# Improving Universal Access to Modern Speech Technology


**Martijn Bartelds**

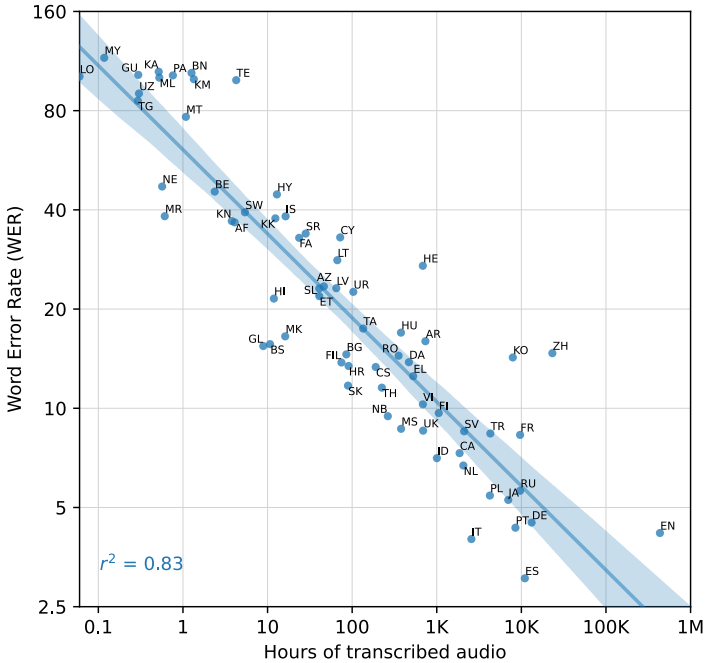
Stanford NLP Group

bartelds@stanford.edu

Increasingly powerful speech models promise “universal” speech processing

**Robust Speech Recognition via Large-Scale Weak Supervision**

Alec Radford [*1] Jong Wook Kim [*1] Tao Xu [1] Greg Brockman [1] Christine McLeavey [1] Ilya Sutskever [1]




"Whisper's speech recognition performance is still quite poor on many languages."
(Radford et al. 2023)

# Scaling Speech Technology to 1,000+ Languages

**Vineel Pratap*** **Andros Tjandra*** **Bowen Shi*** **Paden Tomasello**
**Arun Babu** **Sayani Kundu†** **Ali Elkahky‡** **Zhaoheng Ni**
**Apoorv Vyas** **Maryam Fazel-Zarandi** **Alexei Baevski** **Yossi Adi**
**Xiaohui Zhang** **Wei-Ning Hsu** **Alexis Conneau§** **Michael Auli***

| | Whisper medium | Whisper large-v2 | MMS L-61 noLM | MMS L-61 CC LM | MMS L-61 noLM LSAH | MMS L-61 CC LM LSAH | MMS L-1107 noLM | MMS L-1107 CC LM | MMS L-1107 noLM LSAH | MMS L-1107 CC LM LSAH |
|---|---|---|---|---|---|---|---|---|---|---|
| Amharic | 229.3 | 140.3 | 48.7 | 30.7 | 52.4 | 32.5 | 52.9 | 30.1 | 53.3 | 31.1 |
| Arabic | 20.4 | 16.0 | 34.9 | 19.6 | 35.8 | 19.9 | 44.0 | 23.4 | 41.3 | 21.0 |
| Assamese | 102.3 | 106.2 | 29.5 | 18.8 | 28.4 | 18.6 | 37.6 | 21.2 | 30.5 | 19.2 |
| Azerbaijani | 33.1 | 23.4 | 40.7 | 21.3 | 38.3 | 19.8 | 45.0 | 21.2 | 40.1 | 19.1 |
| Bengali | 100.6 | 104.1 | 19.7 | 11.6 | 20.0 | 12.1 | 25.0 | 12.5 | 23.5 | 12.1 |
| Bulgarian | 21.4 | 14.6 | 23.4 | 13.1 | 23.9 | 13.3 | 27.9 | 12.9 | 25.5 | 13.5 |
| Burmese | 123.0 | 115.7 | 22.2 | 14.2 | 22.3 | 14.5 | 29.2 | 20.2 | 24.5 | 16.0 |
| Catalan | 9.6 | 7.3 | 18.1 | 11.0 | 18.1 | 11.0 | 25.9 | 11.5 | 20.1 | 10.8 |
| Dutch | 9.9 | 6.7 | 26.9 | 13.7 | 26.4 | 14.3 | 38.1 | 14.9 | 27.6 | 14.5 |

Addressing this challenge could improve the digital participation of many speakers worldwide

# What do we need?

Better ways to **reliably measure** speech recognition model performance

# What do we need?

Better ways to **reliably measure** speech recognition model performance

**New methods** for bridging the performance gap between languages

Interspeech 2024
1-5 September 2024, Kos, Greece

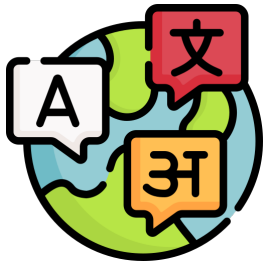
# ML-SUPERB 2.0: Benchmarking Multilingual Speech Models Across Modeling Constraints, Languages, and Datasets

*Jiatong Shi*[1], *Shih-Heng Wang*[2,*], *William Chen*[1,*], *Martijn Bartelds*[3,*], *Vanya Bannihatti Kumar*[1], *Jinchuan Tian*[1], *Xuankai Chang*[1], *Dan Jurafsky*[3], *Karen Livescu*[3,4], *Hung-yi Lee*[2], *Shinji Watanabe*[1]

[1] Carnegie Mellon University, [2] National Taiwan University, [3] Stanford University,
[4] Toyota Technological Institute at Chicago

# Background: Multilingual Speech Processing Benchmark

- Recent multilingual speech processing models
  - Have the capacity to model **hundreds of languages**



# Background: Multilingual Speech Processing Benchmark

- Recent multilingual speech processing models
  - Have the capacity to model **hundreds of languages**
  - However, they are often evaluated using different setups, which **limits the extent to which they can be reliably compared**

# Background: Multilingual Speech Processing Benchmark

- Recent multilingual speech processing models
  - Have the capacity to model **hundreds of languages**
  - However, they are often evaluated using different setups, which **limits the extent to which they can be reliably compared**

→This motivates the need for **multilingual speech processing benchmarks**

# Background: Multilingual Speech Processing Benchmark

We observe great efforts in the community on spoken multilingual benchmarks:

- XTREME-S (Conneau et al. 2022)
- IndicSUPERB (Javed et al. 2023)
- ML-SUPERB (Shi et al. 2023)

# Background: Multilingual Speech Processing Benchmark

- We observe great efforts in the community on spoken multilingual benchmarks:
  - XTREME-S (Conneau et al. 2022)
  - IndicSUPERB (Javed et al. 2023)
  - ML-SUPERB (Shi et al. 2023)

- ML-SUPERB is the most comprehensive benchmark in terms of language coverage, as it includes **143** languages and it evaluates models on:
  - Monolingual/multilingual automatic speech recognition (ASR)
  - Language identification (LID)
  - Joint ASR + LID

# Limitations of ML-SUPERB

- Strictly constrained benchmark settings with self-supervised learning (SSL) pre-trained models
  - Efficient yet not generalizable enough to various settings (Zaiem et al. 2023; Arora et al. 2024)
  - Does not take application requirements or users' budgets into account
- This motivates benchmarking with more **flexible constraints**

# Limitations of ML-SUPERB

- Evaluation metric does not provide insight into performance variations between individual languages and datasets
- This motivates changes to the **evaluation metrics** to place greater focus on **robustness** across languages and datasets

# Introduction of ML-SUPERB 2.0

- We revisit ML-SUPERB:
  - By **relaxing its fixed constraints**
  - By **improving fairness in its evaluation metrics** to focus on **robustness** across languages and **variation** across datasets

# Experimental Design (General Setup)

- ML-SUPERB 2.0 evaluates joint multilingual LID/ASR
- We updated the ML-SUPERB dataset by **correcting** some mistakes*
- Some statistics:
  - 141 languages across 15 datasets
  - Around 300 hours in total (with 85 hours for validation + test sets)
  - We follow the 1-hour configuration presented in ML-SUPERB
  - 20 languages are reserved for few-shot learning experiments, each using 5 utterances for training

* Please refer to our paper for details about the updates to the dataset

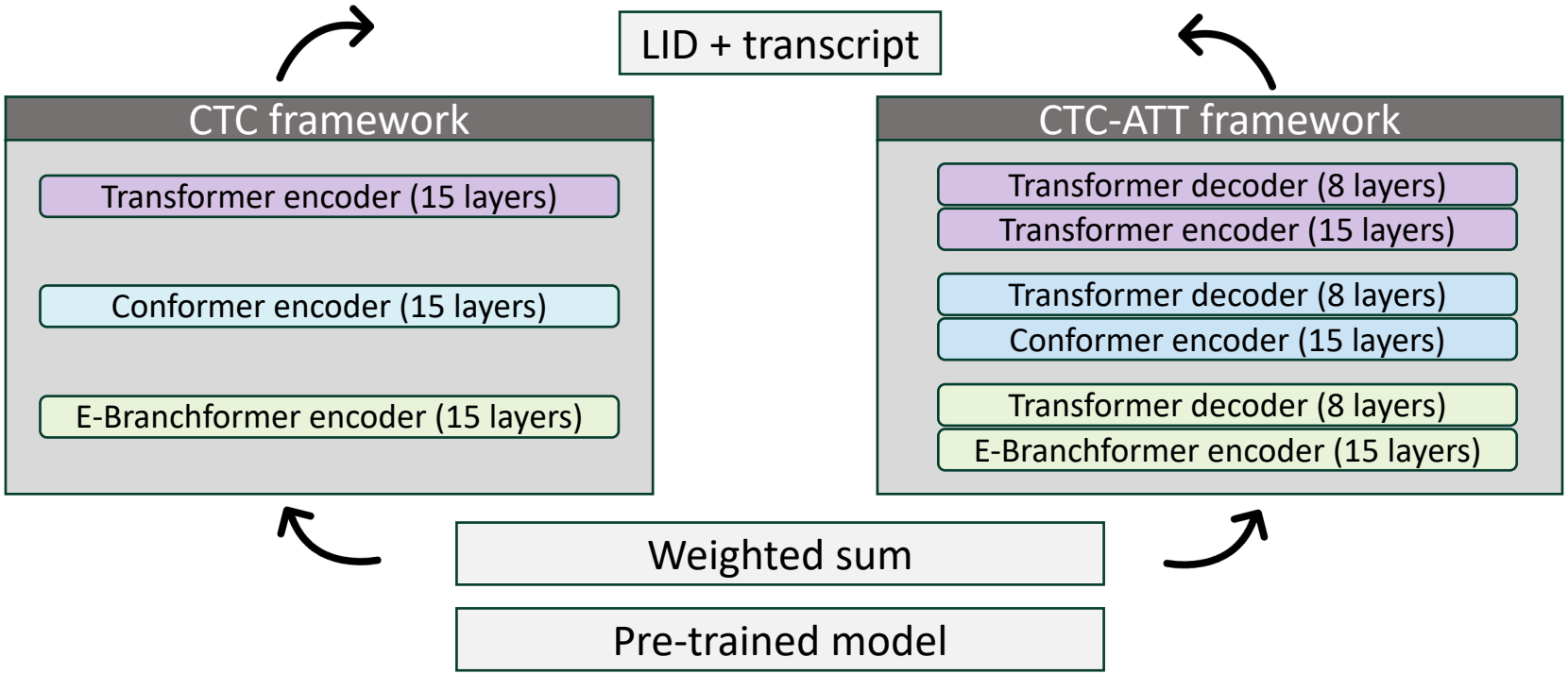
# Experimental Design (General Setup)

- Experimental codebases:
  - ESPnet (Watanabe et al. 2018)
  - S3PRL (Yang et al. 2021)
- Selected pre-trained self-supervised models:
  - XLS-R (Babu et al. 2022)
  - MMS (Pratap et al. 2024)
- In line with the original ML-SUPERB:
  - Limit the number of tunable parameters to 100 million

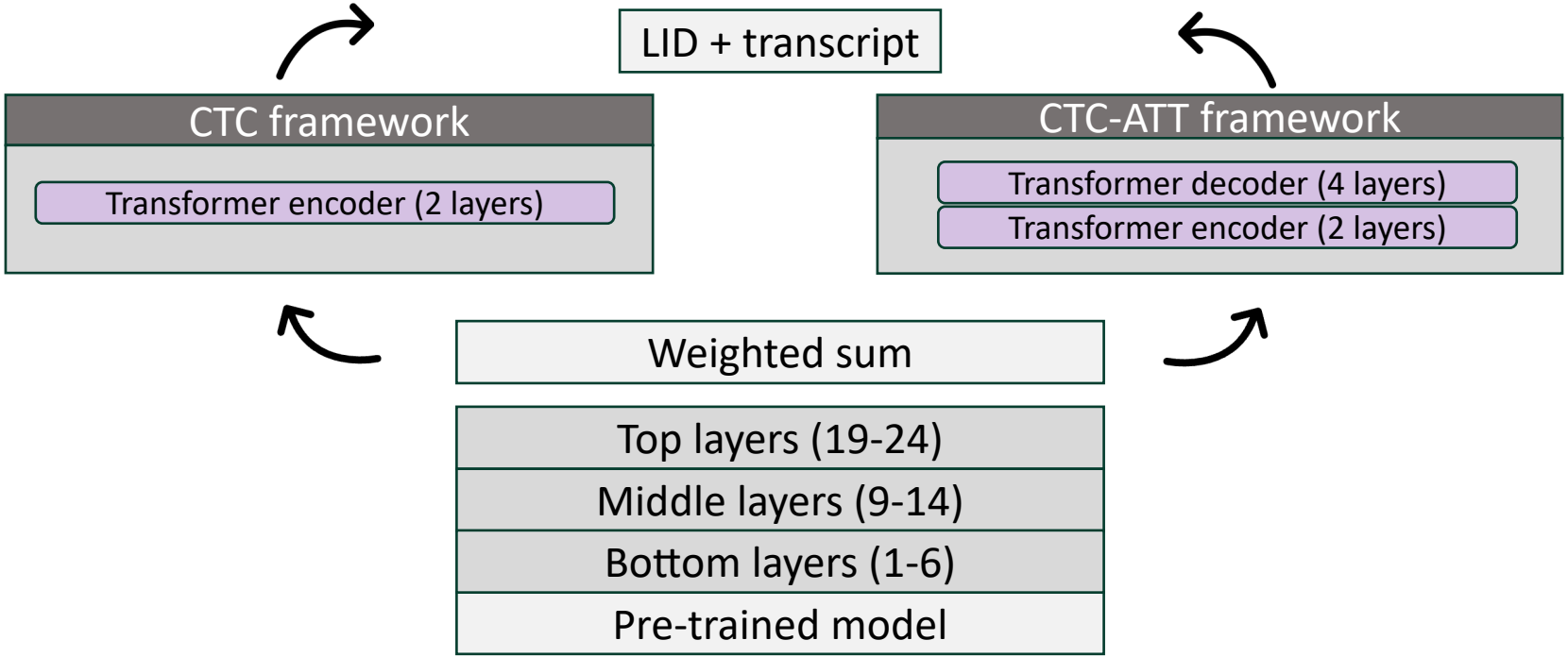
# Experimental Design (General Setup)

- Specifically, we investigate **four new benchmark configurations**:
  - Larger-scale downstream models
  - SSL model fine-tuning
  - Efficient model adaptation strategies
  - Supervised pre-trained models

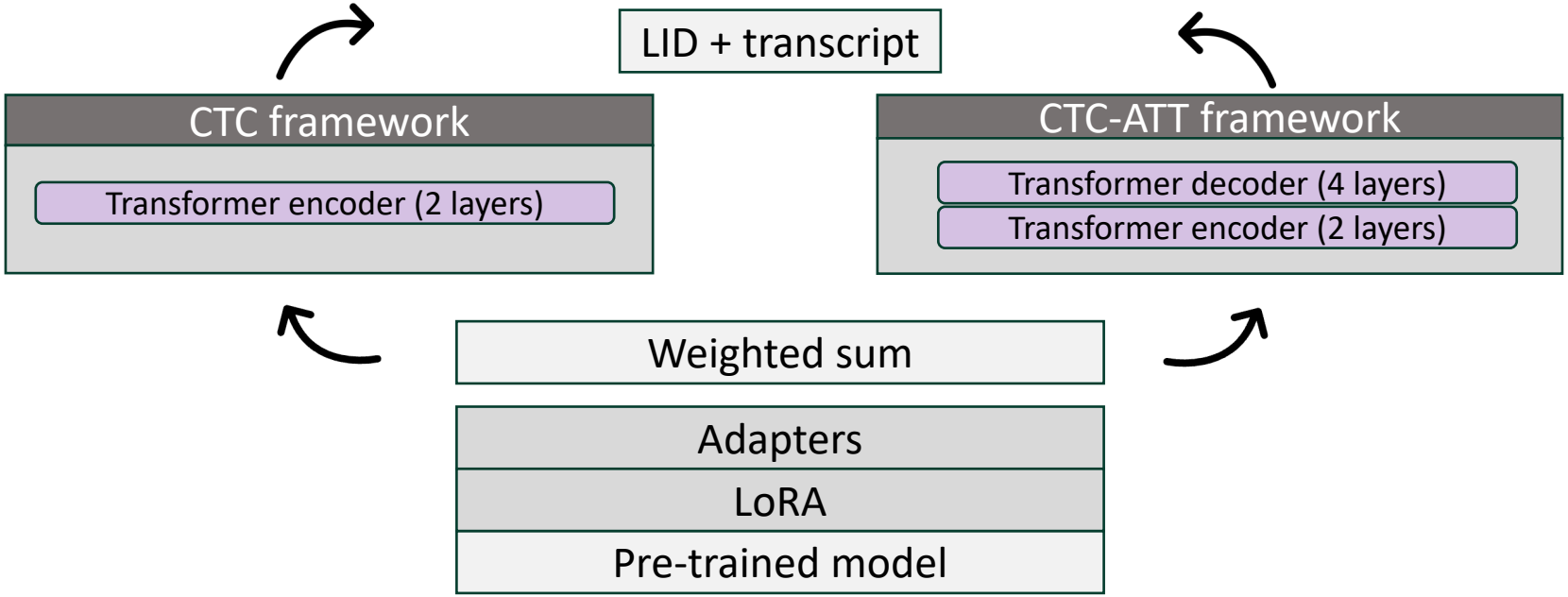
# Larger-scale downstream models



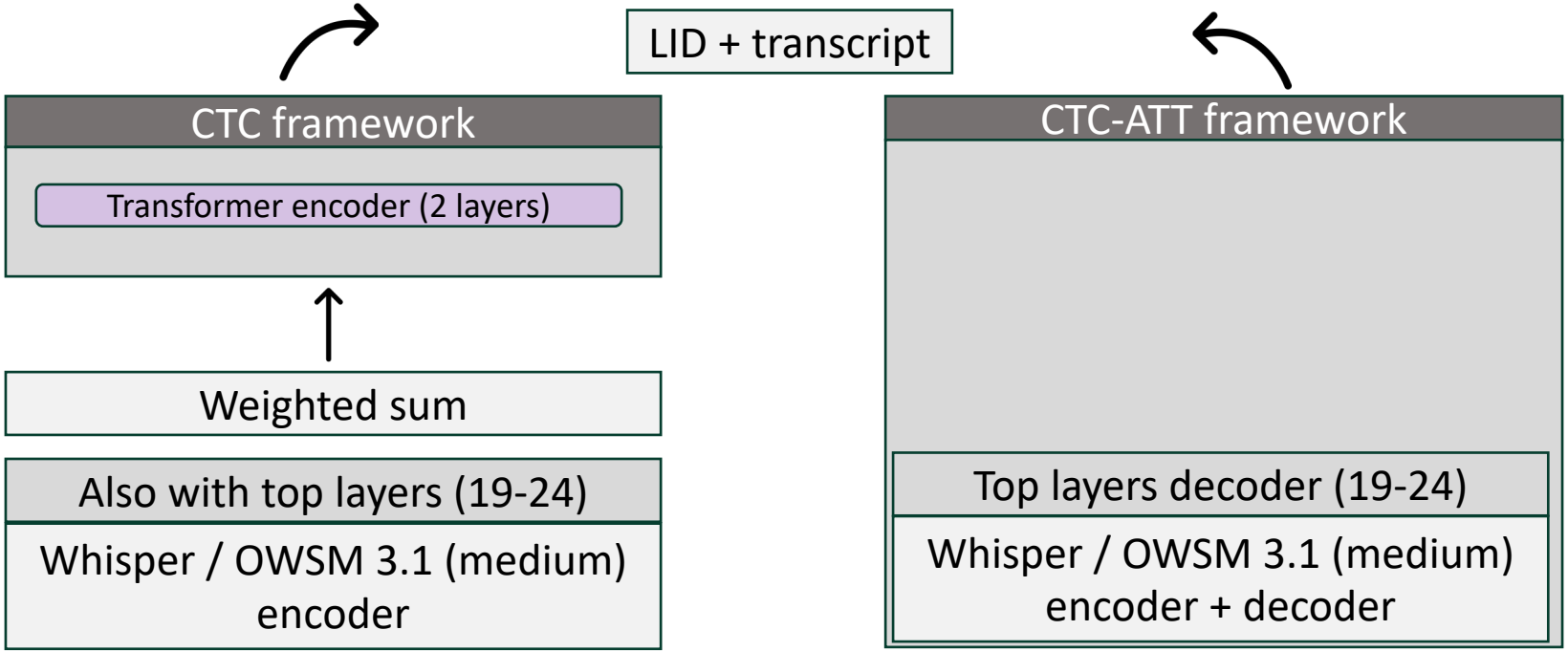
# SSL model fine-tuning



# Efficient model adaptation strategies



# Supervised pre-trained models



# Experimental Design (Configuration Setup)

- For the four benchmark configurations:
  - Hyperparameters follow prior works*
  - We tune the learning rate and select the best-performing model on the validation set

* Please refer to our paper for the complete list of prior works we refer to.

# Experimental Design (Evaluation)

- Base metrics:
  - Accuracy for LID
  - Character error rate (CER) for ASR on two sets (normal and few-shot setting)

# Experimental Design (Evaluation)

- **Place greater focus on measuring robustness:**
  - Macro-average over languages/datasets instead of micro-average CER
    - Compute per-language CER as the macro-average of CERs across all datasets per language
    - Compute the macro-average of the per-language CERs

    → Allows to better understand variation between languages and datasets

    → Languages with more samples do not disproportionally affect the CER

  - Standard deviation of language-specific CERs
  - Measure CER of the worst-performing language
  - Measure CER range between datasets in the same language

# Experimental Results and Discussions

- Effect of introducing four benchmark configurations
- Model ranking for the benchmark configurations
- Supervised ASR versus SSL pre-trained models
- Variation across languages and datasets

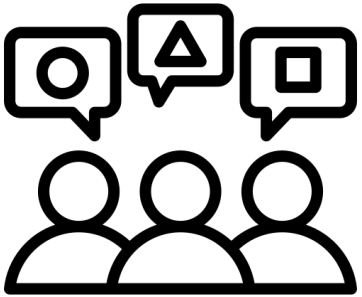
Due to the time limits, we present part of results in the presentation. Please refer to our paper for the full details.

# Effect of Introducing Four Configurations

| Configurations | Details | Accuracy | CER (Normal) |
|---|---|---|---|
| **Original ML-SUPERB** | MMS + Transformer CTC | 90.3 | 24.7 ± 12.3 |
| **Larger Downstream** | MMS + E-Branchformer ATT-CTC | 95.2 | 16.6 ± 11.8 |
| **SSL Model Fine-tuning** | MMS + 9-14 layers partial fine-tuning CTC | **95.6** | **15.5 ± 10.3** |
| **Efficient Model Adaptation** | MMS + LoRA + Transformer ATT-CTC | 94.2 | 18.7 ± 11.5 |
| **Supervised Pre-trained Model** | Whisper Encoder + Transformer CTC | 91.7 | 21.0 ± 12.5 |

Compared to the original ML-SUPERB, we observe **better performance** for LID and ASR across **ALL configurations** (normal setting)

# Model Ranking given Different Configurations



- ML-SUPERB 2.0 is a **better estimate** of model performance compared to the original ML-SUPERB
- However, when considering **different** training settings, the ranking of upstream models can be **different**

# Model Ranking given Different Configurations (Larger-scale Downstream Models)

| | Transformer | Conformer | E-Branchformer |
|---|---|---|---|
| **CTC** | XLS-R | MMS | XLS-R |
| **ATT-CTC** | MMS | MMS | MMS |

XLS-R wins

MMS wins

Compared to the original ML-SUPERB, the performance of XLS-R and MMS **depends on the choice of the downstream model**

# Model Ranking given Different Configurations (Model Fine-tuning)

| | Bottom | Middle | Top |
|---|---|---|---|
| **CTC** | MMS | MMS | MMS |
| **ATT-CTC** | MMS | MMS | MMS |

XLS-R wins

MMS wins

Compared to the downstream model configuration,
XLS-R and MMS **rank differently** when considering fine-tuning approaches

# Model Ranking given Different Configurations (Efficient Model Adaptation)

| | LoRA | Adapter |
|---|---|---|
| **CTC** | XLS-R | XLS-R |
| **ATT-CTC** | MMS | XLS-R |

XLS-R wins

MMS wins

Compared to previous experimental settings,
XLS-R and MMS **rank differently** when considering efficient model adaptation approaches

There is no single way to evaluate an SSL model.
It must always be measured in the context of
a specific downstream model and task

# Supervised ASR vs. SSL Pre-trained Models

- Original ML-SUPERB only focuses on SSL pre-trained models
- ML-SUPERB 2.0 also allows the use of supervised ASR models
  - As long as the test sets from the ML-SUPERB 2.0 dataset are not used in training
- In our paper, we introduce some preliminary analysis on the comparison between supervised ASR and SSL pre-trained models

# Supervised ASR vs. SSL Pre-trained Models

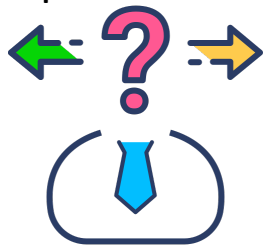
| Pre-trained Model (Module) | Downstream Learning Modules | Accuracy | CER (Normal) |
|---|---|---|---|
| **XLS-R** | Additional transformer encoder + CTC prediction head | 93.7 | 20.7 ± 10.8 |
| **MMS** | Additional transformer encoder + CTC prediction head | 93.6 | 21.0 ± 11.2 |
| **Whisper Encoder** | Additional transformer encoder + CTC prediction head | 91.7 | 21.0 ± 12.5 |
| **Whisper Encoder** | Partial parameters in Whisper encoder (top layers) and additional transformer encoder + CTC prediction head | 83.9 | 26.8 ± 15.0 |
| **Whisper Encoder + Decoder** | Partial parameters in Whisper decoder (top layers) | 85.5 | 25.6 ± 19.4 |

In our experiments, SSL pre-trained models demonstrate slightly **superior performance** compared to supervised ASR pre-trained models

# Variation across Languages and Datasets

- Large standard deviations in both normal and few shot settings

  → This shows that there is **substantial variation** among the language-specific CERs

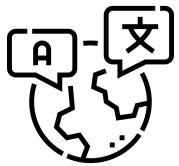


# Variations across Languages and Datasets

- Large standard deviations in both the normal and few-shot settings

  → This shows that there is **substantial variation** among the language-specific CERs

- The impact of language differences is also highlighted by the CER of the worst-performing languages
  - In most cases, Lao or Min Nan Chinese have a CER > 60%



# Variations across Languages and Datasets

- Large standard deviations in both the normal and few-shot settings
  → This shows that there is **substantial variation** among language-specific CERs
- The large impact of language differences is also highlighted by the CER of the worst-performing languages
  - In most cases, Lao or Min Nan Chinese have a CER > 60%
- Large CER differences between datasets in the same language
  → This highlights the **impact of domain or acoustic differences**

# Findings of ML-SUPERB 2.0

- All four configurations **show improvements** over the configuration used in the original ML-SUPERB, which was likely underestimating model performance
- Model ranking depends on the configuration of the benchmark
- There is no single way to evaluate an SSL model. It must always be measured in the context of a specific downstream model and task

- We encourage research on methods that improve language/dataset robustness

Can we develop methods to address the performance gap?

Many languages have limited prospect of obtaining more transcribed speech data

Can standard data augmentation approaches improve low-resource speech recognition performance using real-world data?

Can standard data augmentation approaches improve low-resource speech recognition performance using real-world data?

Self-training and TTS-generated speech

# XLST: Cross-lingual Self-training to Learn Multilingual Representation for Low Resource Speech Recognition

Zi-Qiang Zhang, Yan Song, Ming-Hui Wu, Xin Fang, Li-Rong Dai



## CONTINUOUS PSEUDO-LABELING FROM THE START

**Dan Berrebbi***
Carnegie Mellon University
dberrebb@andrew.cmu.edu

**Ronan Collobert, Samy Bengio,**
**Navdeep Jaitly, Tatiana Likhomanenko**
Apple
{collobert,bengio,njaitly,antares}@apple.com

## GENERATING SYNTHETIC AUDIO DATA FOR ATTENTION-BASED SPEECH RECOGNITION SYSTEMS

*Nick Rossenbach, Albert Zeyer, Ralf Schlüter, Hermann Ney*

Human Language Technology and Pattern Recognition, Computer Science Department,
RWTH Aachen University, 52074 Aachen, Germany
AppTek GmbH, 52062 Aachen, Germany
<surname>@i6.informatik.rwth-aachen.de

## PSEUDO-LABELING FOR MASSIVELY MULTILINGUAL SPEECH RECOGNITION

*Loren Lugosch*[1*], *Tatiana Likhomanenko*[2†], *Gabriel Synnaeve*[2], *Ronan Collobert*[2†]

[1]McGill University / Mila, [2]Facebook AI Research

## SPEAKER AUGMENTATION FOR LOW RESOURCE SPEECH RECOGNITION

*Chenpeng Du, Kai Yu*

MoE Key Lab of Artificial Intelligence
SpeechLab, Department of Computer Science and Engineering
Shanghai Jiao Tong University, China
{duchenpeng, kai.yu}@sjtu.edu.cn

## SPEECH RECOGNITION WITH AUGMENTED SYNTHESIZED SPEECH

*Andrew Rosenberg, Yu Zhang, Bhuvana Ramabhadran, Ye Jia, Pedro Moreno, Yonghui Wu, Zelin Wu*

Google
{rosenberg, ngyuzh, bhuv, jiaye, pedro, yonghui, zelinwu}@google.com

## SELF-TRAINING FOR END-TO-END SPEECH RECOGNITION

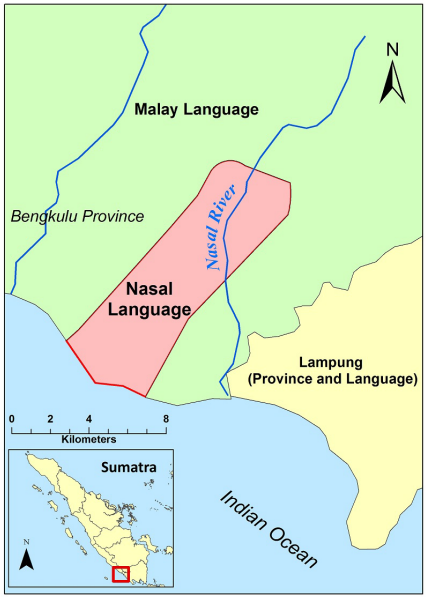
*Jacob Kahn, Ann Lee, Awni Hannun*

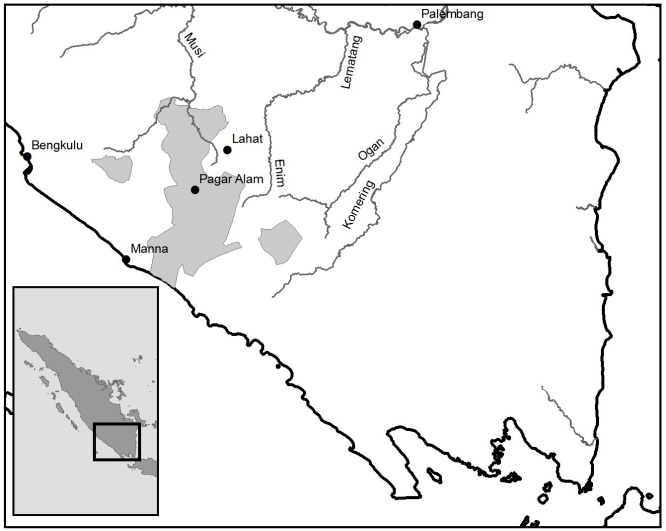
Facebook AI Research

## MAGIC DUST FOR CROSS-LINGUAL ADAPTATION OF MONOLINGUAL WAV2VEC-2.0

*Sameer Khurana*[1], *Antoine Laurent*[2], *James Glass*[1]

[1]MIT Computer Science and Artificial Intelligence Laboratory, Cambridge, MA, USA
[2]LIUM - Le Mans University, France


Images: https://en.wikipedia.org/wiki/Low_German (modified), Anderbeck, K., & Aprilani, H. (2013). The improbable language: Survey report on the Nasal language of Bengkulu, Sumatra. *SIL Electronic Survey Report*, *12*, McDonnell, B. J. (2016). *Symmetrical voice constructions in Besemah: A usage-based approach*. University of California, Santa Barbara.

Improving


West-Frisian: ± 875,000 Gronings: ± 260,000

Images: https://en.wikipedia.org/wiki/Low_German (modified), Anderbeck, K., & Aprilani, H. (2013). The improbable language: Survey report on the Nasal language of Bengkulu, Sumatra. *SIL Electronic Survey Report*, *12*, McDonnell, B. J. (2016). *Symmetrical voice constructions in Besemah: A usage-based approach*. University of California, Santa Barbara.


**Indo-European languages**

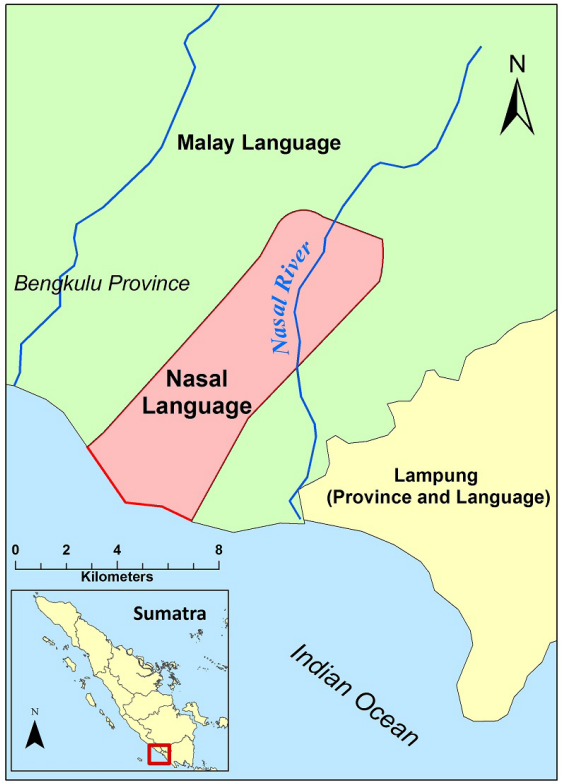
They have regional lexical, grammatical and acoustic variation

**Language use and characteristics**

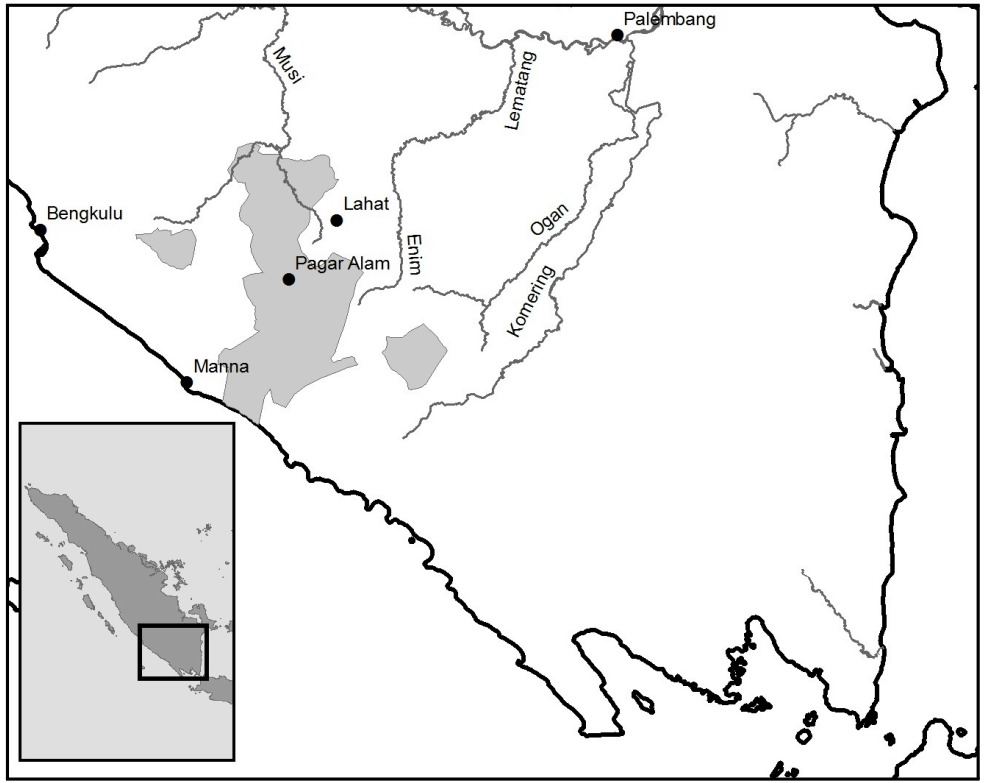
Speakers of both languages speak Dutch and are not mutually intelligible

Images: https://en.wikipedia.org/wiki/Low_German (modified), Anderbeck, K., & Aprilani, H. (2013). The improbable language: Survey report on the Nasal language of Bengkulu, Sumatra. *SIL Electronic Survey Report*, *12*, McDonnell, B. J. (2016). *Symmetrical voice constructions in Besemah: A usage-based approach*. University of California, Santa Barbara.

Improving

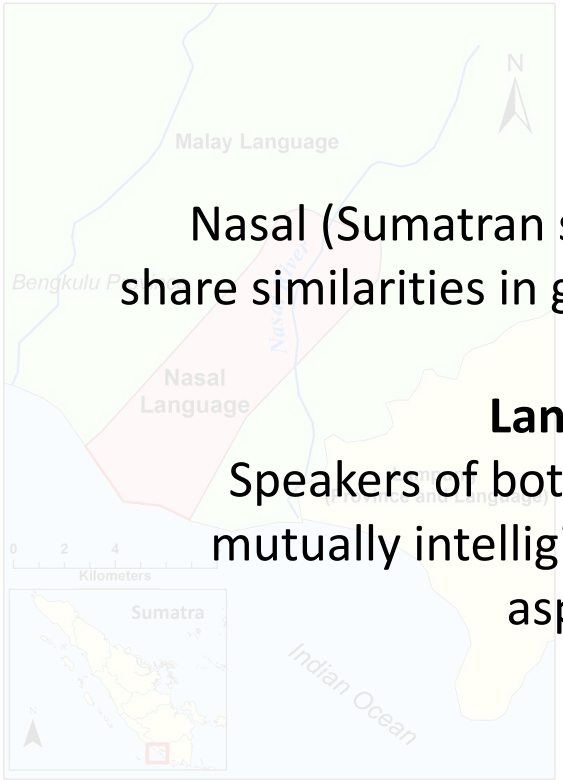



Nasal: ± 3,000

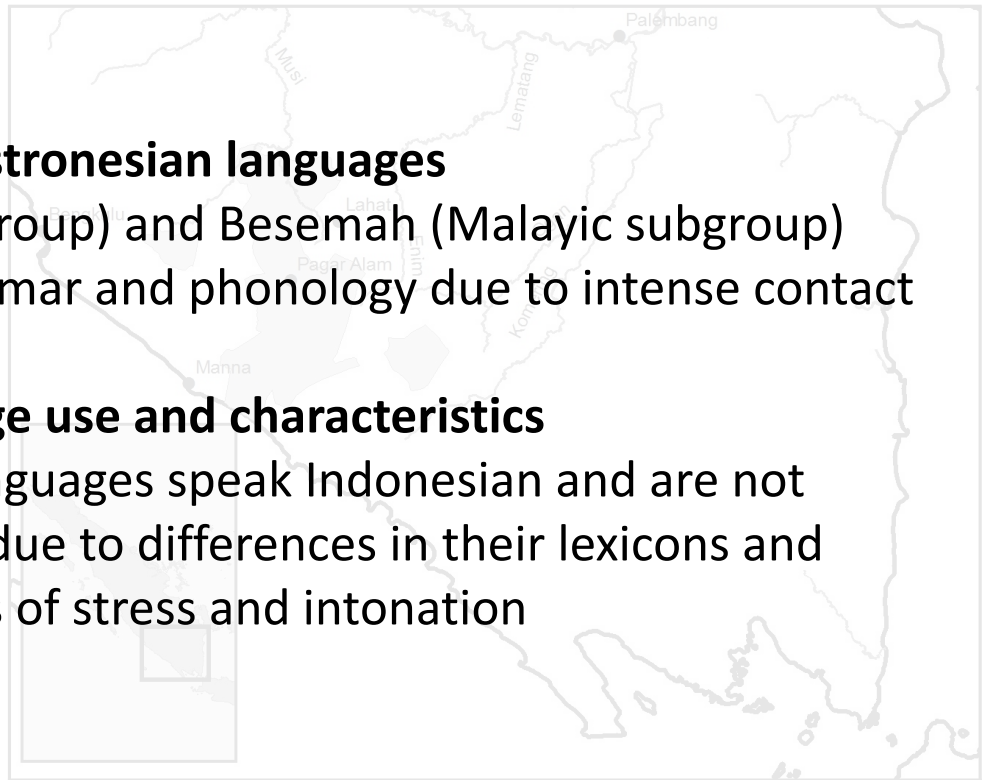



Besemah: ± 500,000

Images: https://en.wikipedia.org/wiki/Low_German (modified), Anderbeck, K., & Aprilani, H. (2013). The improbable language: Survey report on the Nasal language of Bengkulu, Sumatra. *SIL Electronic Survey Report*, *12*, McDonnell, B. J. (2016). *Symmetrical voice constructions in Besemah: A usage-based approach*. University of California, Santa Barbara.




Nasal: ± 3,000




Besemah: ± 500,000

## Austronesian languages

Nasal (Sumatran subgroup) and Besemah (Malayic subgroup) share similarities in grammar and phonology due to intense contact

## Language use and characteristics

Speakers of both languages speak Indonesian and are not mutually intelligible due to differences in their lexicons and aspects of stress and intonation

Images: https://en.wikipedia.org/wiki/Low_German (modified), Anderbeck, K., & Aprilani, H. (2013). The improbable language: Survey report on the Nasal language of Bengkulu, Sumatra. *SIL Electronic Survey Report*, *12*, McDonnell, B. J. (2016). *Symmetrical voice constructions in Besemah: A usage-based approach*. University of California, Santa Barbara.

**4 hours per language**

80 % train (192 mins) / 10 % dev (24 mins) / 10 % test (24 mins)

**4 hours per language**

80 % train (192 mins) / 10 % dev (24 mins) / 10 % test (24 mins)

**TTS synthesis**

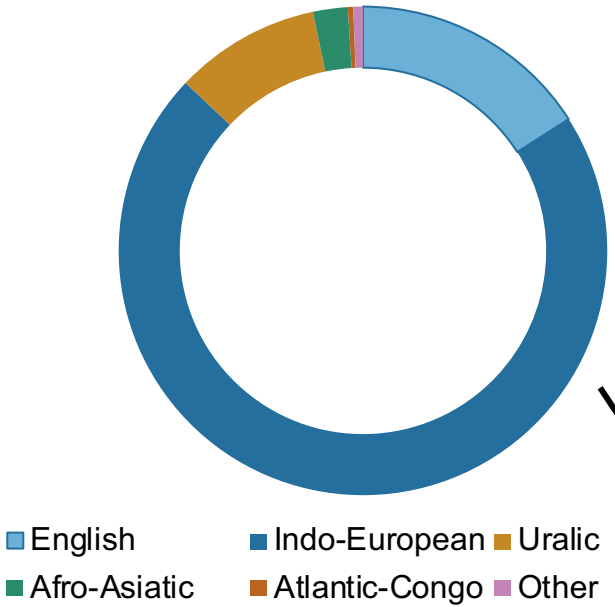
Only for Gronings

**Impact of additional training data**

+8 hours for Gronings

# XLS-R

436,000 hours in 128 languages



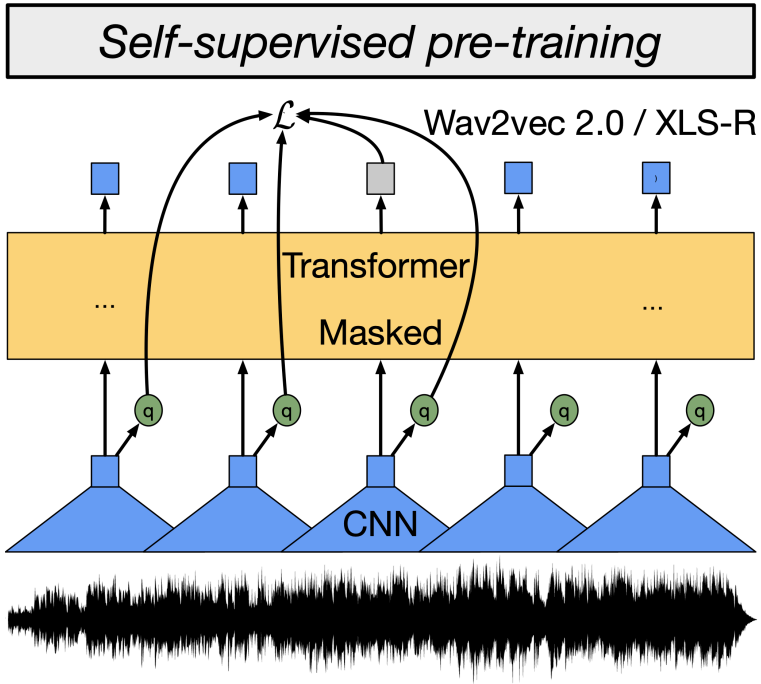




Image: Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2021. XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale.

# Effect of data size

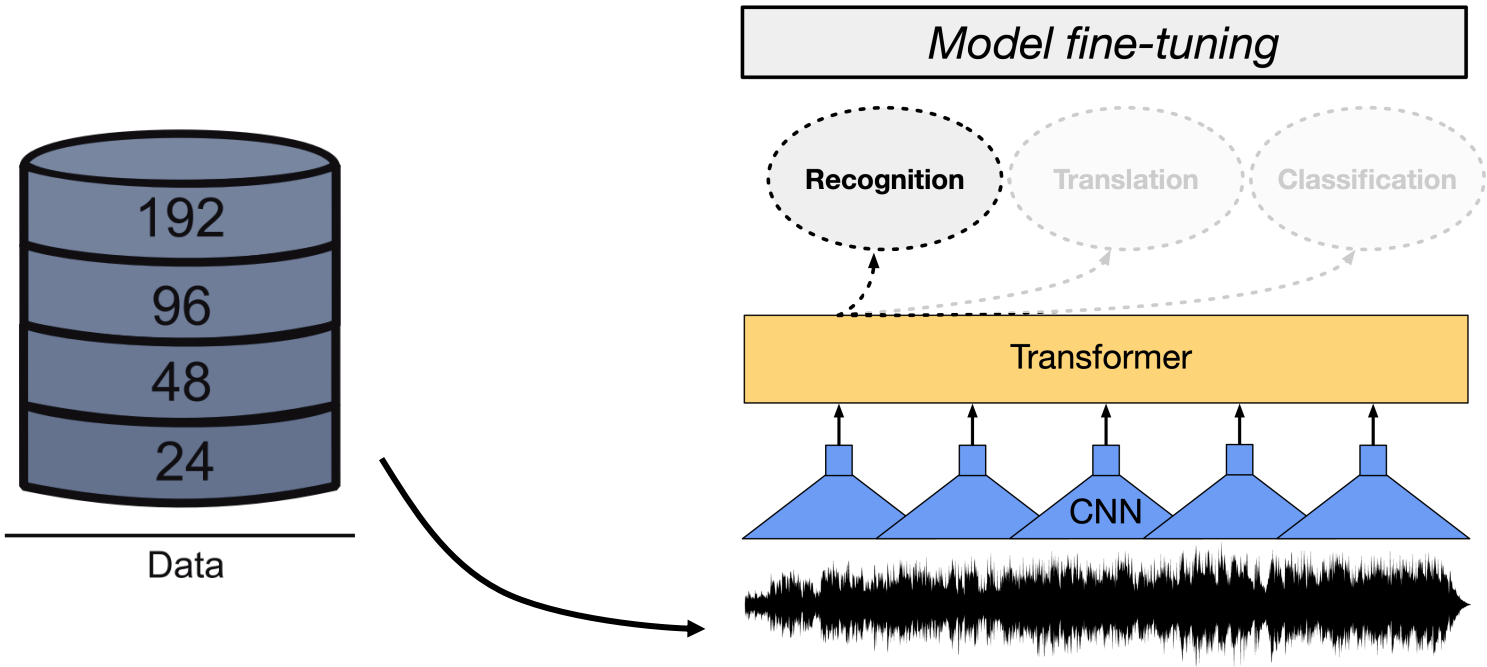
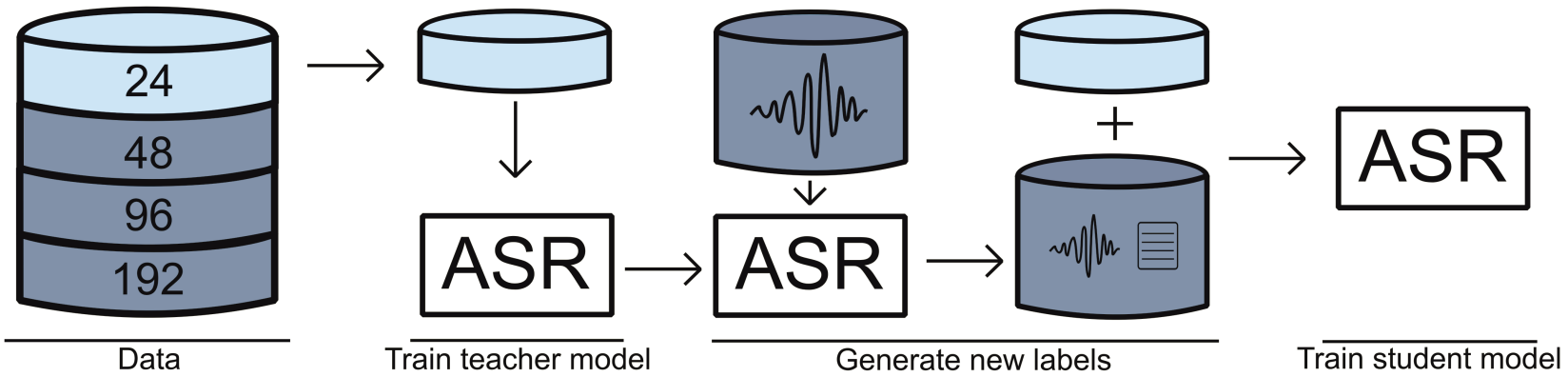
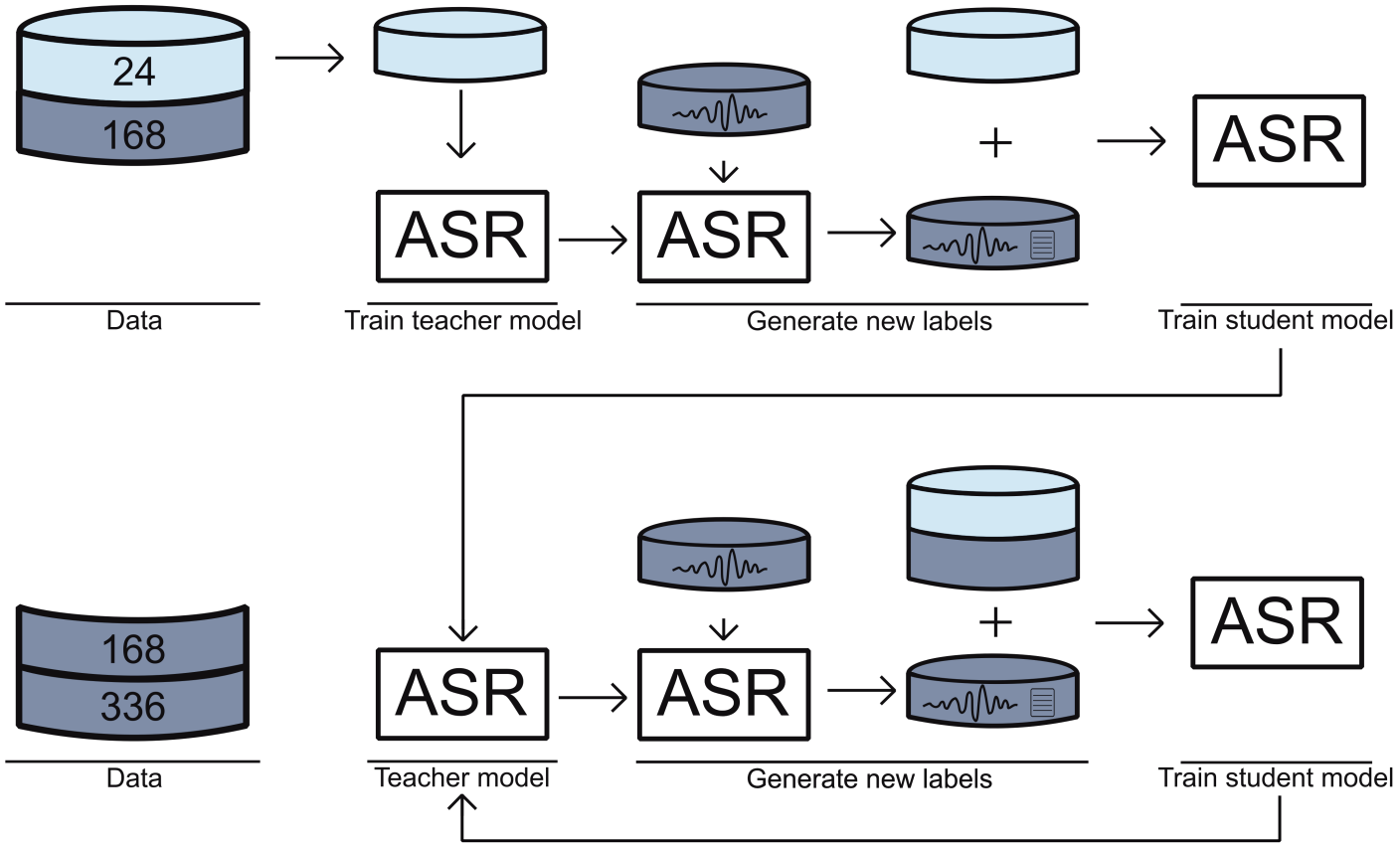



Image (modified): Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2021. XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale.

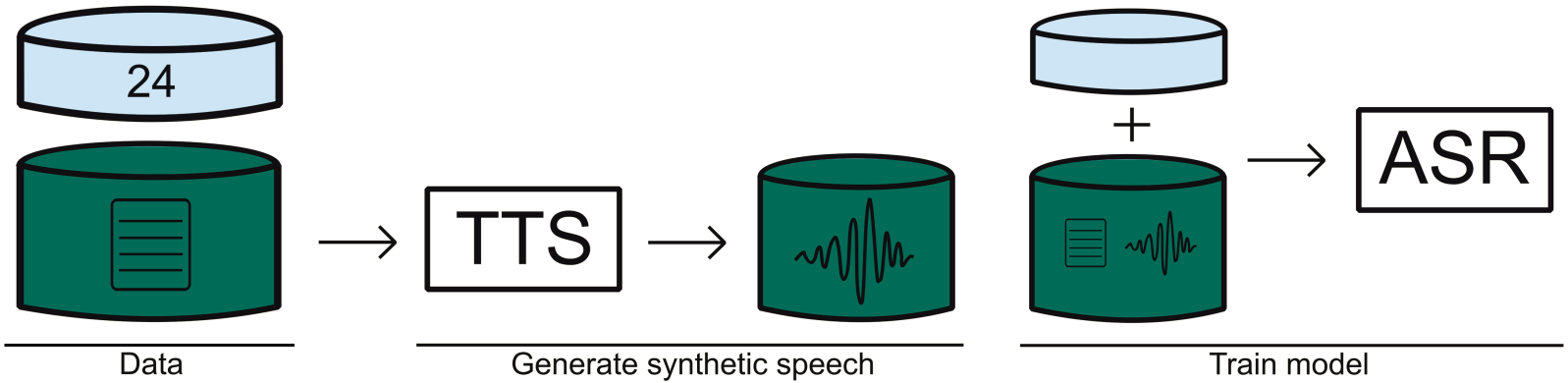
# Self-training

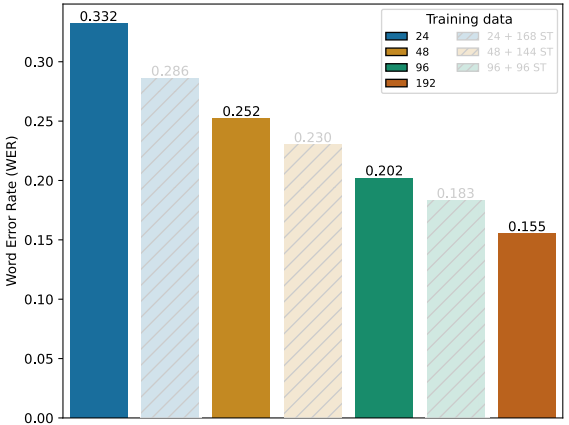


# Self-training on Gronings

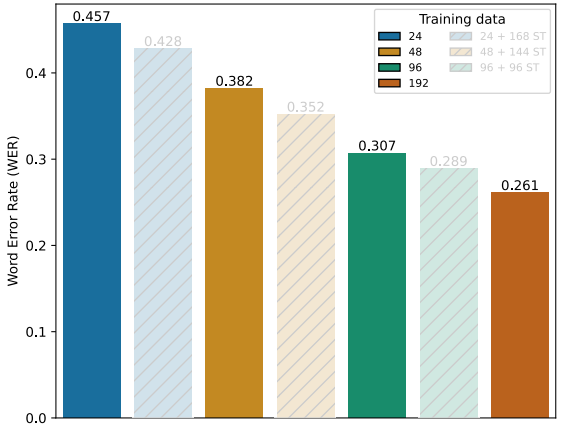


# Gronings synthetic speech

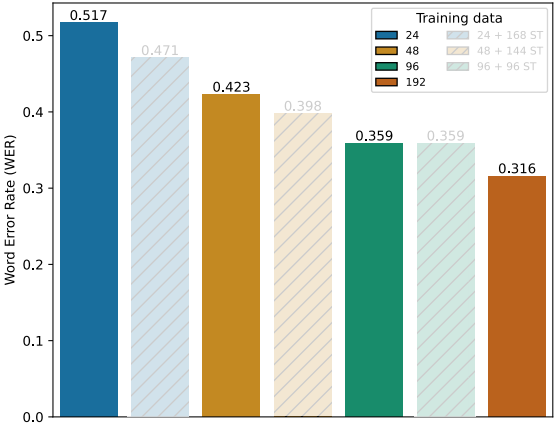




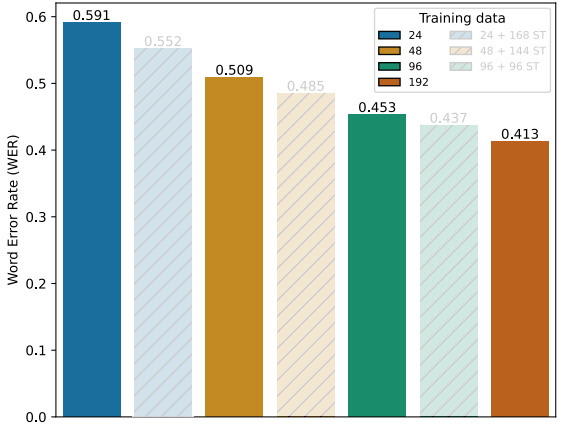

(a) Results for the Gronings test set.



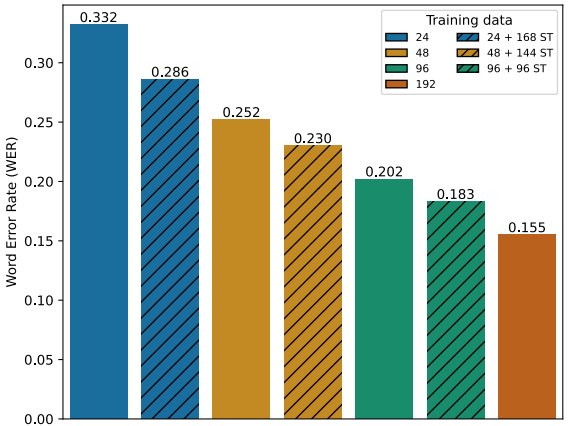

(b) Results for the West-Frisian test set.



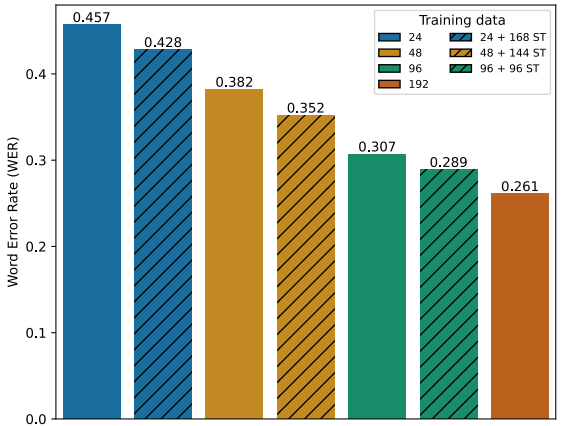

(c) Results for the Besemah test set.



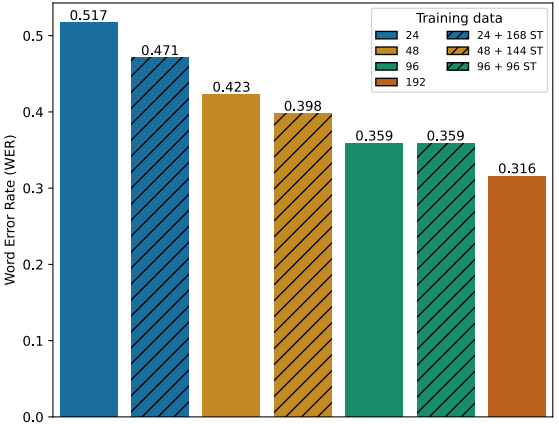

(d) Results for the Nasal test set.



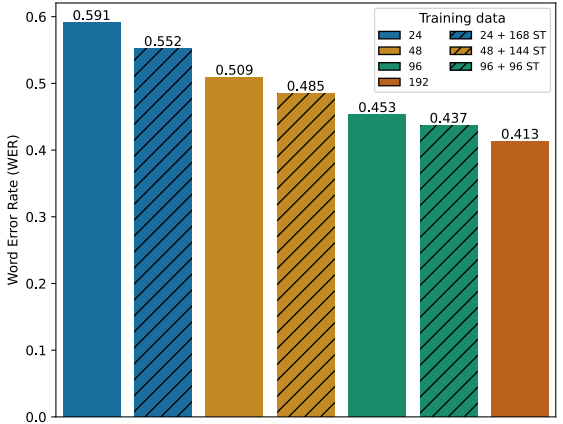

(a) Results for the Gronings test set.



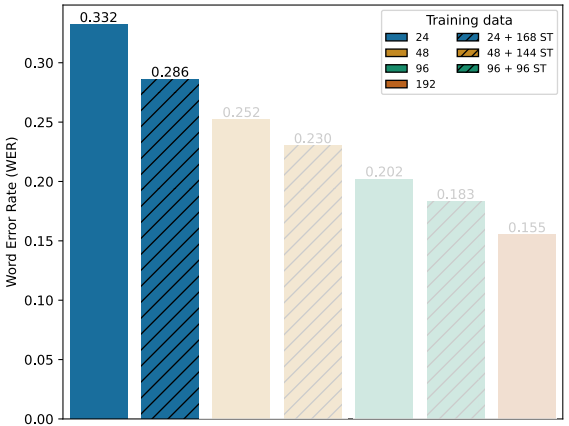

(b) Results for the West-Frisian test set.



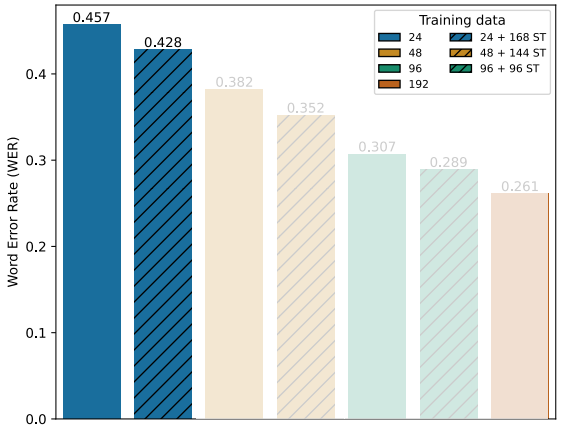

(c) Results for the Besemah test set.



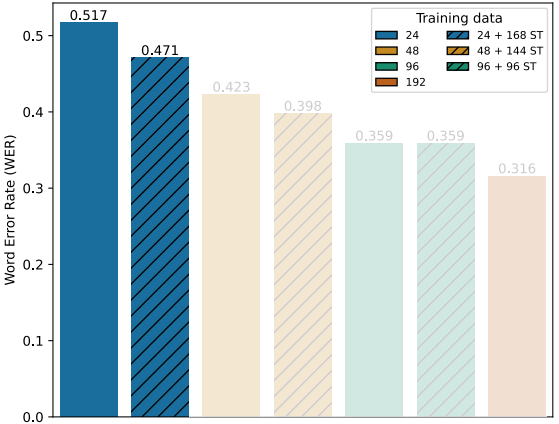

(d) Results for the Nasal test set.



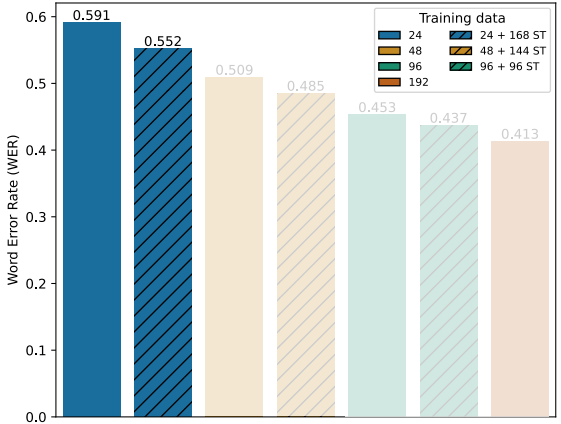

(a) Results for the Gronings test set.




(b) Results for the West-Frisian test set.




(c) Results for the Besemah test set.



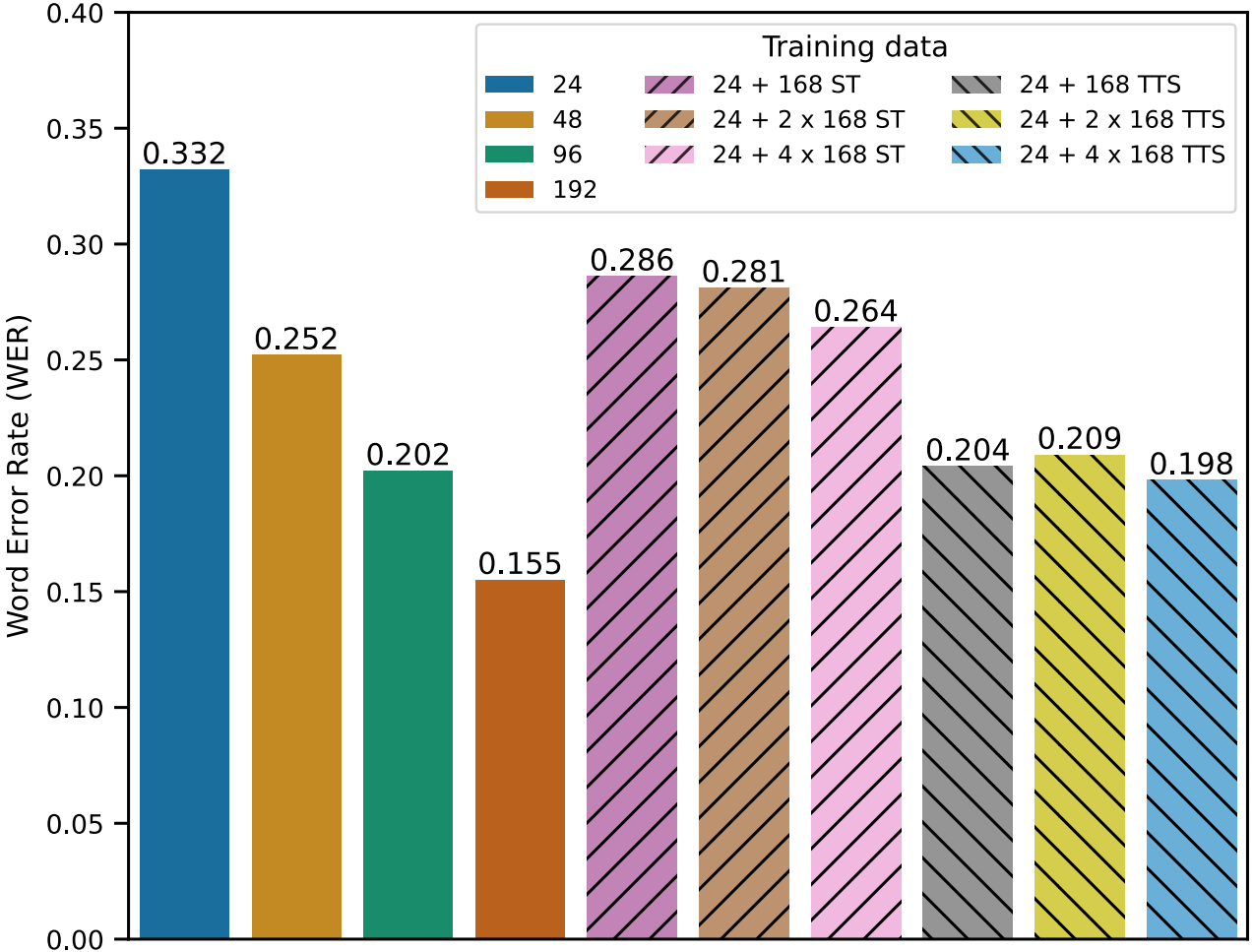

(d) Results for the Nasal test set.

Although the transcripts generated by the 24-minute model might contain some errors, the model likely benefits from the larger amount of training data

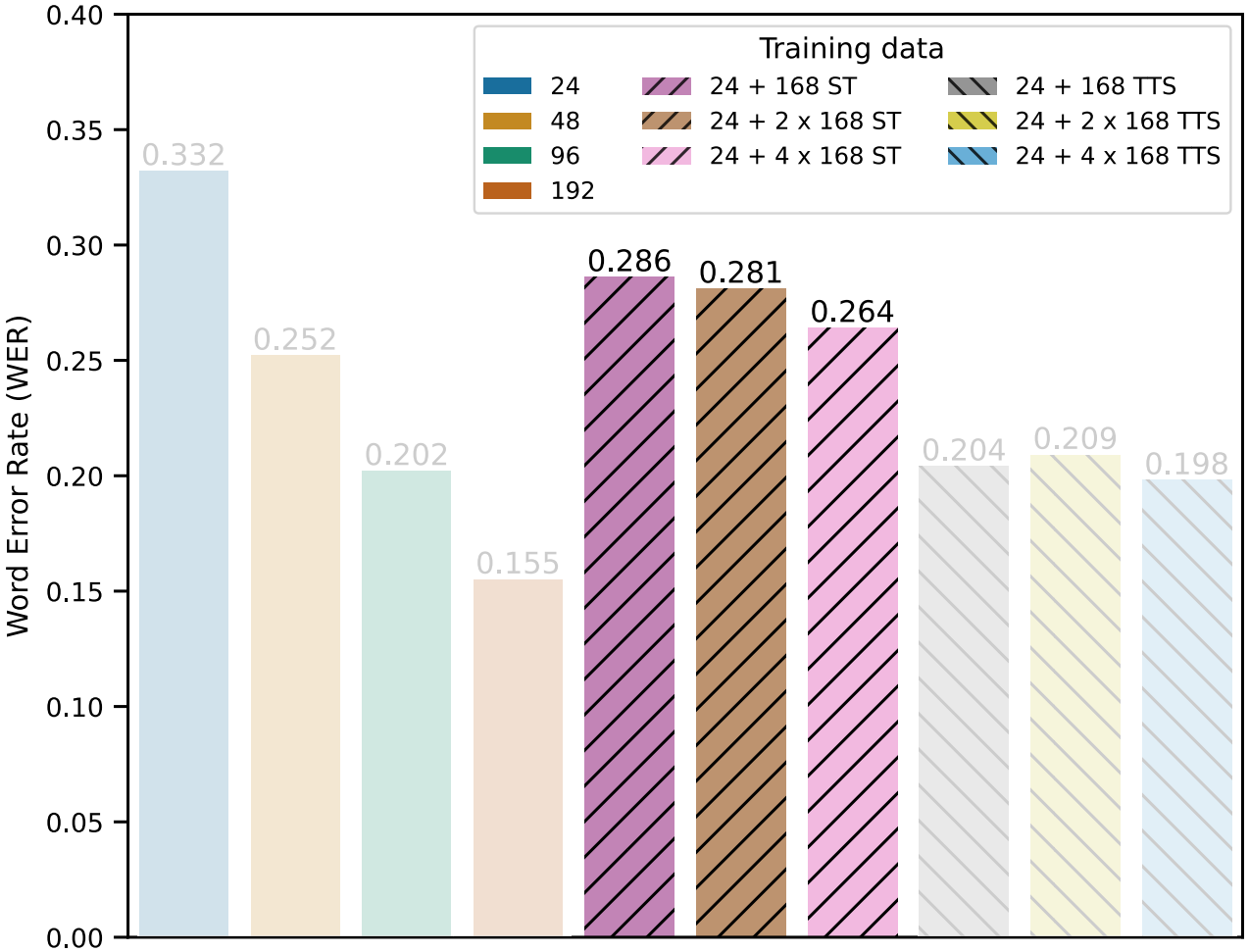
This effect diminishes when more manually transcribed training data is used

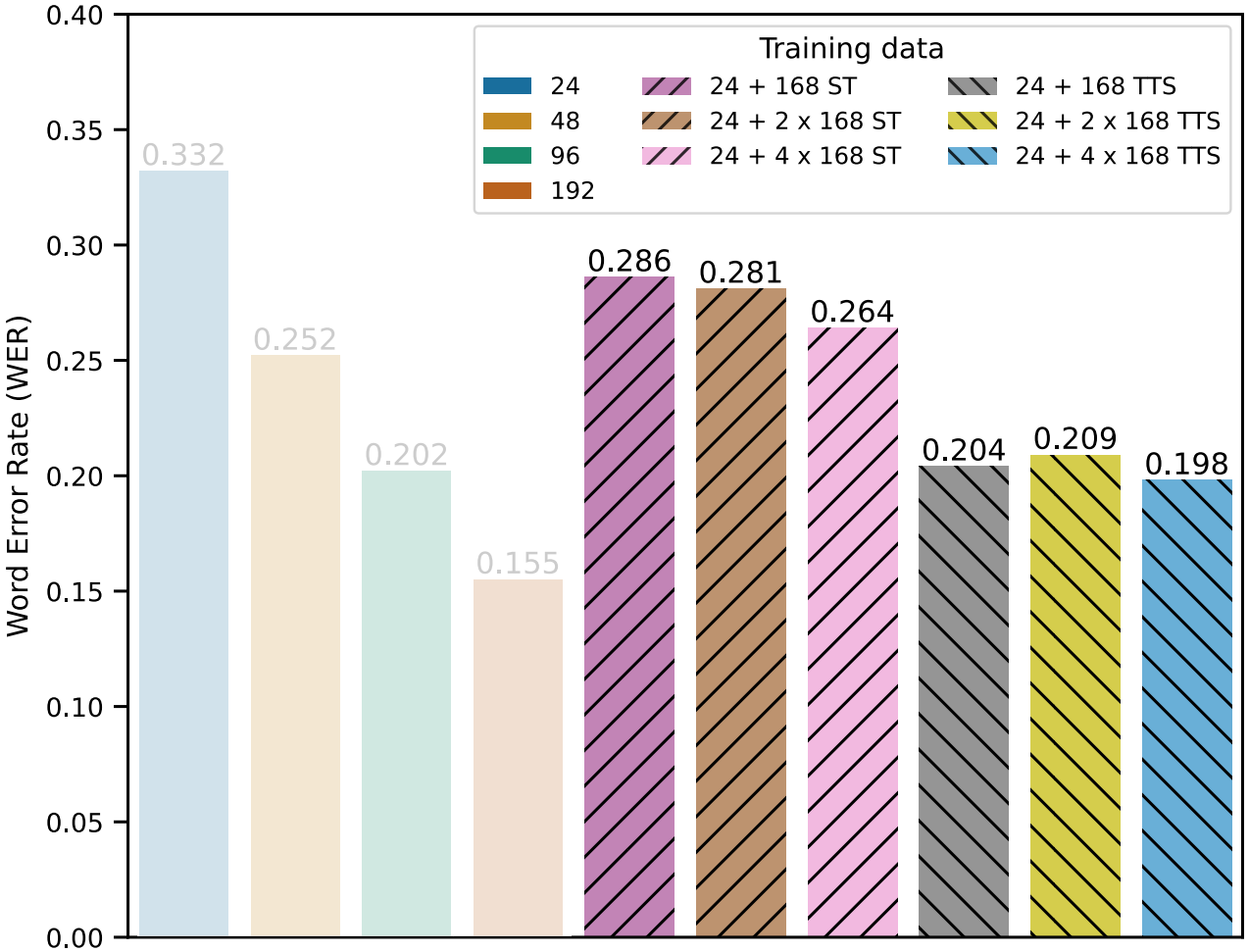
The baseline gets stronger and the added value of adding the labels obtained through self-training becomes less

Improving

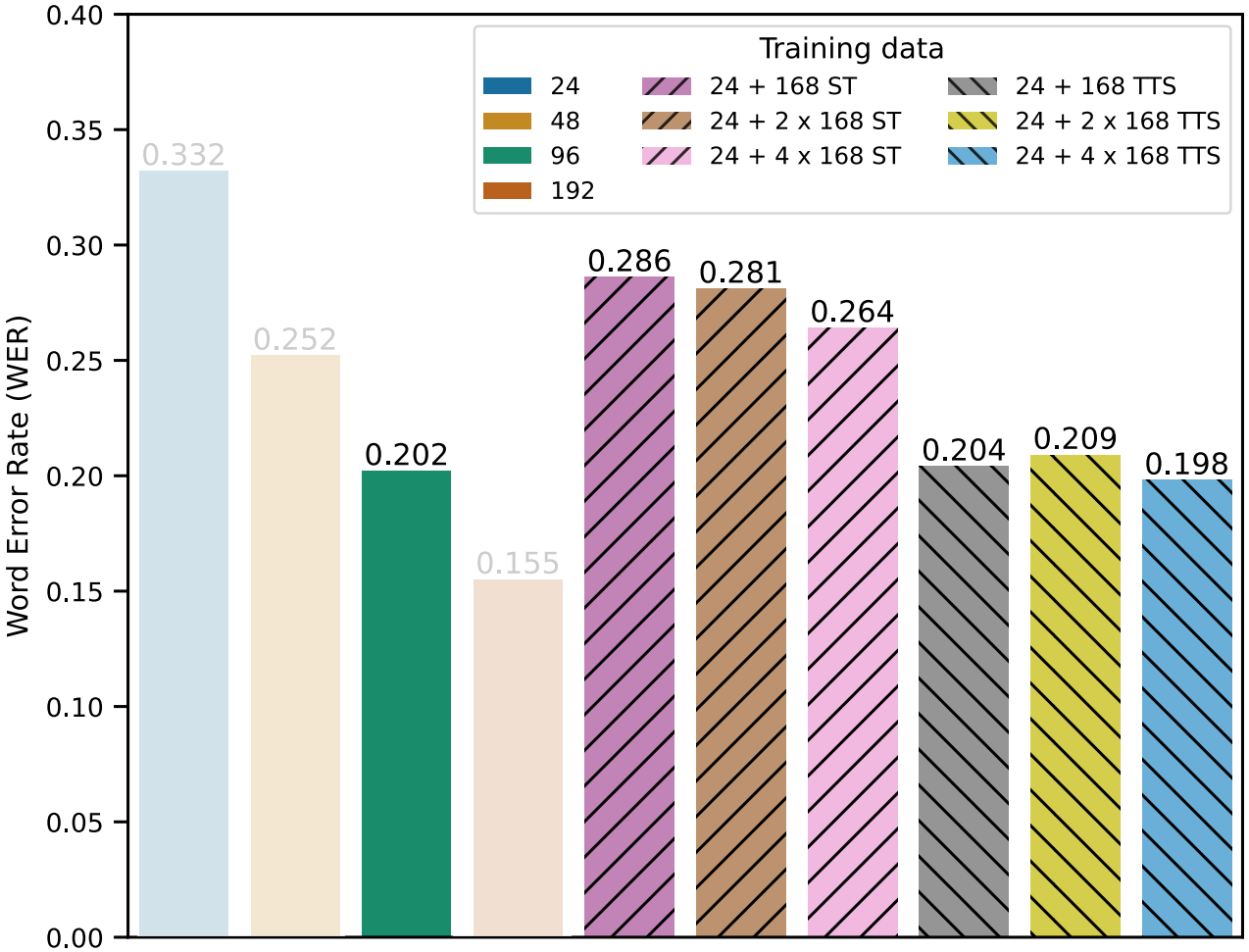

Training data
24
48
96
192
24 + 168 ST
24 + 2 x 168 ST
24 + 4 x 168 ST
24 + 168 TTS
24 + 2 x 168 TTS
24 + 4 x 168 TTS
Word Error Rate (WER)
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
0.332
0.252
0.202
0.155
0.286
0.281
0.264
0.204
0.209
0.198

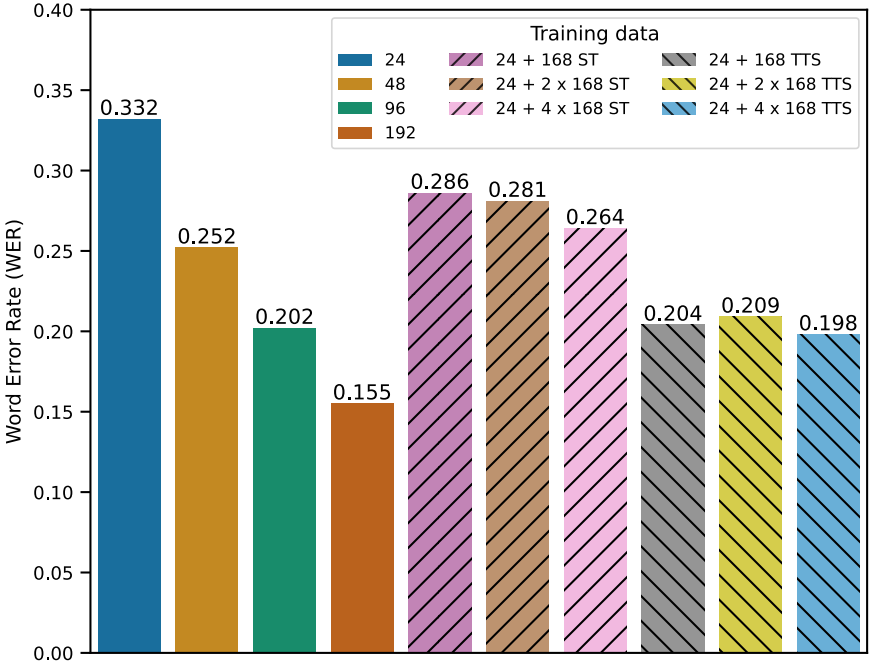
Improving


Training data
24
48
96
192
24 + 168 ST
24 + 2 x 168 ST
24 + 4 x 168 ST
24 + 168 TTS
24 + 2 x 168 TTS
24 + 4 x 168 TTS
Word Error Rate (WER)
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
0.332
0.252
0.202
0.155
0.286
0.281
0.264
0.204
0.209
0.198

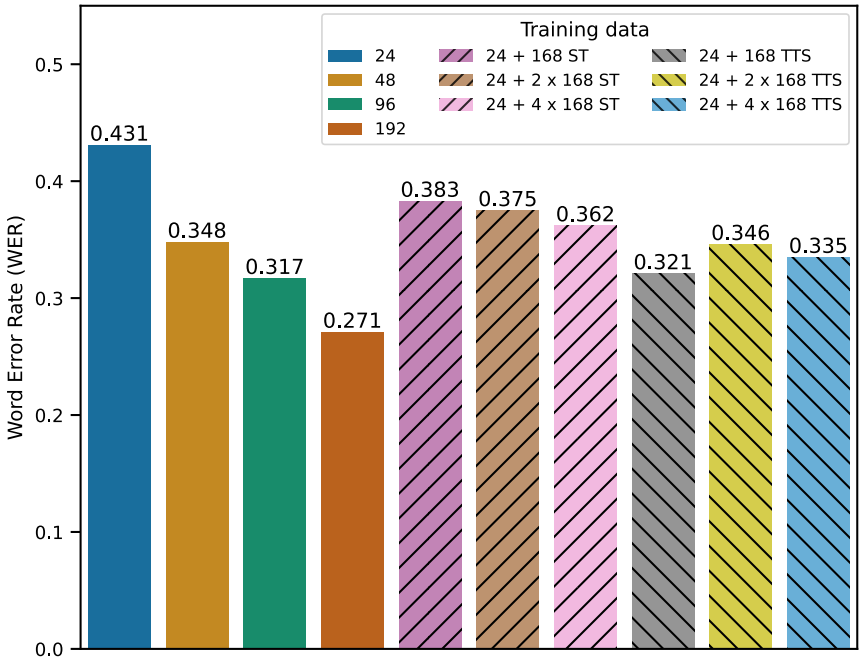

Training data
24
48
96
192
24 + 168 ST
24 + 2 x 168 ST
24 + 4 x 168 ST
24 + 168 TTS
24 + 2 x 168 TTS
24 + 4 x 168 TTS
Word Error Rate (WER)
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
0.332
0.252
0.202
0.155
0.286
0.281
0.264
0.204
0.209
0.198

Improving


Training data
24
48
96
192
24 + 168 ST
24 + 2 x 168 ST
24 + 4 x 168 ST
24 + 168 TTS
24 + 2 x 168 TTS
24 + 4 x 168 TTS
Word Error Rate (WER)
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
0.332
0.252
0.202
0.155
0.286
0.281
0.264
0.204
0.209
0.198




(a) Results for the regular Gronings test set.




(b) Results for the out-of-domain Gronings test set.

# Summary

We found that:

- Data augmentation techniques may serve as a **cost-effective** way to improve low-resource ASR performance in a real-world setting

# Summary

We found that:

- Data augmentation techniques may serve as a **cost-effective** way to improve low-resource ASR performance in a real-world setting
- The largest performance gains were observed when increasing the amounts of manually transcribed data

# Summary

We found that:

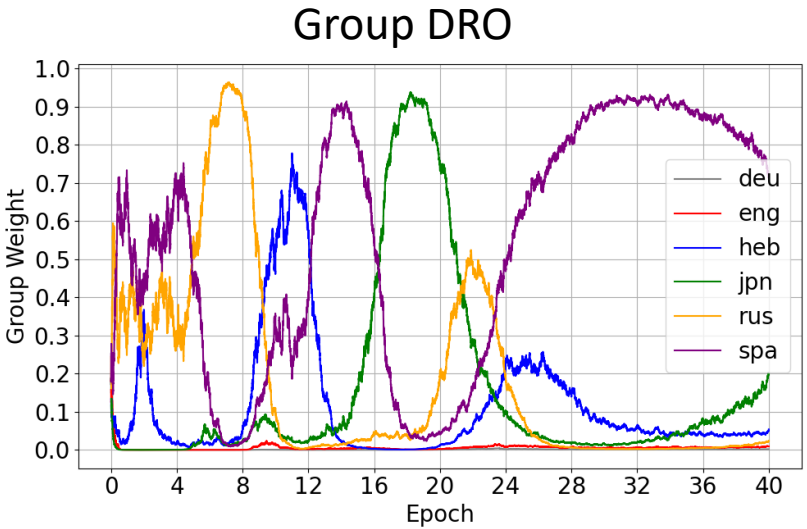
- Data augmentation techniques may serve as a **cost-effective** way to improve low-resource ASR performance in a real-world setting
- The largest performance gains were observed when increasing the amounts of manually transcribed data
- To help researchers to include real-world data into their analysis and applications, we released our datasets in addition to our code and models

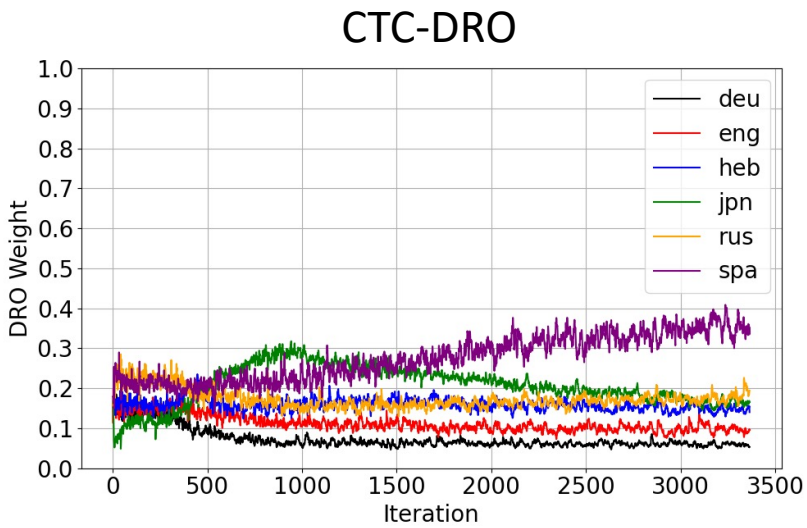
# Broader context

**CTC-DRO: Robust Optimization for Reducing Language Disparities in Speech Recognition**

**Martijn Bartelds** [*1] **Ananjan Nandi** [*1] **Moussa Koulako Bala Doumbouya** [1] **Dan Jurafsky** [1]
**Tatsunori Hashimoto** [1] **Karen Livescu** [2]

# Broader context



| Set # | Model | Type | Max CER (ISO) (↓) | Avg CER (↓) | LID (↑) |
|---|---|---|---|---|---|
| 1 | MMS | Baseline | 60.8 (nan) | 23.4 | **97.4** |
| | | group DRO | 86.6 (nan) | 30.5 | 78.7 |
| | | CTC-DRO | **56.8** (nan) | **22.9** | 95.8 |
| | XLS-R | Baseline | 64.9 (cmn) | 25.2 | **92.6** |
| | | group DRO | 78.4 (nan) | 30.0 | 87.8 |
| | | CTC-DRO | **57.6** (nan) | **22.5** | 89.5 |
| 2 | MMS | Baseline | 49.4 (yue) | 15.8 | **98.4** |
| | | group DRO | 55.5 (yue) | 20.7 | 98.2 |
| | | CTC-DRO | **44.4** (yue) | **15.0** | 96.2 |
| | XLS-R | Baseline | 68.8 (yue) | 19.0 | **94.2** |
| | | group DRO | 58.8 (yue) | 21.6 | 87.0 |
| | | CTC-DRO | **45.0** (yue) | **15.8** | 89.3 |
| 3 | MMS | Baseline | 34.2 (kor) | 16.1 | 98.5 |
| | | group DRO | 34.0 (kor) | 22.0 | **98.7** |
| | | CTC-DRO | **31.3** (khm) | **15.3** | **98.7** |
| | XLS-R | Baseline | 33.2 (khm) | **17.0** | **99.2** |
| | | group DRO | 38.0 (khm) | 25.1 | 97.2 |
| | | CTC-DRO | **32.2** (khm) | 17.7 | 97.9 |
| 4 | MMS | Baseline | 24.0 (snd) | 14.4 | **87.9** |
| | | group DRO | 21.8 (urd) | 14.9 | 91.9 |
| | | CTC-DRO | **18.4** (urd) | **12.9** | 87.3 |
| | XLS-R | Baseline | 29.7 (urd) | 14.6 | 88.4 |
| | | group DRO | 25.6 (slv) | 18.6 | 83.5 |
| | | CTC-DRO | **24.2** (urd) | **13.7** | **88.9** |
| 5 | MMS | Baseline | 90.0 (jpn) | 26.0 | **96.3** |
| | | group DRO | 62.2 (jpn) | 29.2 | 67.0 |
| | | CTC-DRO | **57.5** (jpn) | **24.3** | 90.5 |
| | XLS-R | Baseline | 114.8 (jpn) | 29.9 | 89.0 |
| | | group DRO | 92.9 (jpn) | 36.8 | 57.7 |
| | | CTC-DRO | **71.5** (jpn) | **23.8** | **91.0** |

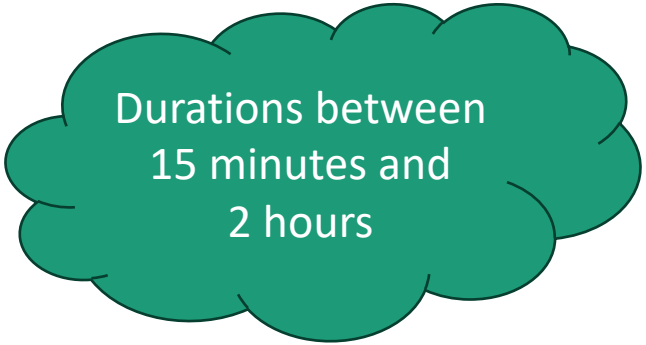
# Broader context

**BLAB: Brutally Long Audio Bench**

**Orevaoghene Ahia[1] Martijn Bartelds[2] Kabir Ahuja[1] Hila Gonen[1]
Valentin Hofmann[1] Siddhant Arora [4] Shuyue Stella Li[1] Vishal Puttagunta[3]
Mofetoluwa Adeyemi[5] Charishma Buchireddy[3] Ben Walls[3] Noah Bennett[3]
Shinji Watanabe[4] Noah A. Smith[1] Yulia Tsvetkov[1] Sachin Kumar[3]**

[1]University of Washington [2]Stanford University [3]The Ohio State University
[4]Carnegie Mellon University [5]University of Waterloo

# Broader context




**BLAB: Brutally Long Audio Bench**

**Orevaoghene Ahia[1] Martijn Bartelds[2] Kabir Ahuja[1] Hila Gonen[1]**
**Valentin Hofmann[1] Siddhant Arora [4] Shuyue Stella Li[1] Vishal Puttagunta[3]**
**Mofetoluwa Adeyemi[5] Charishma Buchireddy[3] Ben Walls[3] Noah Bennett[3]**
**Shinji Watanabe[4] Noah A. Smith[1] Yulia Tsvetkov[1] Sachin Kumar[3]**

[1]University of Washington [2]Stanford University [3]The Ohio State University
[4]Carnegie Mellon University [5]University of Waterloo

# Broader context

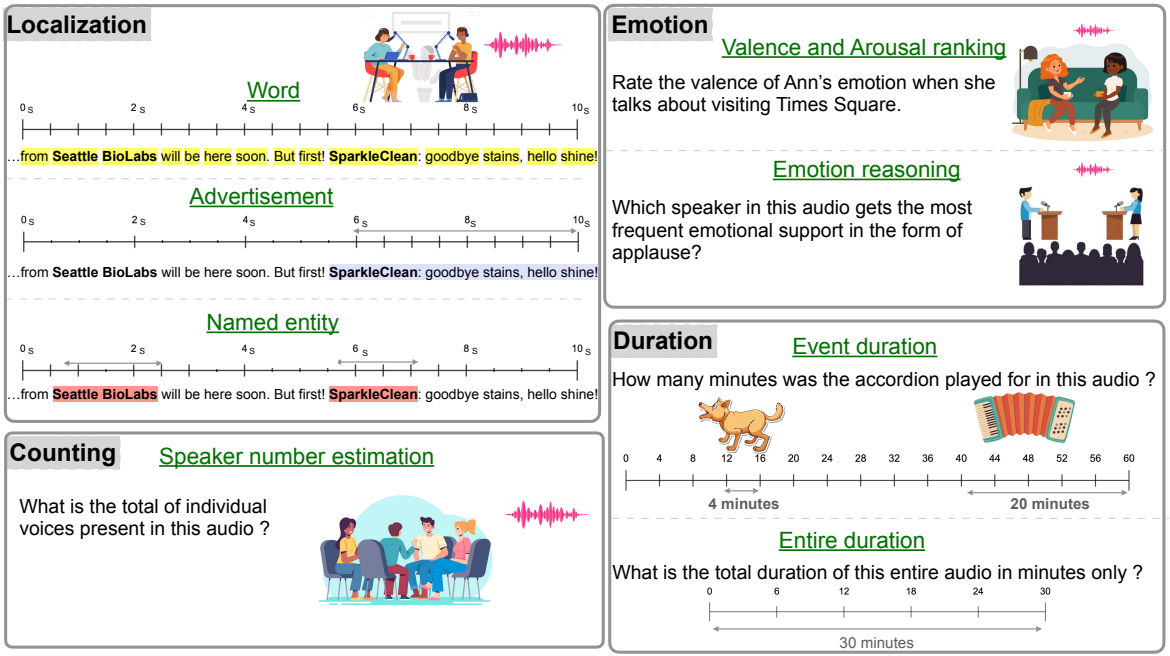



Figure 1: Overview of BLAB, designed to test true multimodal understanding abilities of audio LMs. It contains eight distinct audio tasks across four categories, namely **localization**, **counting**, **emotion**, and **duration estimation**. † All images are designed by Freepik .

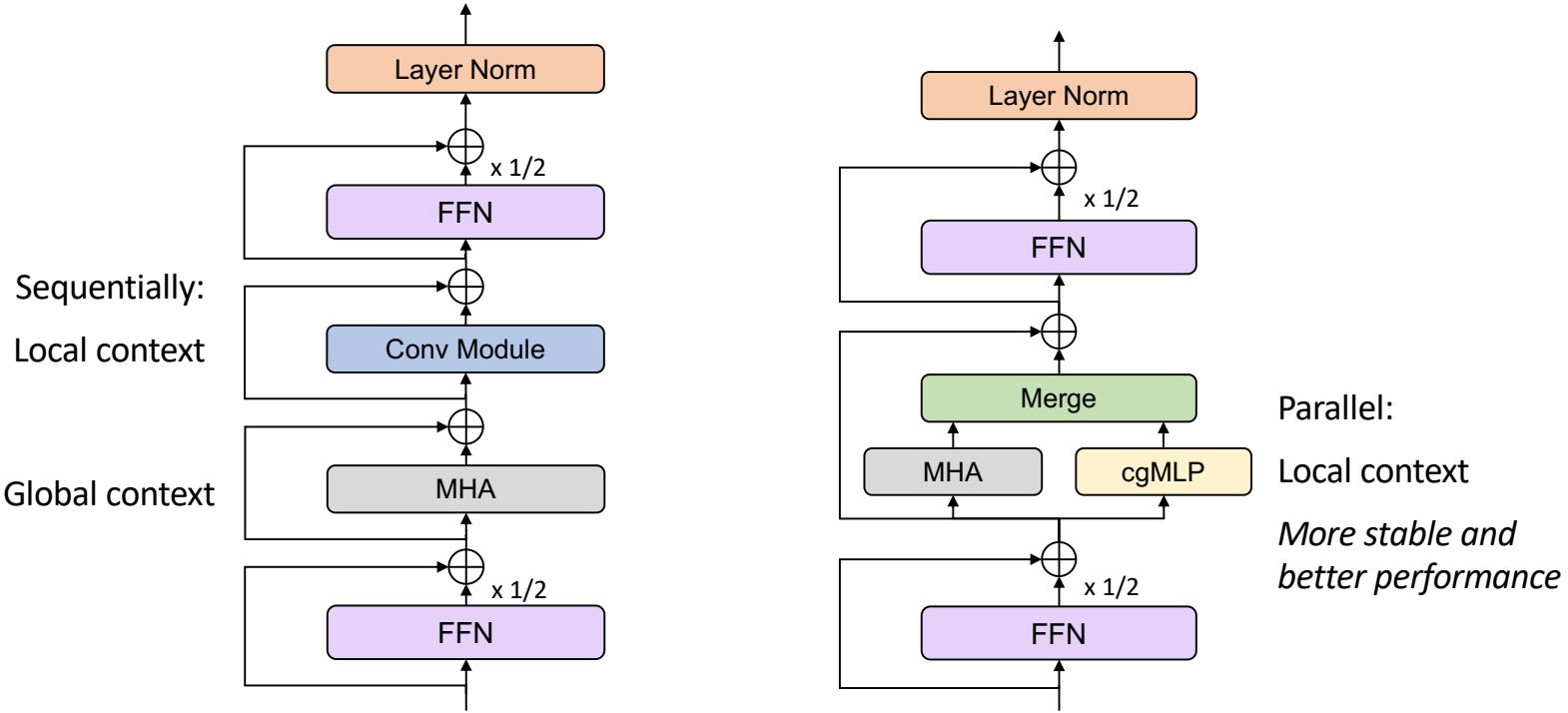
# Broader context

| Task | Metric (↑) | Gemini 2.0 Flash | Gemini 2.0 Pro |
|---|---|---|---|
| Word Localization | word $F_1$ | 1.12 | 0.19 |
| Advertisement localization | Frame-level $F_1$ | 4.93 | 0.15 |
| NE Localization | Frame-level $F_1$ | 2.97 | 2.14 |
| Speaker Number Estimation | EMA | 8.00 | 8.50 |
| Valence and Arousal Ranking | EMA | 26.28 | 32.00 |
| Emotion Reasoning | EMA | 54.54 | 64.29 |
| Entire Duration | EMA (without / with $\pm$2 seconds offset) | 0.50/3.50 | 0.00/2.50 |
| Event Duration | EMA (with / without $\pm$2 seconds offset) | 1.49/4.95 | 1.49/3.96 |

Table 2: Performance comparison of Gemini audio LMs across all BLAB long audio tasks. Both models exhibit similar performance, generally achieving low performance across tasks.

# Contact information


bartelds@stanford.edu | martijnbartelds.nl | @barteldsmartijn




(a) *Conformer [10]*

(b) *E-Branchformer [13]*

Peng et al., 2024