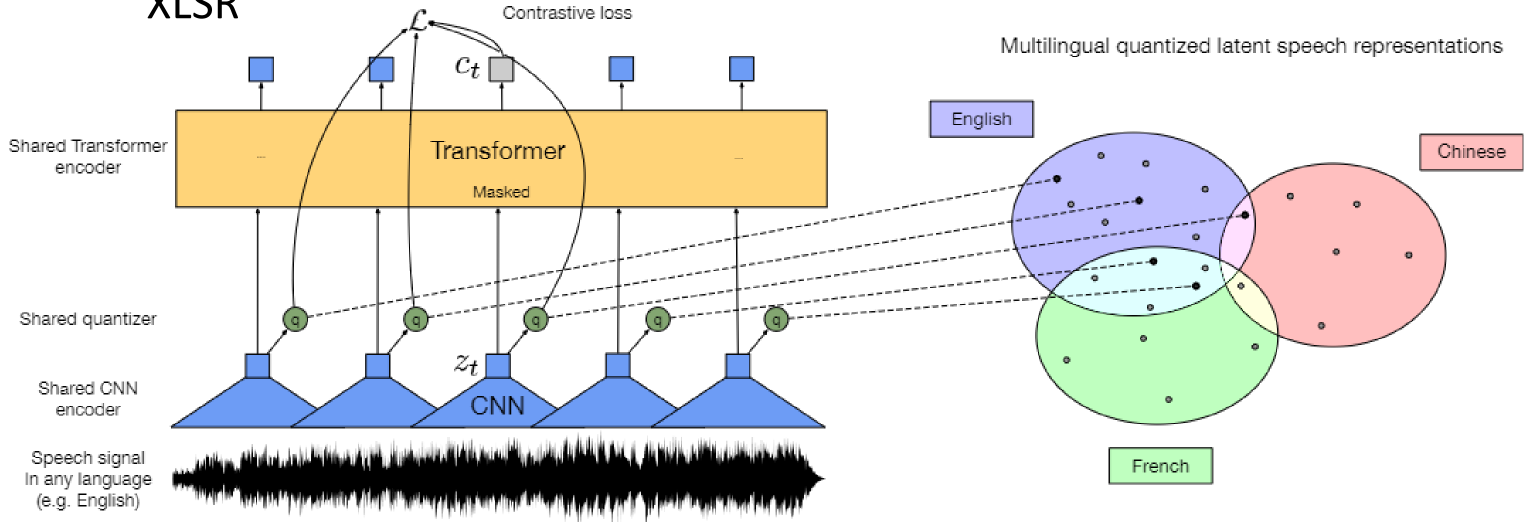
- There are many foundation models nowadays

But how can we know their code-switching abilities?

HuBERT



XLSR



https://arxiv.org/abs/2106.07447

https://arxiv.org/abs/2006.13979

Benchmarking Code-switching Abilities

- Most existing benchmarks need fine-tuning on downstream tasks
  - Cannot tell where the performances are derived. Pre-training, or fine-tuning?
- Can we have a benchmark without fine-tuning on downstream tasks?

Fine-tuning?
Seen code-switching data?

Pre-training?
Faithful Representations?

ZERO RESOURCE CODE-SWITCHED SPEECH BENCHMARK USING SPEECH UTTERANCE PAIRS FOR MULTIPLE SPOKEN LANGUAGES

ICASSP 2024

https://arxiv.org/abs/2310.03018

Task

- Objective: Assessing the semantic and syntactic understanding
- Given two code-switching utterances that are similar in content
  - One is “correct”
    - No semantic inconsistency, logical error, and grammatical errors.
  - The other one is “wrong”
    - Contains semantic inconsistency, logical error, and grammatical errors.
- The model should give higher score for the correct utterance

EX: 這不溶於water vs. 這不溶於fire

Task

- Objective: Assessing the semantic and syntactic understanding
- Given two code-switching utterances that are similar in content
  - One is “correct”
    - No semantic inconsistency, logical error, and grammatical errors.
  - The other one is “wrong”
    - Contains semantic inconsistency, logical error, and grammatical errors.
- The model should give higher score for the correct utterance

EX: 這不溶於water vs. 這不溶於fire

Need both multilingual and cross-lingual understanding

Task

- Objective: Assessing the semantic and syntactic understanding
- Given two code-switching utterances that are similar in content
  - One is "correct"
    - No semantic inconsistency, logical error, and grammatical errors.
  - The other one is "wrong"
    - Contains semantic inconsistency, logical error, and grammatical errors.
- The model should give higher score for the correct utterance

EX: 這不溶於water vs. 這不溶於fire

Chinese understanding ⟷ English understanding

Cross-lingual understanding

Need both multilingual and cross-lingual understanding

Dataset Construction

Step 1: Generate code-switch sentence from a monolingual sentence.

Prompt:
You are a code-switch sentence generator. Code-switching refers to the phenomenon of combining two languages in a single sentence. You will receive a sentence. You have to generate a code-switch sentence based on the given sentence. Quote the output in quotation marks.
Based on the sentence [input sentence], generate a code-switched sentence switching between two languages, Chinese and English. No other languages besides Chinese and English are allowed. Don't just repeat the original sentence in another language.

Output:
input sentence: "不溶于水。" (translation: Does not dissolve in water.)
output sentence: "这是不溶于water的。" (translation: This does not dissolve in water.)

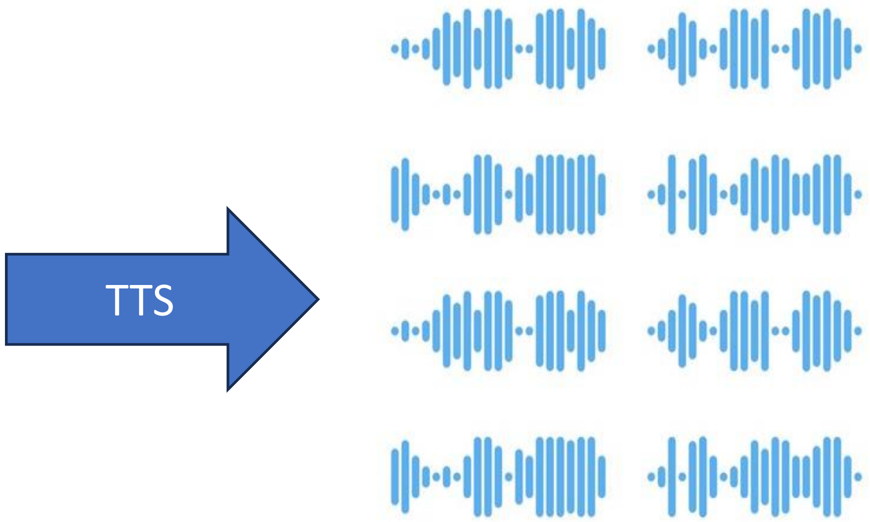
Step 2: Generate meaningless or erroneous code-switched sentence based on a given code-switch sentence.

Prompt:
Code-switching refers to the phenomenon of combining two languages in a single sentence. Given a code-switched sentence, randomly switch or replace at most three words so that the sentence becomes meaningless or erroneous but still remains as a code-switched sentence. [input sentence (correct)]

Output:
input sentence (correct): "这是不溶于water的。" (translation: This does not dissolve in water.)
output sentence (wrong): "这是不溶于fire的。" (translation: This does not dissolve in fire.)

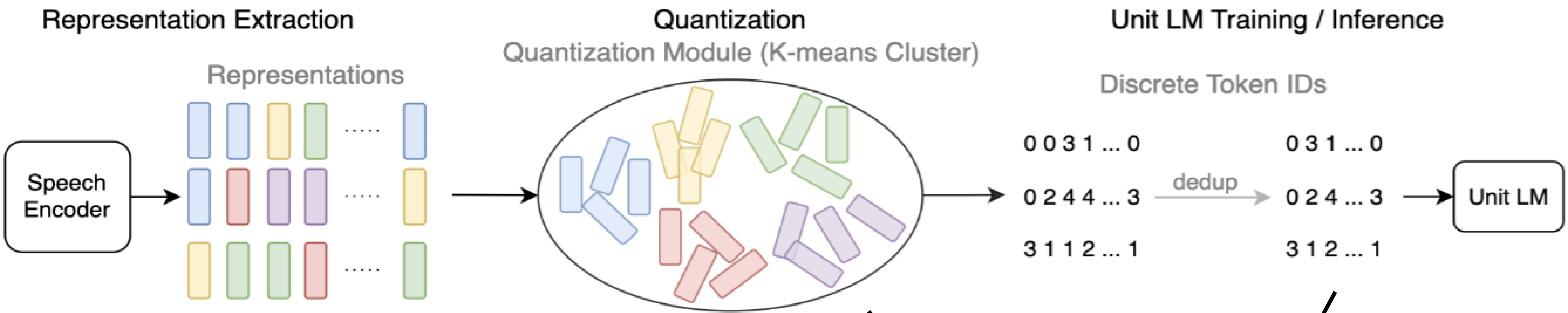
Verified by human annotators

- No existing dataset for this task
- Construct with LLM
- Language: zh/es/fr-en



Speech pairs

Training Pipeline



No code-switching data

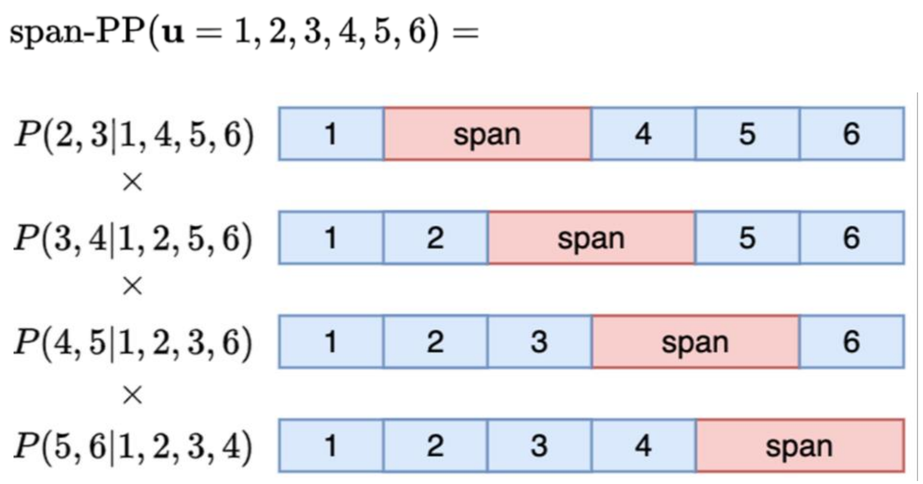
So we are assessing the quality of representations

Evaluation

- Accuracy for the classification
- Classified with the score (no need for training classification head):

$$\text{span-PP}_{w,s}(\mathbf{u}) = \prod_{i=1+j\cdot s} P(u_i \cdots u_{i+w} | u_1 \cdots u_{i-1} u_{i+w+1} \cdots u_T)$$

Can be viewed as the prob. of the utterance



Assessed Models

- Speech foundation models (all based on self-supervised learning, SSL)
  - HuBERT XL, base
  - **mHuBERT**
  - Wav2vec2.0 Large
  - **XLSR-53, XLS-R 0.3B, XLS-R 1B**
- Text-based language models (serving as toplines)
  - XLMR base
  - XGLM 1.7B

Results & Findings

Speech encoder	# param. (B)	km: 100 cluster mono speech (hr)	Unit LM (RoBERTa) mono speech (hr)	dedup	es-en Acc ↑	fr-en Acc ↑	zh-en Acc ↑	avg Acc ↑
Multilingual Speech Encoders								
XLSR-53 (53 lang)	0.3	es, fr, zh, en 25 each	es, fr, zh, en 100 each	V	33.74	45.25	47.20	42.06
XLS-R 0.3B (128 lang)	0.3	es, fr, zh, en 25 each	es, fr, zh, en 100 each	V	75.16	59.30	43.18	59.21
XLS-R 1B (128 lang)	1	es, fr, zh, en 25 each	es, fr, zh, en 100 each	V	33.30	38.66	39.22	37.06
mHuBERT (es, fr, en)	0.09	es, fr, en 33 each	es, fr, en 133 each	V	29.55	30.42	40.33	33.43
Monolingual Speech Encoders								
Wav2vec 2.0 LARGE (ll60k)	0.3	en 100	en 400	V	13.11	25.35	42.41	26.96
HuBERT X-LARGE (ll60k)	1	en 100	en 400	V	24.54	25.60	38.60	29.58
HuBERT Base (LS960)	0.09	en 100	en 400	V	22.26	25.30	40.24	29.27
XLM-RoBERTa Base (text-base)	0.125	-	-	-	54.62	55.12	55.16	54.97
XGLM 1.7B (text-base)	1.7	-	-	-	90.91	88.38	92.03	90.44

Results & Findings

Speech encoder	# param. (B)	km: 100 cluster mono speech (hr)	Unit LM (RoBERTa) mono speech (hr)	dedup	es-en Acc ↑	fr-en Acc ↑	zh-en Acc ↑	avg Acc ↑
Multilingual Speech Encoders								
XLSR-53 (53 lang)	0.3	es, fr, zh, en 25 each	es, fr, zh, en 100 each	V	33.74	45.25	47.20	42.06
XLS-R 0.3B (128 lang)	0.3	es, fr, zh, en 25 each	es, fr, zh, en 100 each	V	75.16	59.30	43.18	59.21
XLS-R 1B (128 lang)	1	es, fr, zh, en 25 each	es, fr, zh, en 100 each	V	33.30	38.66	39.22	37.06
mHuBERT (es, fr, en)	0.09	es, fr, en 33 each	es, fr, en 133 each	V	29.55	30.42	40.33	33.43
Monolingual Speech Encoders								
Wav2vec 2.0 LARGE (ll60k)	0.3	en 100	en 400	V	13.11	25.35	42.41	26.96
HuBERT X-LARGE (ll60k)	1	en 100	en 400	V	24.54	25.60	38.60	29.58
HuBERT Base (LS960)	0.09	en 100	en 400	V	22.26	25.30	40.24	29.27
XLM-RoBERTa Base (text-base)	0.125	-	-	-	54.62	55.12	55.16	54.97
XGLM 1.7B (text-base)	1.7	-	-	-	90.91	88.38	92.03	90.44

Multilingual pre-training helps!

Results & Findings

Speech encoder	# param. (B)	km: 100 cluster mono speech (hr)	Unit LM (RoBERTa) mono speech (hr)	dedup	es-en Acc ↑	fr-en Acc ↑	zh-en Acc ↑	avg Acc ↑
Multilingual Speech Encoders								
XLSR-53 (53 lang)	0.3	es, fr, zh, en 25 each	es, fr, zh, en 100 each	V	33.74	45.25	47.20	42.06
XLS-R 0.3B (128 lang)	0.3	es, fr, zh, en 25 each	es, fr, zh, en 100 each	V	75.16	59.30	43.18	59.21
XLS-R 1B (128 lang)	1	es, fr, zh, en 25 each	es, fr, zh, en 100 each	V	33.30	38.66	39.22	37.06
mHuBERT (es, fr, en)	0.09	es, fr, en 33 each	es, fr, en 133 each	V	29.55	30.42	40.33	33.43
Monolingual Speech Encoders								
Wav2vec 2.0 LARGE (ll60k)	0.3	en 100	en 400	V	13.11	25.35	42.41	26.96
HuBERT X-LARGE (ll60k)	1	en 100	en 400	V	24.54	25.60	38.60	29.58
HuBERT Base (LS960)	0.09	en 100	en 400	V	22.26	25.30	40.24	29.27
XLM-RoBERTa Base (text-base)	0.125	-	-	-	54.62	55.12	55.16	54.97
XGLM 1.7B (text-base)	1.7	-	-	-	90.91	88.38	92.03	90.44

Smaller models perform better

Results & Findings

Speech encoder	# param. (B)	km: 100 cluster mono speech (hr)	Unit LM (RoBERTa) mono speech (hr)	dedup	es-en Acc ↑	fr-en Acc ↑	zh-en Acc ↑	avg Acc ↑
Multilingual Speech Encoders								
XLSR-53 (53 lang)	0.3	es, fr, zh, en 25 each	es, fr, zh, en 100 each	V	33.74	45.25	47.20	42.06
XLS-R 0.3B (128 lang)	0.3	es, fr, zh, en 25 each	es, fr, zh, en 100 each	V	75.16	59.30	43.18	59.21
XLS-R 1B (128 lang)	1	es, fr, zh, en 25 each	es, fr, zh, en 100 each	V	33.30	38.66	39.22	37.06
mHuBERT (es, fr, en)	0.09	es, fr, en 33 each	es, fr, en 133 each	V	29.55	30.42	40.33	33.43
Monolingual Speech Encoders								
Wav2vec 2.0 LARGE (ll60k)	0.3	en 100	en 400	V	13.11	25.35	42.41	26.96
HuBERT X-LARGE (ll60k)	1	en 100	en 400	V	24.54	25.60	38.60	29.58
HuBERT Base (LS960)	0.09	en 100	en 400	V	22.26	25.30	40.24	29.27
XLM-RoBERTa Base (text-base)	0.125	-	-	-	54.62	55.12	55.16	54.97
XGLM 1.7B (text-base)	1.7	-	-	-	90.91	88.38	92.03	90.44

More languages in pre-training helps

Results & Findings

Speech encoder	# param. (B)	km: 100 cluster mono speech (hr)	Unit LM (RoBERTa) mono speech (hr)	dedup	es-en Acc ↑	fr-en Acc ↑	zh-en Acc ↑	avg Acc ↑
Multilingual Speech Encoders								
XLSR-53 (53 lang)	0.3	es, fr, zh, en 25 each	es, fr, zh, en 100 each	V	33.74	45.25	47.20	42.06
XLS-R 0.3B (128 lang)	0.3	es, fr, zh, en 25 each	es, fr, zh, en 100 each	V	75.16	59.30	43.18	59.21
XLS-R 1B (128 lang)	1	es, fr, zh, en 25 each	es, fr, zh, en 100 each	V	33.30	38.66	39.22	37.06
mHuBERT (es, fr, en)	0.09	es, fr, en 33 each	es, fr, en 133 each	V	29.55	30.42	40.33	33.43
Monolingual Speech Encoders								
Wav2vec 2.0 LARGE (ll60k)	0.3	en 100	en 400	V	13.11	25.35	42.41	26.96
HuBERT X-LARGE (ll60k)	1	en 100	en 400	V	24.54	25.60	38.60	29.58
HuBERT Base (LS960)	0.09	en 100	en 400	V	22.26	25.30	40.24	29.27
XLM-RoBERTa Base (text-base)	0.125	-	-	-	54.62	55.12	55.16	54.97
XGLM 1.7B (text-base)	1.7	-	-	-	90.91	88.38	92.03	90.44

Huge gaps between text and speech models

Quick Summary

- Not showing good understanding for speech models
  - Huge gap between text-based models and speech foundation models
- Some points helpful for this task
  - Multilingual pre-training
  - Smaller model size
  - Wide coverage of pre-training languages

Besides SSL Models...

- Other multilingual models like Whisper, SeamlessM4T may not perform well on code-switching ASR

Index	Model	Prompting strategy	CSZS-correct MER(↓)
1	TCS	-	70.59
2	SeamlessM4T *medium*	-	71.12
3	SeamlessM4T *large*	-	69.52
4	SeamlessM4T v2	-	**63.69**
5	Whisper-large-v3	`nonconcat`	60.17

https://arxiv.org/abs/2401.00273

Takeaways

- Code-switching speech processing is challenging
- Current speech foundation models don’t exhibit strong semantic and syntactic understanding

References

- Yang, Shu-wen et al. “**A Large-Scale Evaluation of Speech Foundation Models**”
- Hsu, Wei-Ning et al. “**HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units**”
- Conneau, Alexis et al. “**Unsupervised Cross-lingual Representation Learning for Speech Recognition**”
- Huang, Kuan-Po et al. “**Zero Resource Code-switched Speech Benchmark Using Speech Utterance Pairs For Multiple Spoken Languages**”
- Yang, Chih-Kai et al. “**Investigating Zero-Shot Generalizability on Mandarin-English Code-Switched ASR and Speech-to-text Translation of Recent Foundation Models with Self-Supervision and Weak Supervision**”

THANKS
FOR
LISTENING