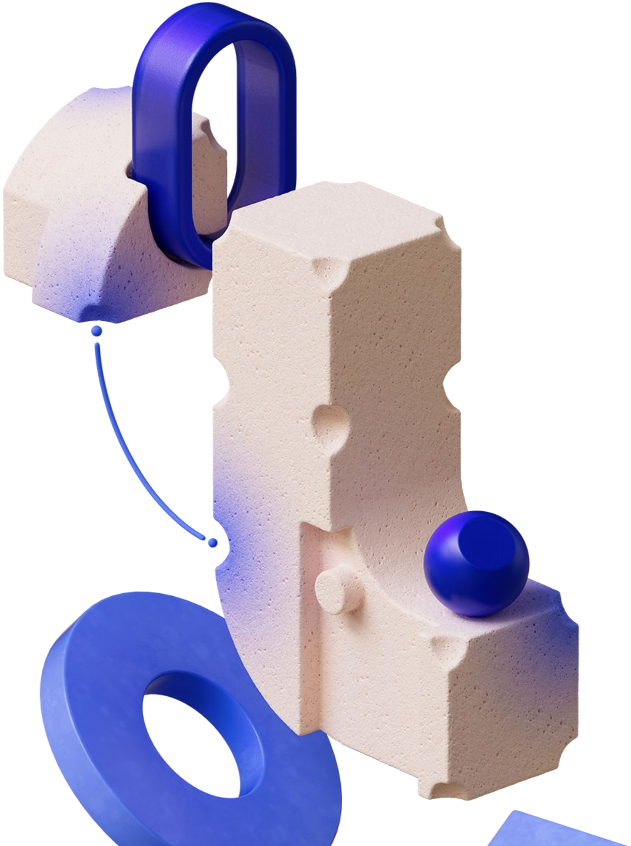
Google DeepMind



Improving Multilingual Speech Recognition and Language Identification

Min Ma
Research Scientist

April 2025

Multilingual ASR with Semi-supervised Learning

1

Reimagining Babel: The AI-Powered Ascent to Universal Understanding

Rebuilt Babel Tower in 21st Century?

- The World's Linguistic Landscape:
  - 7,164 living languages so far
  - 3,866 languages uses an established writing system
- The Digital Babel: a *universal* automatic speech recognition (ASR) system that transcribes speech from *any language*
  a. **Algorithms:** Deep Learning
  b. **Data:** Massive Datasets
  c. **Compute:** GPU/TPU-Accelerated Computing"

Digital Babel: Massively Multilingual Speech Recognition as Foundation

Imaging it works:

- Seamlessly:
  - Multilingual speech recognition
  - Multilingual speech translation
  - Multilingual speech understanding
  - Multilingual speech synthesis
  - ...
- Let's start with Massively Multilingual Automatic Speech Recognition:
  - Feed all the data in!
  - Wait, we also need effective model architecture and learning algorithm

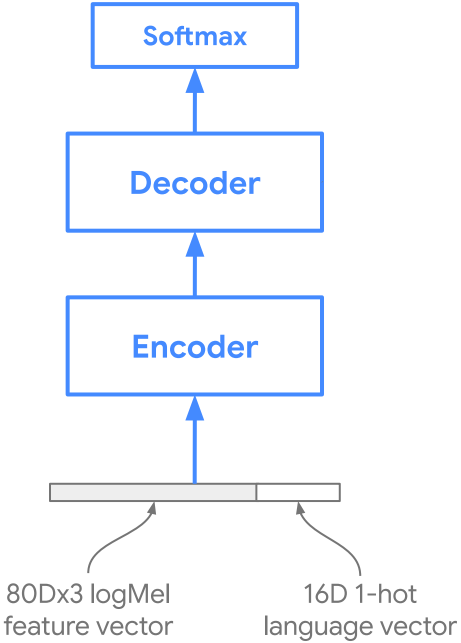
Multilingual Speech Recognition Overview

- **Challenges:**
  - Building Automatic Speech Recognition (ASR) models across many languages is difficult due to large variations and heavily unbalanced data.
  - While multilingual models often help low-resource languages through positive transfer, high-resource languages frequently experience performance degradation.
  - This degradation is often caused by interference from diverse multilingual data and reduced capacity per language.

- **Dataset Overview:**
  - **Scale:** 15 languages from 9 distinct families.
  - **Size:** Data per language varied significantly, from 7.6K to 53.5K hours.
  - **Total:** 359.2K hours of speech data (over 231 million utterances). This scale poses a challenge for multilingual models to outperform strong monolingual baselines.

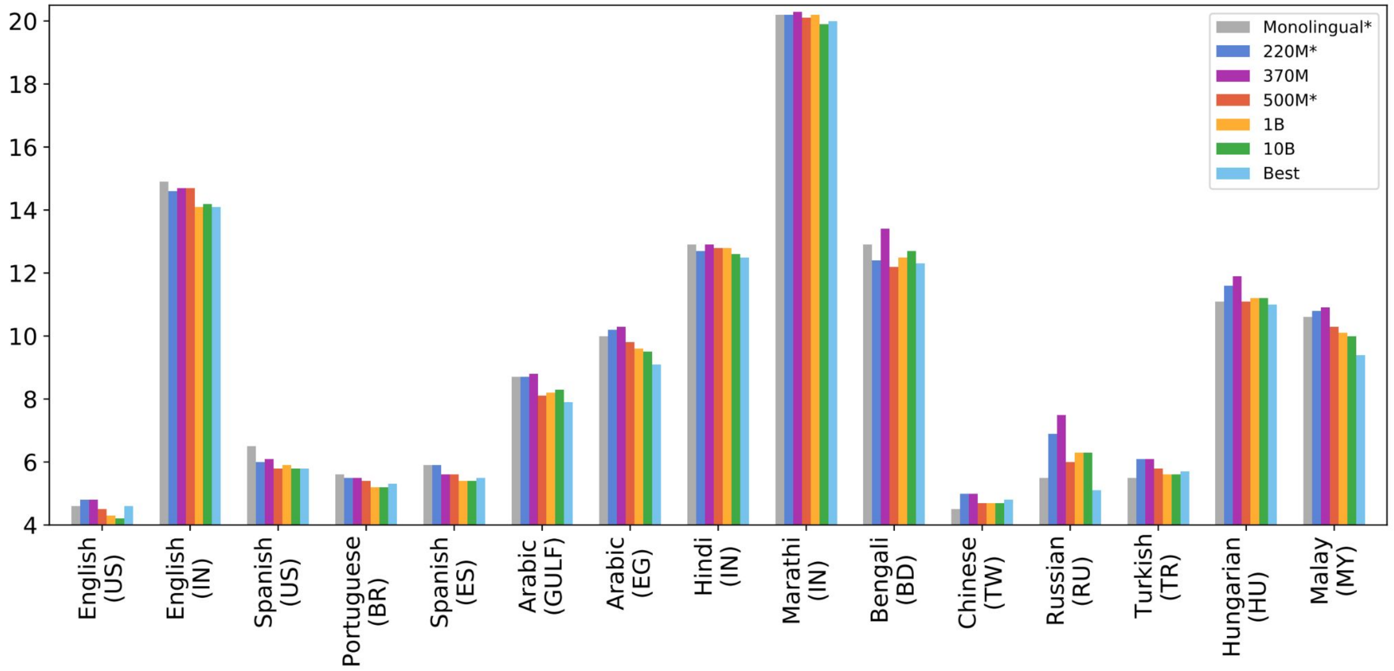
Multilingual Speech Recognition Model Architecture

- 1-hot LangID as additional input
- Conformer Encoder
  - full-context
  - 2X time reduction
- Transformer Decoder
  - masked self-attention
  - cross attention to encoder



Multilingual Speech Recognition Evaluation

- A single multilingual model with average word error rate (WER) of 8.9% vs. 15 mono models' 9.3%



Increasing Model Capabilities Can Help

- **Adding depth works better than width**
  - deeper encoder is better.
  - deeper decoder is better.
  - increase both with more on depth does the best.

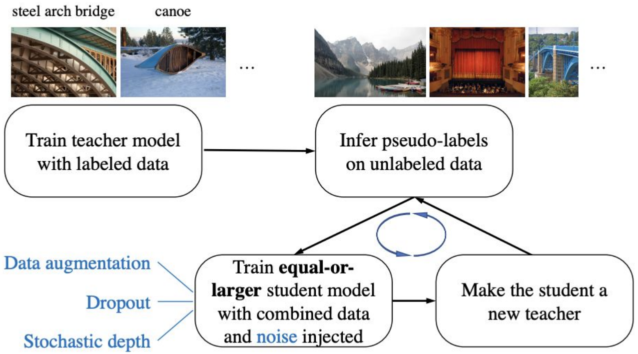
Remaining Questions

Lack of supervised speech data:

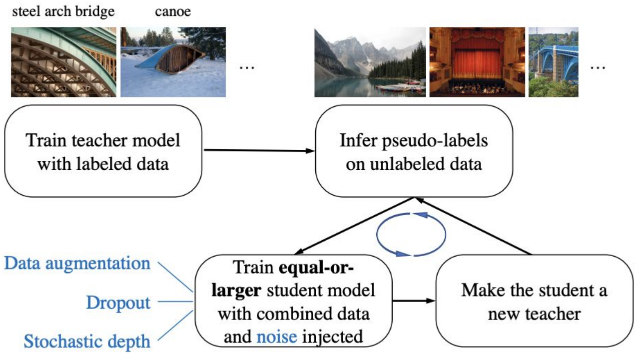
- High-performing multilingual ASR systems typically require vast amounts of labeled data (transcribed audio).

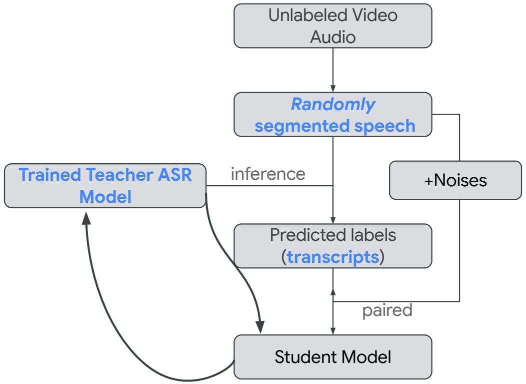
- Creating large labeled datasets is expensive, time-consuming, and often a barrier to progress, especially for diverse domains and languages
  - **Idea 1: can we convert unlabeled speech → labeled speech automatically?**
  - **Idea 2: can we change the way of learning from unlabeled speech?**

Idea 1: Noisy Student Training (with augmented speech segments)



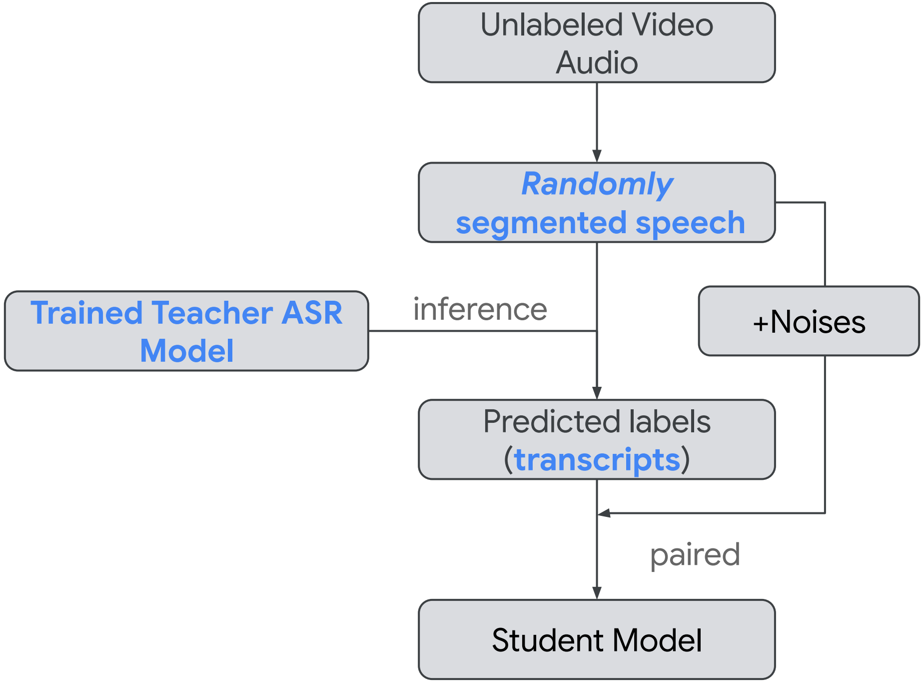
Noisy Student Training (with augmented speech segments)





- **Iterative Self-Training:** uses unlabeled data through sequential teacher-student model training.
- **Teacher-Student Framework:**
  - Teacher: Generates pseudo-labels on unlabeled data from clean input.
  - Student: Trained on these labels with **heavy data augmentation**.
- **Robustness via Augmentation:** Student learns to handle noisy/varied inputs, improving generalization.
  - SpecAugment.

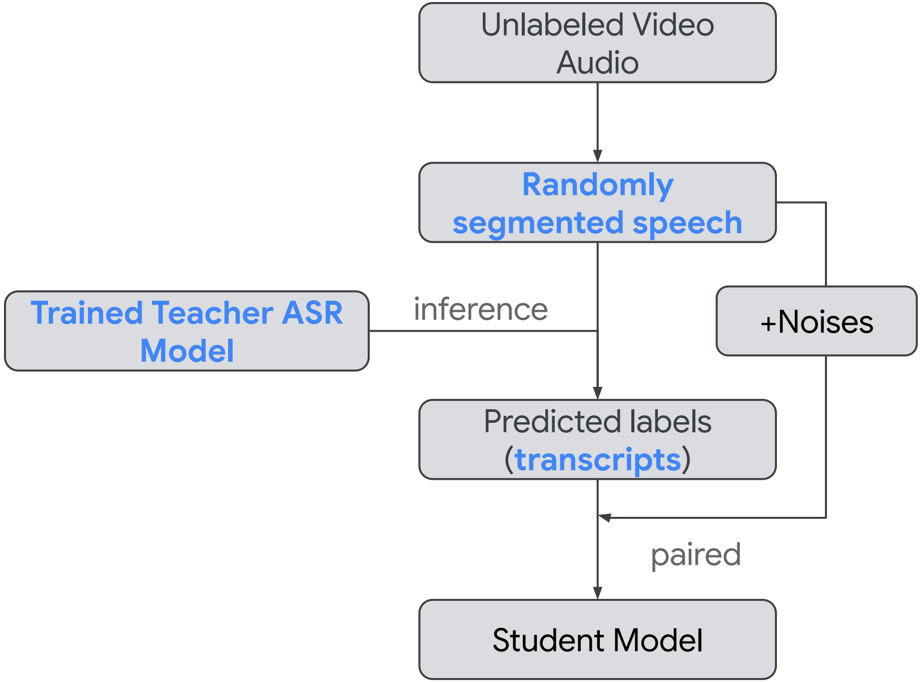
Idea 1: can we convert unlabeled speech ⟶ labeled speech automatically?



- **Iterative Self-Training:** uses unlabeled data through sequential teacher-student model training.
- **Teacher-Student Framework:**
  - Teacher: Generates pseudo-labels on unlabeled data from clean input.
  - Student: Trained on these labels with **heavy data augmentation**.
- **Robustness via Augmentation:** Student learns to handle noisy/varied inputs, improving generalization.
  - SpecAugment.
- **Pseudo-Label Quality:**
  - Filtering: Low-confidence labels are removed.
  - Balancing: Label distribution matches labeled data.
- **Benefit:** Improved ASR accuracy and robustness by effectively utilizing unlabeled data through augmented self-training with quality control.

What would the Noisy Student Training enable us to do?

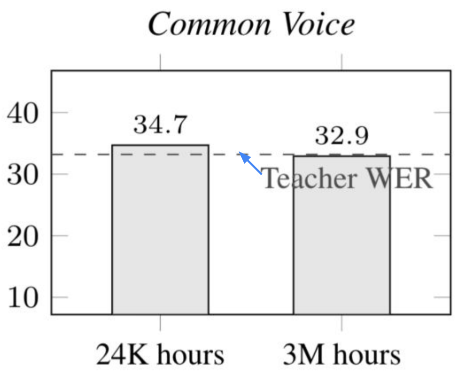
- **Reduced requirement of large amount of human transcribed speech**
  - Especially valuable for **low-resource speech**.
  - Technically, all the available unlabeled speech can be used to train student model.
- **Enabled heterogeneous model architectures of teacher vs. student models**
  - Teacher can be non-streaming ASR model, while student can be streaming.
    - Less latency.
    - On-device model serving.
  - Implicitly knowledge transfer from full context representations learned by teacher model.



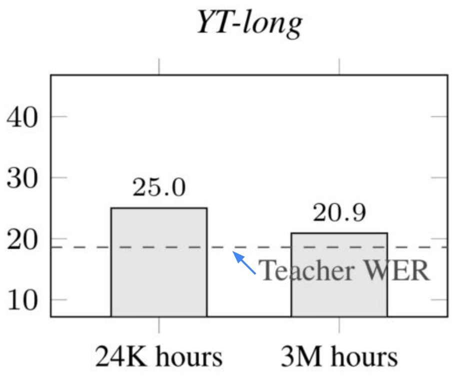
Experimental Results

On monolingual language:

Short-form speech



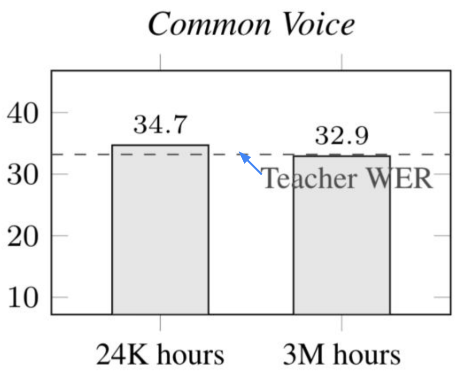
Long-form speech



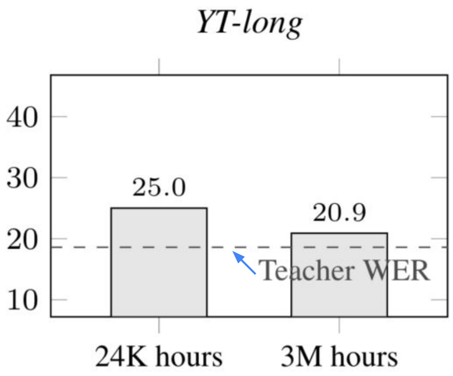
Experimental Results

On monolingual language:

Short-form speech



Long-form speech



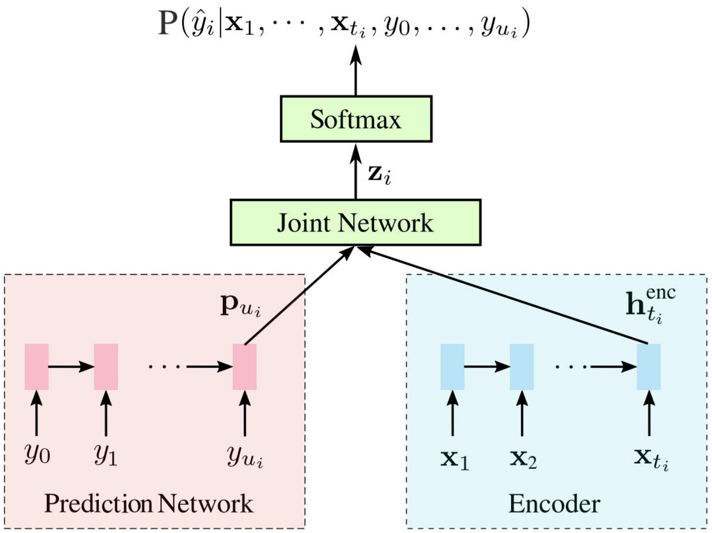
Scaling up for multiple languages:

Language	Test Sets	Streaming Baseline	Non-Streaming teacher model	Streaming student model
French	YT-long Common Voice	34.5 36.2	18.6 33.2	**25.0** **34.7**
Spanish	YT-long Common Voice	35.9 22.0	18.6 11.2	**28.0** **16.5**
Portuguese	YT-long Common Voice	30.8 30.9	22.8 25.8	**28.3** **28.9**
Italian	YT-long Common Voice	24.0 30.0	16.2 27.3	**20.8** **23.6**

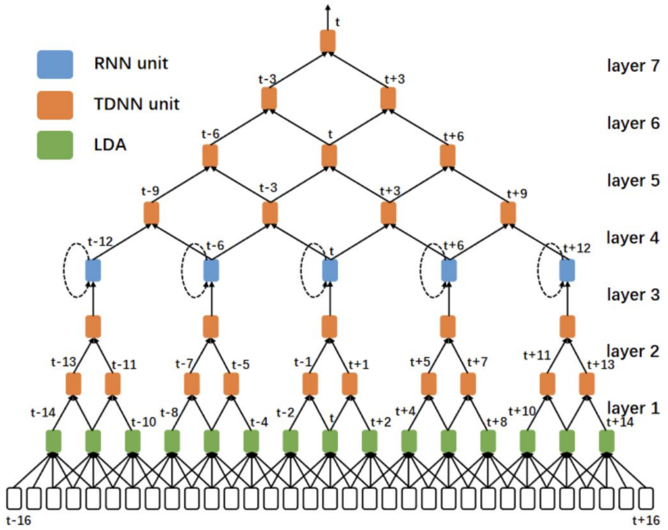
Analysis of different teacher models

Popular ASR Model Architectures:

- **Recurrent Neural Network Transducer (RNN-T)**



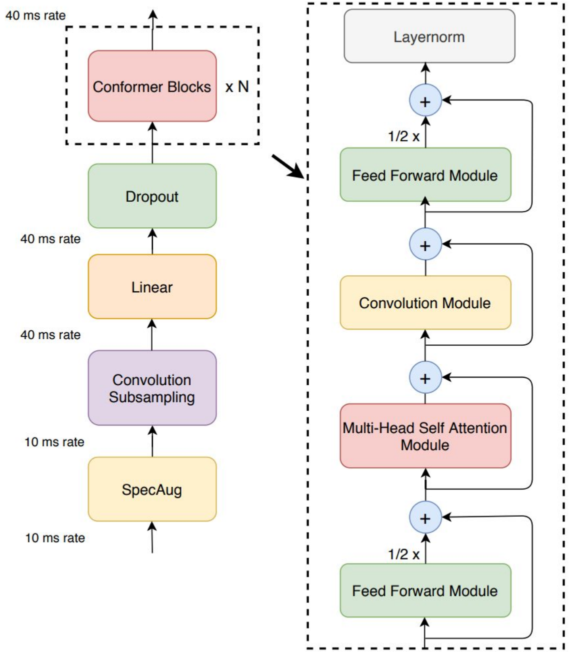
- **Time Delay Neural Network (TDNN)**



Analysis of different teacher models

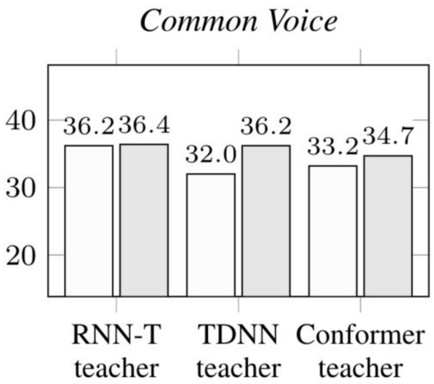
Popular ASR Model Architectures (cont.):

- **Conformer:**

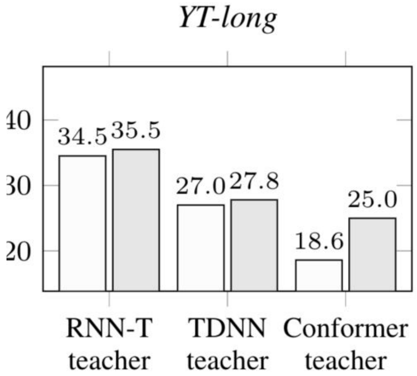


Exploring different teacher models:

Short-form speech



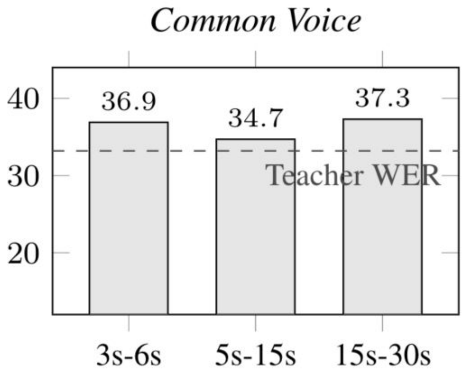
Long-form speech



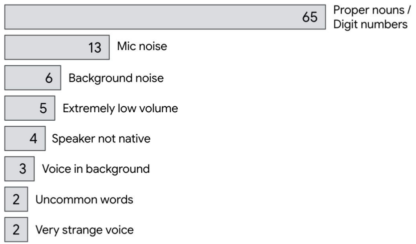
Sensitivity Analysis of WER w.r.t. duration range of random segments

Exploring different teacher models:

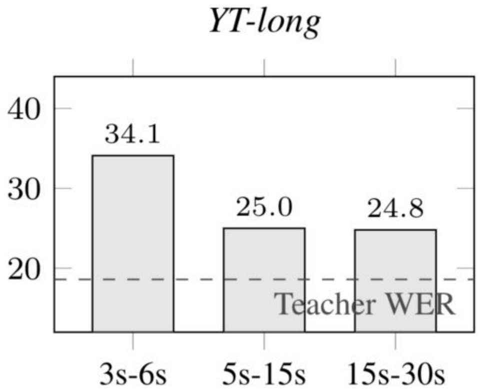
Short-form speech



Among 100 sentences:



Long-form speech



WER (del/ins/sub)

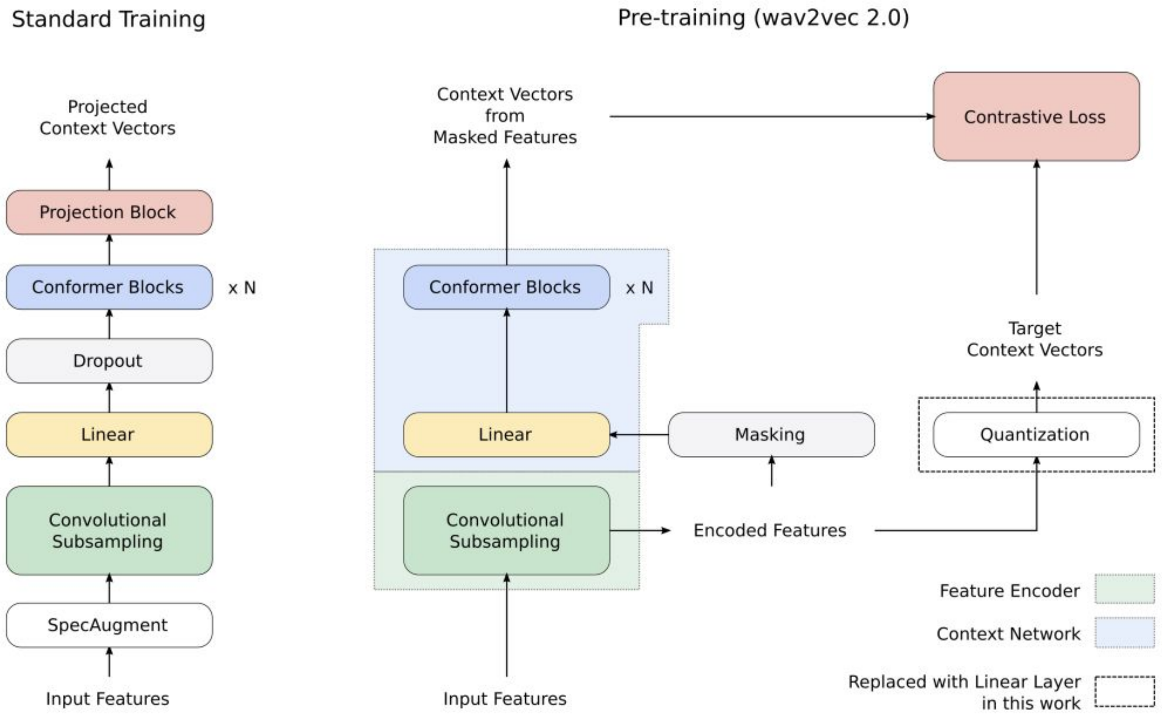
34.1 (13.4/4.4/16.3)	**25.0 (13.4/2.6/9.0)**	**24.8 (12.5/2.7/9.6)**

Idea 2: can we change the way of learning from unlabeled speech?

Improved training recipe (BigSSL):

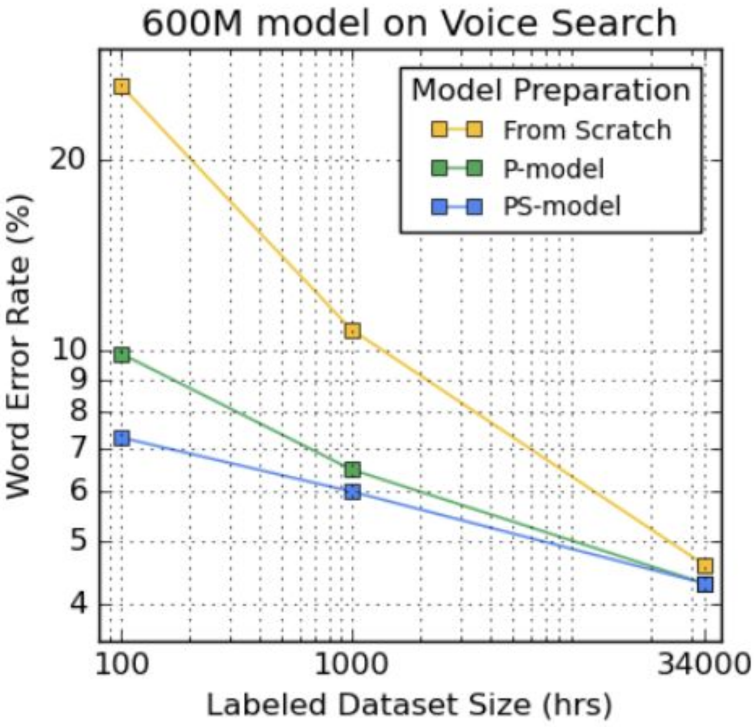
- Pre-training
- Self-training
- Scaling

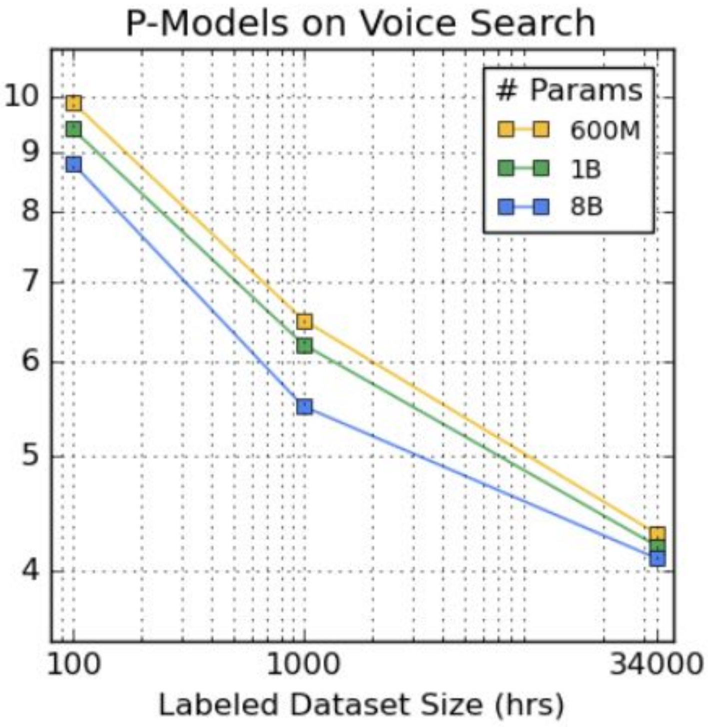
Model	# Params (B)
Conformer XL	0.6
Conformer XXL	1.0
Conformer G	8.0



The BigSSL Results: Pre-training, Self-training, and Scaling

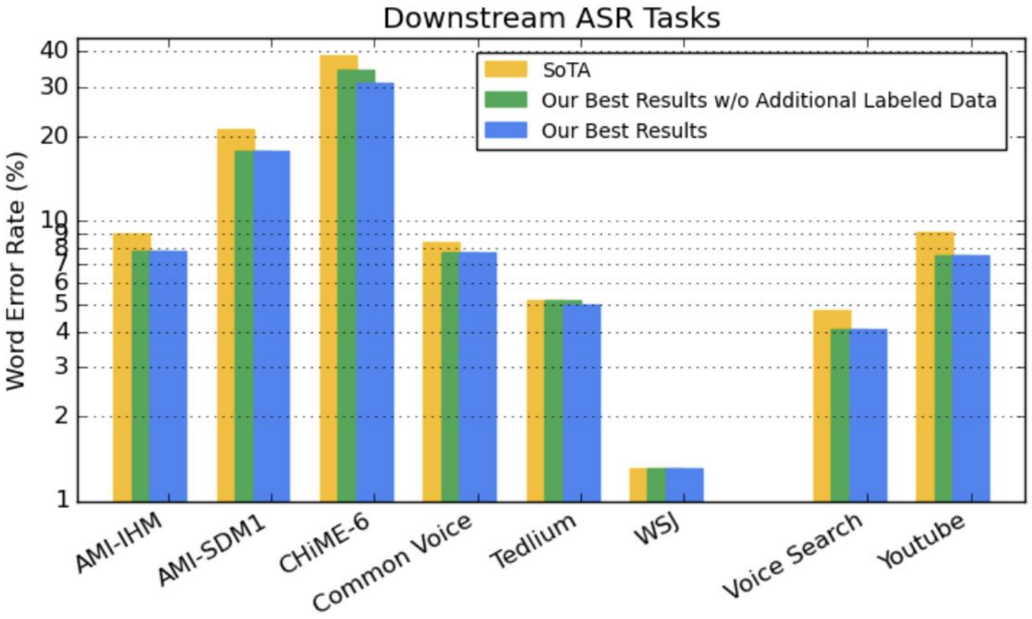
Improvements w.r.t. training strategies, model capability, and labeled dataset size

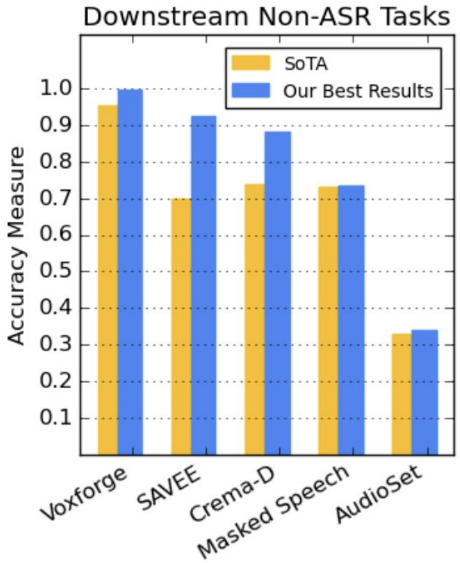




The BigSSL Results: Pre-training, Self-training, and Scaling

Improved ASR and Beyond:





The BigSSL Application: Streaming Model Serving

The fine-tuned PS-model is used to generate NST data for training the student streaming model:

Model	Fine-tuning Mixture			Test-short	Test-long
	Telephony	Video	NST		
Streaming Model					
Baseline	N/A	N/A	N/A	33.11	15.53
Fine-tuned	1.0	-	N/A	22.45	21.41
	0.8	0.2	N/A	22.64	19.99
ConformerXL-RNNT-P	0.8	0.2	N/A	22.24	14.55
ConformerXL-RNNT-PS					
Baseline	N/A	N/A	N/A	27.20	10.97
Fine-tuned	0.8	0.2	N/A	**21.24**	**10.72**
Student Streaming Model	0.8	-	0.2	22.97	16.75

Recap of Section 1:

- Multilingual Self-Supervised Learning (SSL) is just as successful in speech understanding [...*In particular, on an ASR task with 34k hours of labeled data, by fine-tuning an 8 billion parameter pre-trained Conformer model we can match state-of-the-art (SoTA) performance with only* ***3%*** *of the training data and significantly improve SoTA with the full training set...*]

Multilingual Data & Evaluation

2

How can we evaluate pretrained models in over 100+ languages?

Our desiderata for spoken language technologies evaluation:

- **Rich:** N-way parallel permits comparison across languages.
- **Robust:** High domain coverage
- **Realistic:** Natural read-speech, not synthesized
- **Ready:** Standardized splits for train/dev/test

Closest pre-existing datasets:

- **CMU Wilderness** (700 languages, but narrow Bible domain coverage, segments not parallel)
- **CommonVoice** (93 languages with transcripts, but not parallel)

When it comes to Multilingual Speech Data:

Table 1: *Compare FLEURS to common public multilingual speech benchmarks.*

Dataset	#Languages	Total Duration	Domains	Speech Type	Transcripts	Parallel text	Parallel speech
BABEL [13]	17	1k hours	Conversational	Spontaneous	Yes	No	No
CommonVoice [12]	93	15k hours	Open domain	Read	Yes	No	No
CMU Wilderness [15]	700	14k hours	Religion	Read	Yes	Yes	Yes
MLS [8]	8	50.5k hours	Audiobook	Read	Yes	No	No
CoVoST-2 [11]	22	2.9k hours	Open domain	Read	Yes	Yes	No
Voxlingua-107 [14]	107	6.6k hours	YouTube	Spontaneous	No	No	No
Europarl-ST [16]	6	500 hours	Parliament	Spontaneous	Yes	Yes	No
MuST-C [17]	9	385 hours	TED talks	Spontaneous	Yes	Yes	No
mTEDx [18]	9	1k hours	TED talks	Spontaneous	Yes	Yes	No
VoxPopuli [9]	24	400k hours	Parliament	Spontaneous	Partial	Partial	Partial
CVSS [19]	22	1.1k hours	Open domain	Read/Synthetic	Yes	Yes	Yes

When it comes to Multilingual Speech Data:

Table 1: *Compare FLEURS to common public multilingual speech benchmarks.*

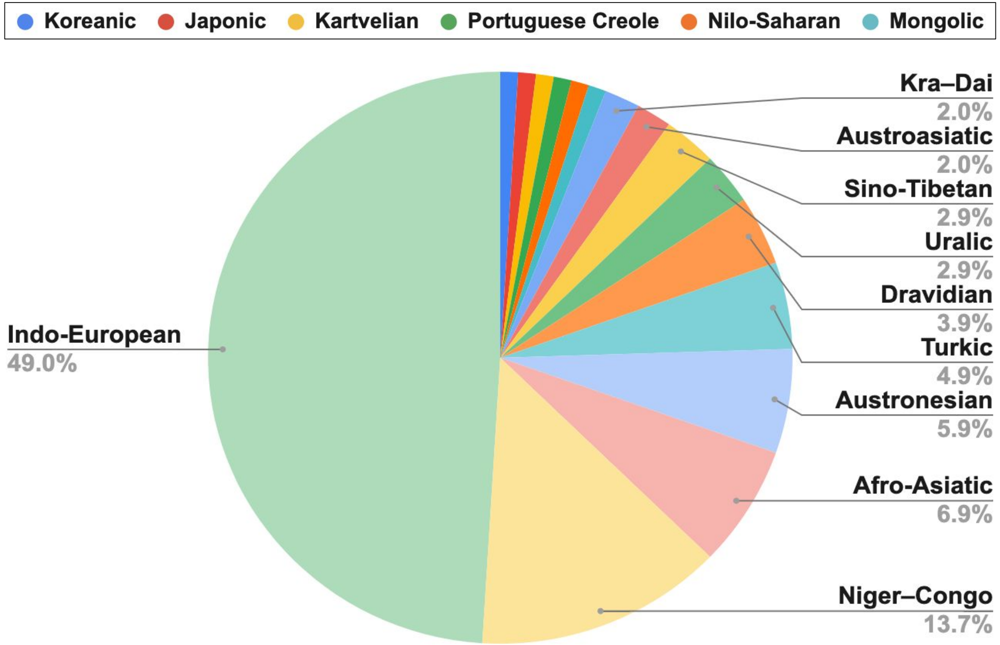
Dataset	#Languages	Total Duration	Domains	Speech Type	Transcripts	Parallel text	Parallel speech
BABEL [13]	17	1k hours	Conversational	Spontaneous	Yes	No	No
CommonVoice [12]	93	15k hours	Open domain	Read	Yes	No	No
CMU Wilderness [15]	700	14k hours	Religion	Read	Yes	Yes	Yes
MLS [8]	8	50.5k hours	Audiobook	Read	Yes	No	No
CoVoST-2 [11]	22	2.9k hours	Open domain	Read	Yes	Yes	No
Voxlingua-107 [14]	107	6.6k hours	YouTube	Spontaneous	No	No	No
Europarl-ST [16]	6	500 hours	Parliament	Spontaneous	Yes	Yes	No
MuST-C [17]	9	385 hours	TED talks	Spontaneous	Yes	Yes	No
mTEDx [18]	9	1k hours	TED talks	Spontaneous	Yes	Yes	No
VoxPopuli [9]	24	400k hours	Parliament	Spontaneous	Partial	Partial	Partial
CVSS [19]	22	1.1k hours	Open domain	Read/Synthetic	Yes	Yes	Yes
FLEURS (this work)	102	1.4k hours	Wikipedia	Read	Yes	Yes	Yes

FLEURS-102

FLEURS-102 is based on FLORES-101.

- **102 language n-way parallel** speech & text dataset
- **~12 hours per language**
- **Open-domain** (text from Wikipedia)
- **Natural read-speech** - no TTS/synthesized speech
- **Balanced speaker gender** distribution across the dataset
- **Pre-split into train/dev/test** with a target ratio 7:1:2
- **Easily accessible** through and

102 Languages across 16 Language Families



- Western European (WE)
- Eastern Europe (EE)
- Central-Asia/Middle-East/North-Africa (CMN)
- Sub-Saharan Africa (SSA)
- South Asia (SA)
- South-East Asia (SEA)
- Chinese, Japanese, and Korean languages(CJK)

Evaluation of Key Tasks enabled by FLEURS

- FLEURS-102 enables evaluation of many spoken language technologies tasks:

Language ID

Speech Recognition

Speech Translation

Retrieval

Automatic Speech Recognition

- Speech-only pre-training outperforms Multimodal pre-training:
  - **12.9 CER** vs **13.4 CER** (w2v-BERT vs mSLAM).
- Languages seen in pre-training yield better results at test time.
- Good performance for zero-shot languages with related languages in pre-training.
  - Malayalam: 8.6 CER
  - Kannada: 7.0 CER
  - Gujarati: 9.3 CER

w2v-BERT: Chung et al., 2021
mSLAM: Bapna et al., 2022

Speech Language Identification

- Multimodal pre-training outperforms speech-only pre-training:
  - **73.3%** vs **71.4%** avg accuracy (mSLAM vs w2v-BERT).

- Most Challenging: South Asian languages!

- Good mSLAM performance for zero-shot languages with related languages in pre-training.
  - **Luxembourgish : 96.90%** (Dutch, French, German in pre-training).
  - **Korean: 95.29%** (Japanese, Cantonese, Mandarin in pre-training).
  - **Filipino: 75%** (Cebuano and Indonesian in pre-training).

Cross-Modal Speech-Text Retrieval

- We present a new task of cross-modal speech-text retrieval
  - **Speech-to-Text Retrieval task**: given audio, retrieve its most probable transcription
  - **Text-to-Speech Retrieval task**: given a text, retrieve the closest audio

- Multimodal models are good at retrieving the correct transcript or speech utterance
  - Precision@1 for Speech to Text Retrieval: **76.9**
  - Precision@1 for Text to Speech Retrieval: **74.4**

The Need for Cross-Lingual Speech Understanding

Design of benchmark:

- **Task Families & Examples:**
  - **Recognition:** Multilingual LibriSpeech (MLS), CommonVoice, BABEL.
  - **Classification:** VoxLingua107 (language identification), VoxCeleb (speaker identification).
  - **Translation:** CoVoST 2 (speech-to-text translation).
  - **Retrieval:** FLEURS-S (speech-to-speech retrieval).
- **Baselines:**
  - The paper establishes the first comprehensive baselines using:
    - **XLS-R:** A large-scale speech-only pre-trained model.
    - **mSLAM:** A model pre-trained on both speech and text data.
- **Accessibility:**
  - All datasets and fine-tuning scripts are readily available through the **Hugging Face platform**.
  - This ease of access encourages wider adoption and faster iteration in the research community.

XTREME-S: The Cross-Lingual Evaluation Benchmark

- XTREME-S evaluates multilingual representation learning in 100+ languages.
- Includes FLEURS: a new 102-language n-way parallel speech dataset.

Speech Recognition

- *FLEURS*: 102 European languages, 1000h for training, read-speech.
- *MLS*: 8 European languages, 10h training, read-speech (books).
- *VoxPopuli*: 16 languages, 500h for high-res, 1-10h for low-resource sessions.

$$0.4 * \left(100 - \frac{\text{Fleurs} + \text{MLS} + \text{VP}}{3}\right)_{\text{(WER)}} +$$

$$0.4 * \text{CoVoST-2}_{\text{(BLEU)}} + 0.2 * \left(\frac{\text{F-LID} + \text{M-14}}{2}\right)_{\text{(Acc)}}$$

Speech Translation

- CoVoST-2: 21 XX->En language pairs, 264h Fr->En, 1h Ja->En, read-speech (Wiki).

Speech Classification

- MInds-14: 14 languages, Intent classification, 50h, commercial system in e-banking.
- *FLEURS*: LangID classification.

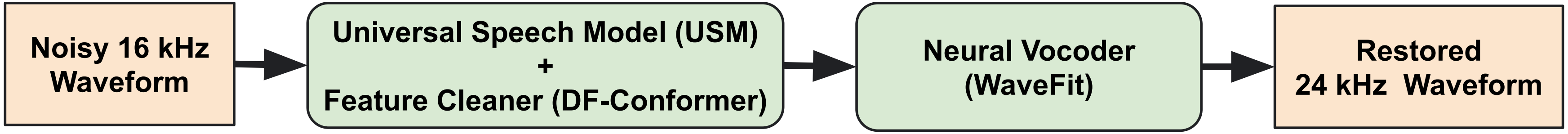
Speech Retrieval

- *FLEURS*: speech-text retrieval.

FLEURS-R

- **Motivations:**
  - **Cleaner Speech**: Natural read-speech, polished by speech restoration algorithm + ASR filter.
  - **Competitive Benchmarks** for multilingual TTS.

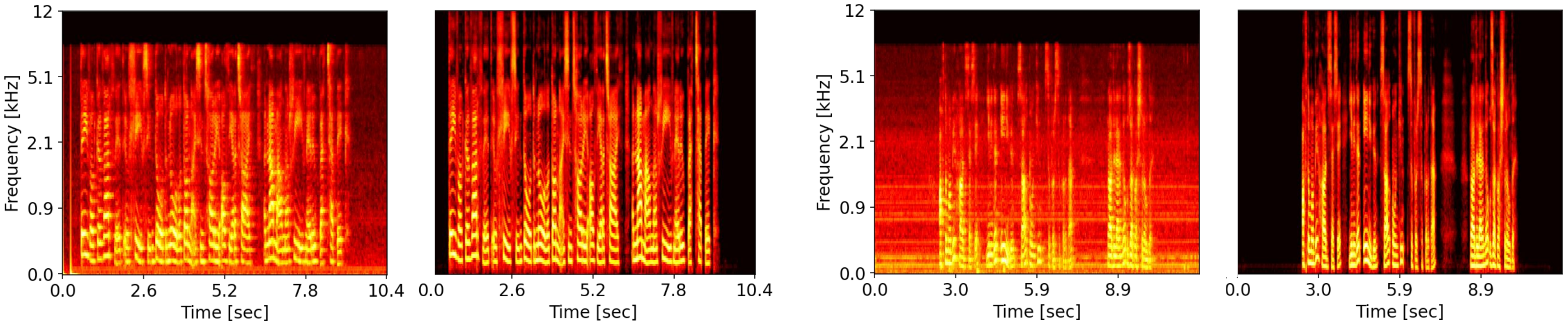
Miipher: Parametric re-synthesis style (TTS-like) speech restoration model



- wav2vec-BERT encoder has been replaced by USM to support multilingual and unknown languages.
  - USM encoder pre-trained over 300 languages, consists of 32 layers (2B params).
- Used intermediate features from 13th layer.
- Feature cleaner and WaveFit trained on 54 locales noisy-clean paired data.
- Not fine-tuned, to preserve *speaker* acoustic characteristics.
- No extra conditionings or extra encoders, *i.e.* no speaker and PnG-BERT text encoders.

Miipher Restoration Performance Visualization

- Before/after examples for unknown languages in FLEURS dataset (unknown: no noisy-clean paired data exists).



Welsh

(Western Europe)

Welsh

(Western Europe)

Azerbaijani

(Central Asia / Middle East / North Africa)

Azerbaijani

(Central Asia / Middle East / North Africa)

Evaluation Metrics

- Semantic Correctness:
  - Evaluated by character error rate (CER).
  - Inference made by a grapheme ASR model (Maestro-U).

- Naturalness of Synthesized Speech:
  - Challenging to obtain human ratings of synthesized speech across 102 languages efficiently and effectively.
  - Evaluated by automatic approximation of Mean Opinion Score, which was predicted by an existing model SQuId.
    - SQuId was pre-trained in 101 text languages and 51 spoken languages, followed by decoder trained on MOS ratings (by humans) in 42 locales.

Experiments

- Better speech restored by Miipher:
  - Maintained CER when decoding FLEURS-R and FLEURS by the same ASR model.
  - Improved SQuId scores for naturalness:
    - 3.72 on FLEURS test.
    - **3.92** on FLEURS-R test
- TTS Benchmark:
  - Trained a TTS model (Virtuoso 2) on original FLEURS speech as a baseline.
  - Trained Virtuoso 2 on FLEURS-R for comparison.
  - Two types of synthesized speech by two TTS models decoded on their test splits:
    - Improved SQuId scores for naturalness:
      - 3.79 by TTS(FLEURS).
      - **3.89** by TTS(FLEURS-R)
    - CER: both got worse CERs, which might due to the ASR model did not adapt to the synthesized data.

Recap of Section 2

- You can use FLEURS-102 for robust multilingual eval for many speech technologies!
  - **Rich:** N-way parallel permits comparison across languages.
  - **Robust:** High domain coverage
  - **Realistic:** Natural read-speech, not synthesized
  - **Ready:** Standardized splits for train/dev/test
- Dataset Available on Hugging Face and Tensorflow

https://huggingface.co/datasets/google/fleurs
https://www.tensorflow.org/datasets/catalog/xtreme_s
https://huggingface.co/datasets/google/fleurs-r

Multimodality for Multilingual ASR and LangID

3

XLS-R: Multilingual Wav2vec 2.0 at Scale

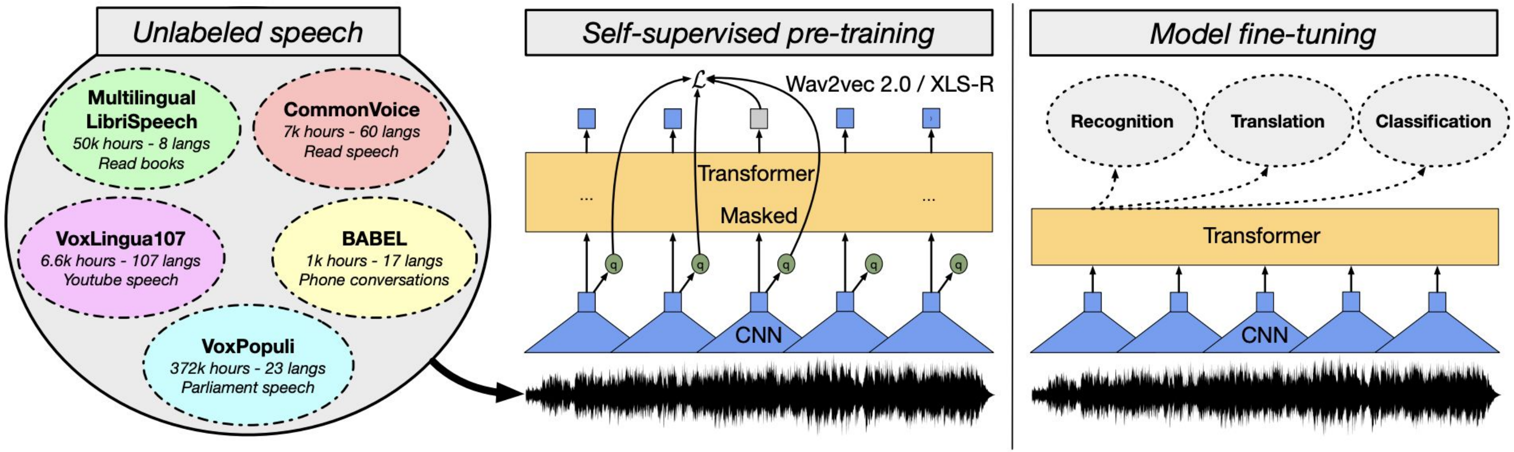


Figure 1: **Self-supervised cross-lingual representation learning.** We pre-train a large multilingual wav2vec 2.0 Transformer (XLS-R) on 436K hours of unannotated speech data in 128 languages. The training data is from different public speech corpora and we fine-tune the resulting model for several multilingual speech tasks.

- Big gains on Translation, Recognition and Classification.

mSLAM: Multilingual Speech+Text Pre-training

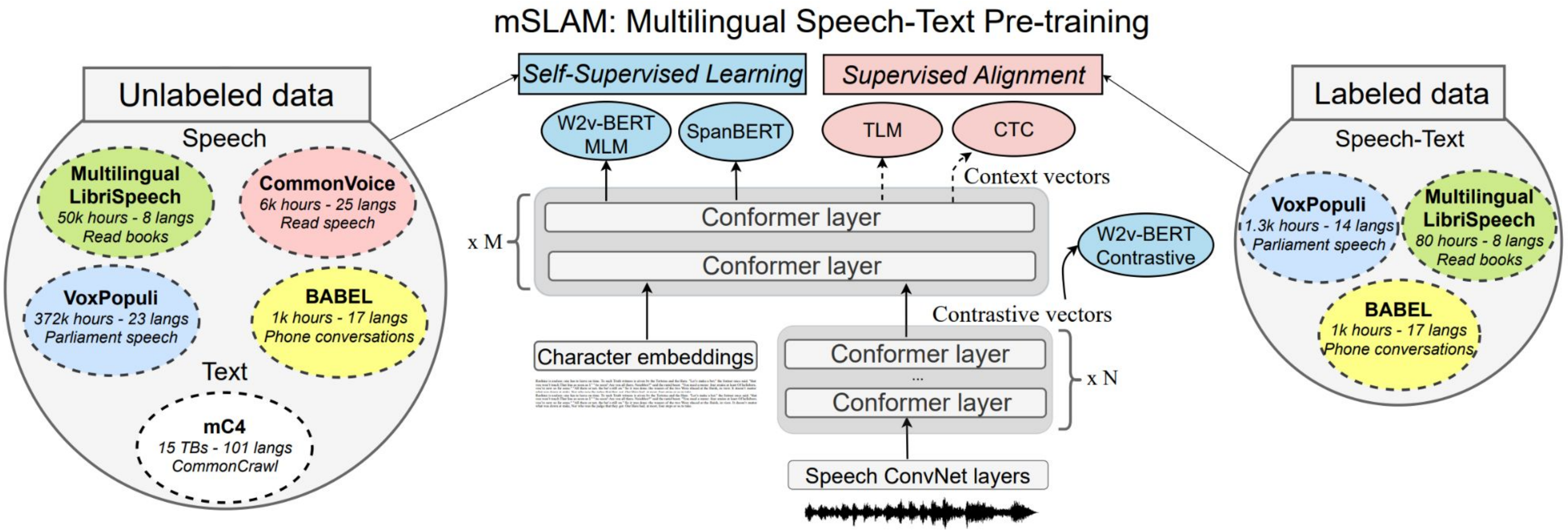


Figure 1: Multilingual Speech-Text Pretraining We pre-train a large multilingual speech-text Conformer on 429K hours of unannotated speech data in 51 languages, 15TBs of unannotated text data in 101 languages, as well as 2.3k hours of speech-text ASR data.

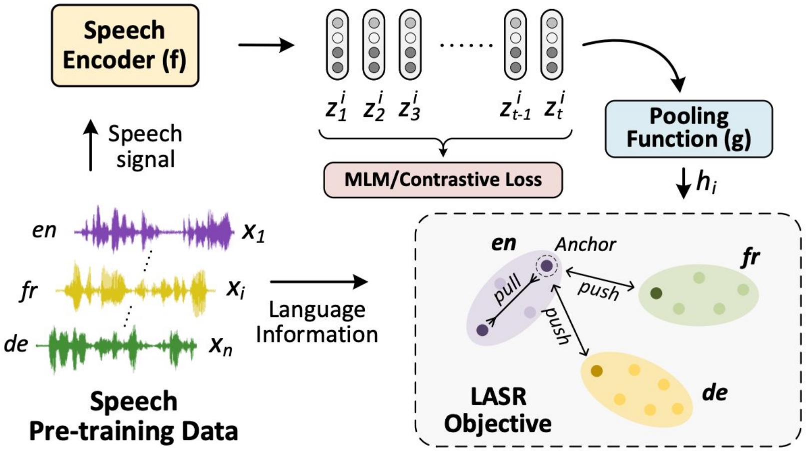
Pre-training Objectives

- Pre-trained to optimize:
  - SpanBERT loss: on unlabeled text, from masked language modeling (MLM) task;
  - w2v-BERT loss: on unlabeled speech, from MLM of discretized acoustic tokens;
  - TLM loss: on paired speech-text, from two MLM tasks to predict masked text or speech, given concatenated speech and text.
  - STM loss: on paired speech-text and non-paired speech-text, from a binary classification to decide if input speech-text is matched or not.

- Gains:
  - Significant improvements on semantic speech tasks Speech translation, Speech intent classification, Speech LangID and text classification tasks.

What if Including LangID into Pre-training?

- Label Aware Speech Representation
- To include both Semantic & Non-Semantic Objectives during Pre-training for Learning Universal Speech Encoders.



$$\mathcal{L}_{\texttt{tri}} = \sum_i \max\left[0, \gamma + d(\boldsymbol{h}_i, \boldsymbol{h}_i^+) - d(\boldsymbol{h}_i, \boldsymbol{h}_i^-)\right], \quad (1)$$

$$\mathcal{L}_{\texttt{hard}} = \sum_i \max[0, \gamma + \max_{j \in i^+} d(\boldsymbol{h}_i, \boldsymbol{h}_j) - \min_{j \in i^-} d(\boldsymbol{h}_i, \boldsymbol{h}_j)] \quad (2)$$

$$\mathcal{L}_{\texttt{LASR}} = \mathcal{L}_{\texttt{SSL}} + \lambda \cdot \mathcal{L}_{\texttt{hard}}. \quad (3)$$

$$\mathcal{L}_{\texttt{ge2e}} = \sum_i 1 - \sigma(\max_{j \in i^+} d(\boldsymbol{h}_i, \boldsymbol{h}_j)) + \sigma(\min_{j \in i^-} d(\boldsymbol{h}_i, \boldsymbol{h}_j)). \quad (4)$$

LASR helps LangID

Method	Accuracy
MLM	75.8
Hard-Triplet	74.3
MLM + Triplet (Eq. 1)	75.1
MLM + GE2E (Contrastive) (Eq. 4)	76.2
MLM + Hard-Triplet (Eq. 2)	80.4

Method	Languages						Avg
	de	en	es	fr	it	nl	
w2v-BERT	4.0	6.2	4.0	4.7	8.9	10.6	7.2
+ LASR	4.0	6.2	4.8	4.8	8.9	10.0	7.2
BEST-RQ	3.9	6.2	3.8	4.8	8.8	9.3	7.0
+ LASR	4.1	6.2	4.3	4.8	9.0	9.6	7.1

$$\mathcal{L}_{\texttt{tri}} = \sum_i \max\left[0, \gamma + d(\boldsymbol{h}_i, \boldsymbol{h}_i^+) - d(\boldsymbol{h}_i, \boldsymbol{h}_i^-)\right], \quad (1)$$

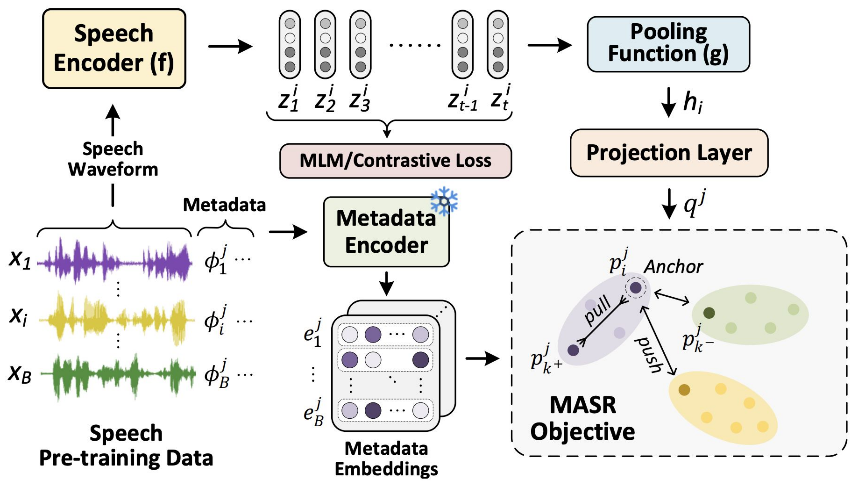
$$\mathcal{L}_{\texttt{hard}} = \sum_i \max[0, \gamma + \max_{j \in i^+} d(\boldsymbol{h}_i, \boldsymbol{h}_j) - \min_{j \in i^-} d(\boldsymbol{h}_i, \boldsymbol{h}_j)] \quad (2)$$

$$\mathcal{L}_{\texttt{LASR}} = \mathcal{L}_{\texttt{SSL}} + \lambda \cdot \mathcal{L}_{\texttt{hard}}. \quad (3)$$

$$\mathcal{L}_{\texttt{ge2e}} = \sum_i 1 - \sigma(\max_{j \in i^+} d(\boldsymbol{h}_i, \boldsymbol{h}_j)) + \sigma(\min_{j \in i^-} d(\boldsymbol{h}_i, \boldsymbol{h}_j)). \quad (4)$$

Generalize LangID Label to Multi-Labels

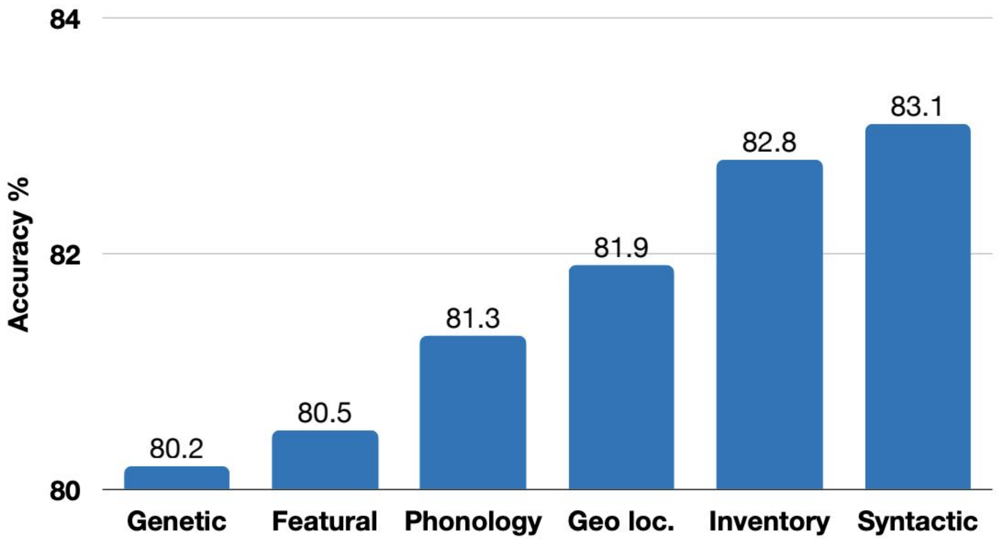
- MASR: MULTI-LABEL AWARE SPEECH REPRESENTATION
- To include multiple meta-labels and soft similarity for the purpose of speech representation learning.



- A metadata encoder: leverages external knowledge resources to generate a representation for each type of metadata.
- A projection layer: integrates multiple types of metadata information.
  - a metadata-specific transformation function.
  - concatenating the metadata encoder and the projection layer outputs.

Various Lang2vec representations for LangID

- Defined by Littell et al., we experimented with syntactic, geographic, phonetic, featural, genetic and inventory based embeddings.



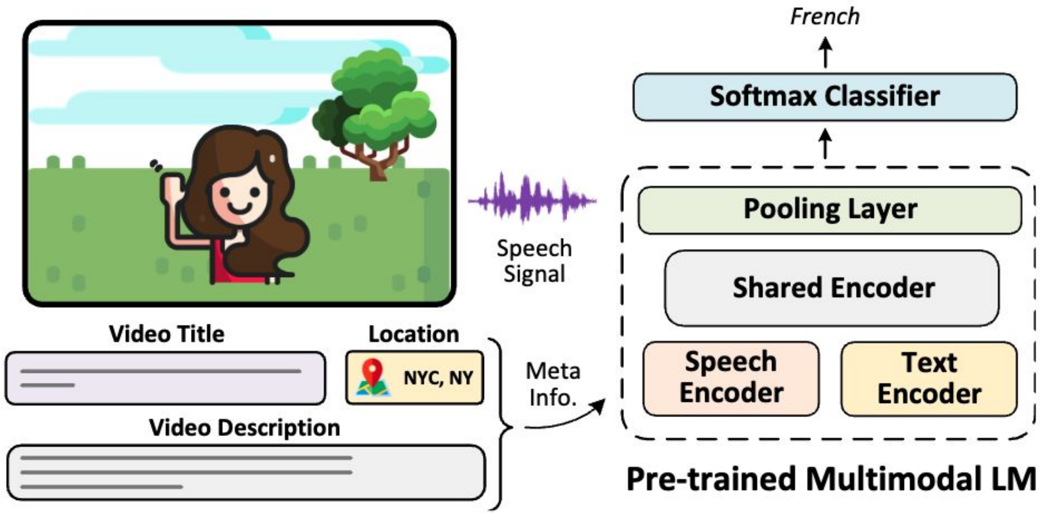
MASR helps LangID and other Non-semantic Tasks

Method	FLEURS					Dhwani				
	O (48)	NO (54)	Overall			O (5)	NO (17)	Overall		
	Acc	Acc	Acc	F1	EER↓	Acc	Acc	Acc	F1	EER↓
w2v-BERT [21]	87.7	69.6	78.0	77.7	0.5	78.8	49.9	58.0	42.6	15.4
+ LASR [19]	88.9	74.3	81.3	80.4	0.5	78.1	52.2	59.5	44.2	**15.2**
+ MASR	**91.4**	**76.3**	**83.4**	**81.3**	**0.4**	**81.0**	**52.4**	**60.7**	**46.2**	15.3
BEST-RQ [10]	85.6	65.2	75.4	72.8	0.9	76.2	46.4	54.7	39.8	16.9
+ LASR [19]	90.6	73.4	81.6	79.7	0.5	77.0	48.6	57.7	43.0	16.1
+ MASR	**91.3**	**74.6**	**83.7**	**81.5**	**0.3**	**80.6**	**50.6**	**59.6**	**44.2**	**15.8**

Method	LangID	Speaker Verification	Emotion Recogition	Audio Classification		
	VoxForge	ASVSpoof2019	Iemocap	Mask Challenge	Esc50-human	Esc50-cough
BEST-RQ [10]	94.6	94.0	54.0	61.1	72.0	90.9
+ LASR	**96.0**	**97.9**	**60.7**	58.1	63.6	87.9
+ MASR	94.0	94.3	60.0	**63.0**	80.3	**94.0**
+ GeoMASR	95.7	96.1	58.8	62.2	**81.8**	92.5
+ TextMASR	90.8	93.2	57.5	61.2	**81.8**	87.9

Massively Multimodal Language Identification

- An end-to-end language identification model that captures signals from multiple modalities.
- Side information were encoded with text encoder.
- +7% improvement over speech language identification.



Thank you.