2025 ICASSP HYDERABAD, INDIA
Celebrating Signal Processing

Towards a Single ASR Model That Generalizes to Disordered Speech

https://arxiv.org/abs/2412.19315

Jimmy Tobin, Katrin Tomanek, **Subhashini Venugopalan**

Project Euphonia

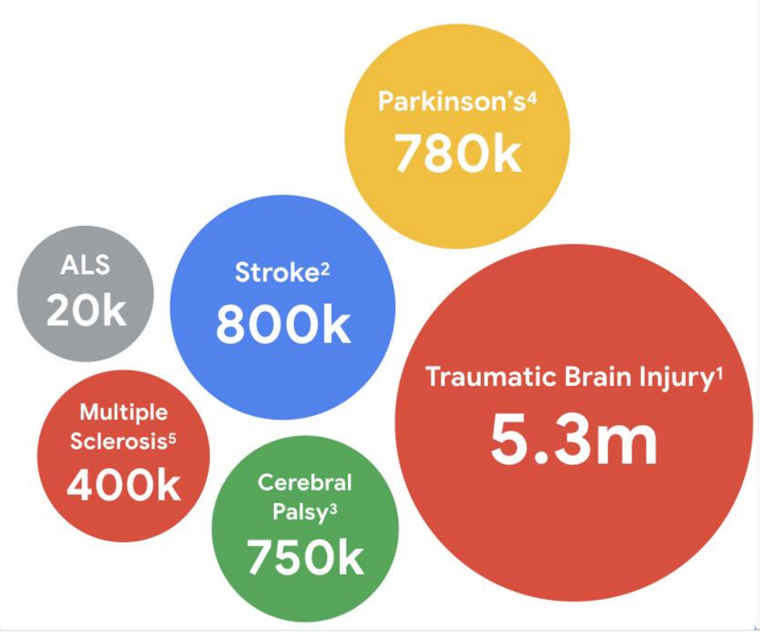
Improve ASR to help people with **speech disorders** who have difficulty being understood by other people and technology.

Our goal is to help these users **communicate** and **gain independence**.

https://sites.research.google/euphonia/about/

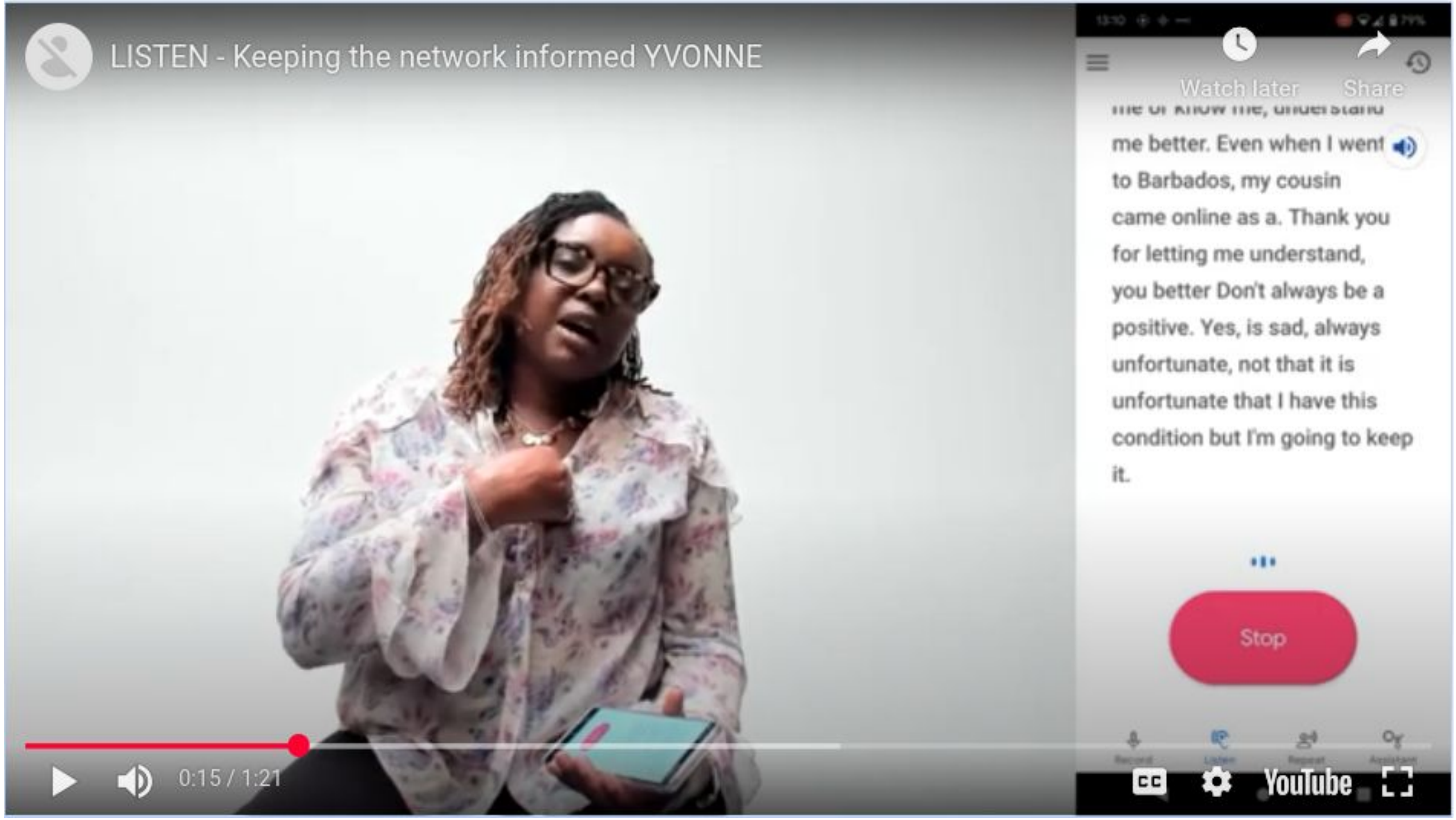
Condition prevalence (US)

Millions of users have neurological conditions that cause speech impairments, in the US and around the world.



Google

Project Relate - Personalize their on-device ASR model



Google

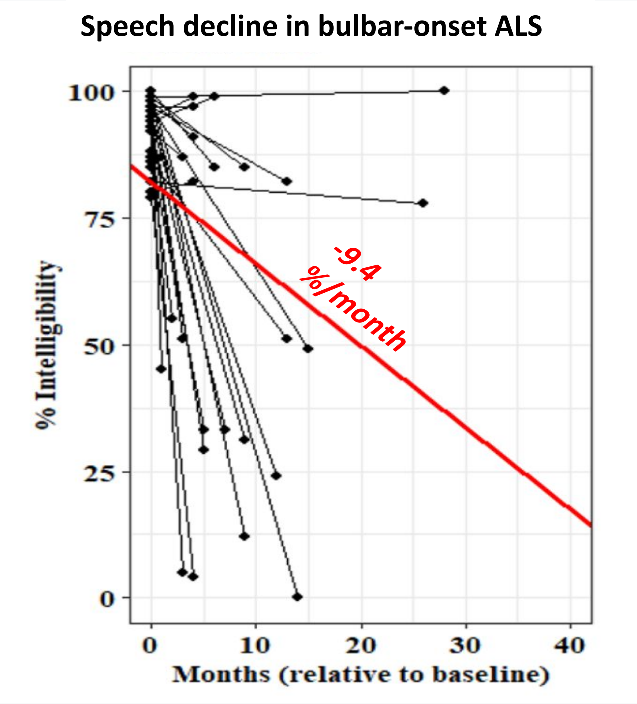
Can we get the default production ASR model to work well for speakers with impairments?

(can we reduce the need for personalization)

Limitations of personalization

Through feedback from users of Project Relate gathered by our speech-language pathologist team, we have identified some challenges to personalization:

- **Enrollment**: For some users, recording speech prompts can be physically demanding because of muscular weakness and fatigue. Cognitive impairment may also lead to incorrectly recorded prompts.
- **Degenerating speech intelligibility**: Diseases like ALS cause people's speech to decline in a unpredictable way. Continuous recording is needed to mitigate this [Tomanek et. al. ICASSP'2023]



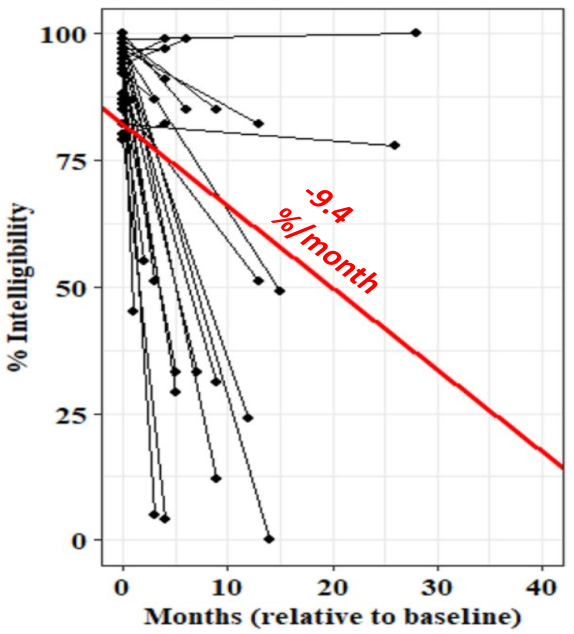
Eshghi et al., 2022

Limitations of personalization

Through feedback from users of Project Relate gathered by our speech-language pathologist team, we have identified some challenges to personalization:

- **Enrollment**: For some users, recording speech prompts can be physically demanding because of muscular weakness and fatigue. Cognitive impairment may also lead to incorrectly recorded prompts.
- **Degenerating speech intelligibility**: Diseases like ALS cause people's speech to decline in a unpredictable way. Continuous recording is needed to mitigate this [Tomanek et. al. ICASSP'2023]
- **Conversation**: Personalization models trained on short, transactional phrases, and not trained for conversations. Conversations have more varied vocabulary, are longer in length, and contain more named entities / rare words.

Speech decline in bulbar-onset ALS



Eshghi et al., 2022

Objectives

- **Speaker Independent ASR (SI-ASR)**: Need a speaker-independent (unpersonalized) ASR model which works well on disordered speech.
- **Generalize to conversational speech**: Should generalize well to conversational speech.
- **No regression on standard speech evals**: The ideal best-case scenario is to ensure there is no regression on standard ASR benchmarks so the same model can be used for all users to provide a good experience.

How?

Joint training

Combine disordered speech datasets with standard speech data and learn to weight

Real Conversation

Evaluate generalization to conversational speech

Large and on-device

Train on large and on-device models to measure generalizability of approach

How?

- **Euphonia combined disordered speech dataset**: Consider the entire corpus of disordered speech data available for training, along with the standard speech datasets.
  - Learn how to weight the disordered speech data (it is out of distribution)
- **Conversational speech test set**: Gather a dataset for conversational speech to evaluate generalization.
- **Large and on-device model**: Train both large and on-device models with the same process to measure generalizability of the approach.

Speaker Independent ASR (SI-ASR) dataset

For evaluation and training of speaker-independent (SI) ASR models for impaired speech we split the full Euphonia prompted speech corpus such that

- there is no overlap on *the speaker and phrase level*

between the training and the test set.

- The testset consists of 5700 utterances from 200 speakers with different severities and types of speech impairment.
- The training set consists of ~950k utterances from ~1200 speakers.

Set	# speakers	# utterances	# hours
Test	200	5699	9
Train	1246	956645	~1158
Dev	24	358	0.64

Real Conversation Test Set

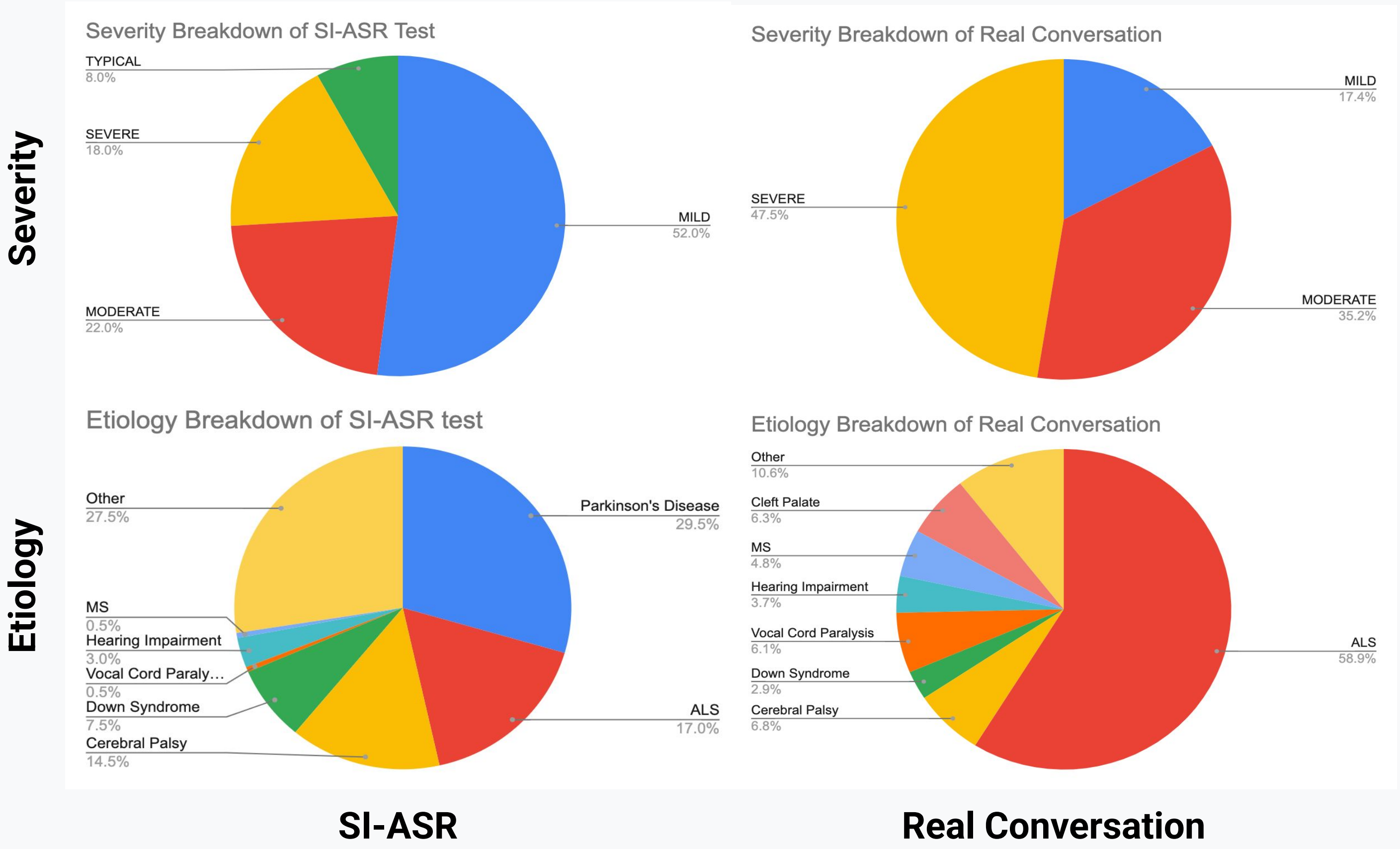
The Real Conversation test set was compiled from Trusted Tester's real world usage of Euphonia's data collection app. Speech-language pathologists

- scrubbed data of any PII,
- transcribed the speech, and
- verified as "organic usage in a conversation".

Note: All speakers in this test set were removed from the Speaker Independent ASR training set.

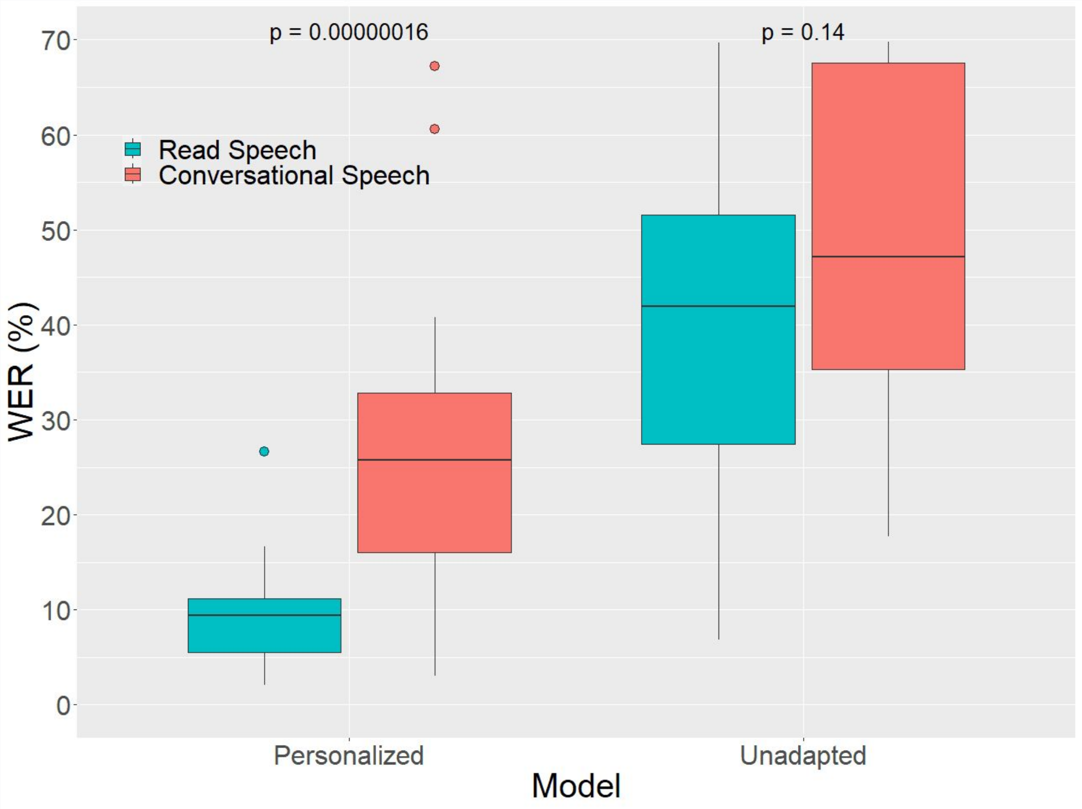
# Speakers	29
# Utterances	1515

SI-ASR test set vs Real Conversation test set



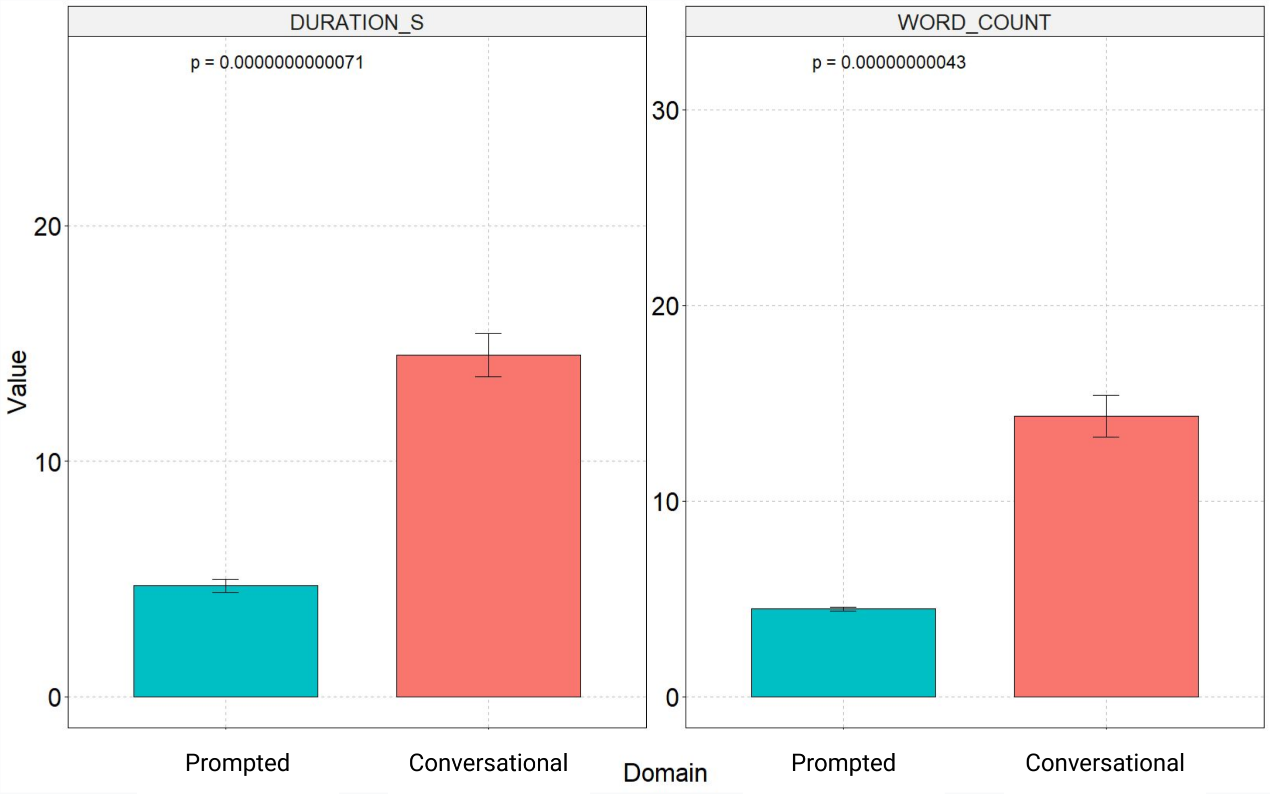
Google

Real Conversations is a harder use case



Comparisons for the same 29 speakers in the real conversation test set and speaker-independent prompted speech test set

Duration and #words are much longer in conversation

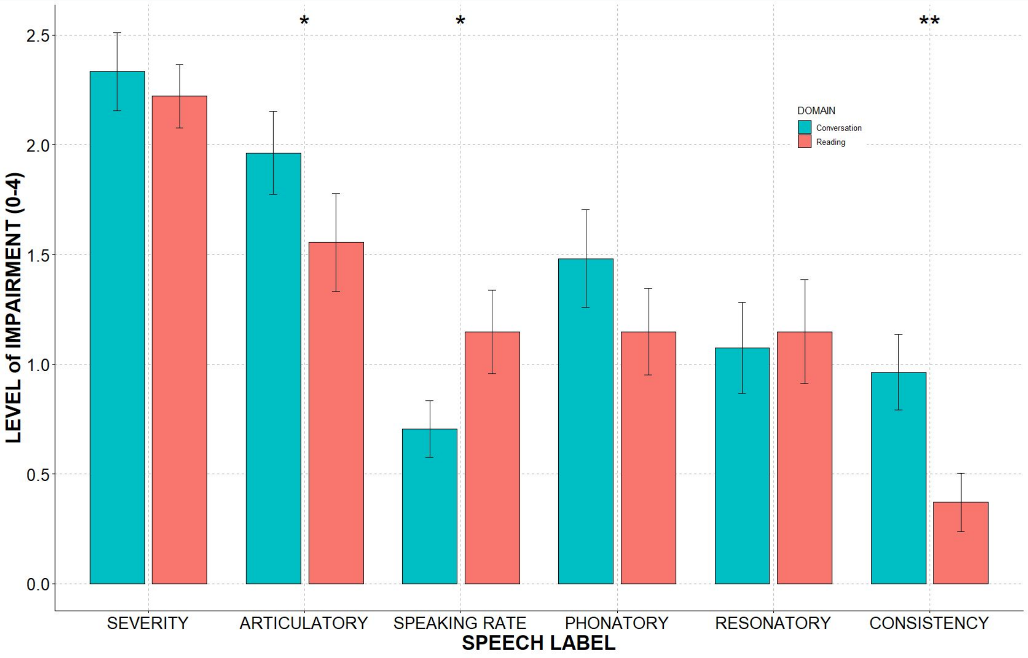


Real Conversation vs Prompts (cont)

Speech-Language Pathologists assessed the level of impairment (0-4) across different speech characteristics for the group of 29 speakers.

During conversational speech, people with disordered speech demonstrated…

- greater impairments to articulation & voice (phonatory)
- more inconsistency in their speech patterns
- a more typical rate of speech

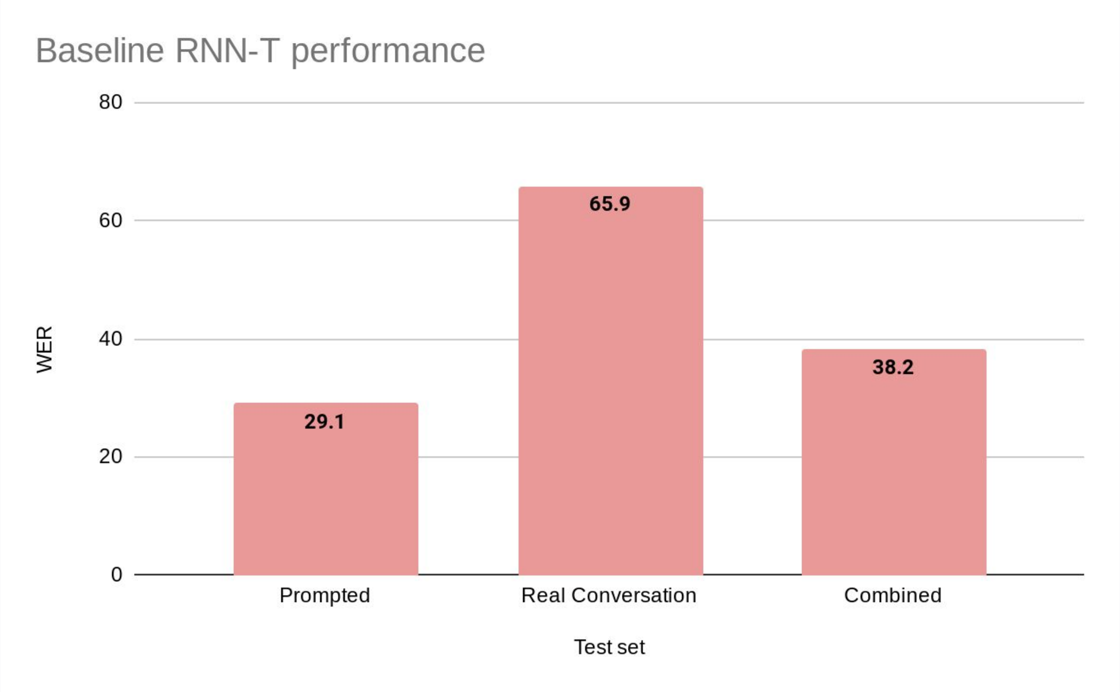


Google

Performance at baseline

Baseline on-device (RNN-T) model

On-device model is an **RNN-T** model that was trained on a wide range of standard speech including conversational data.

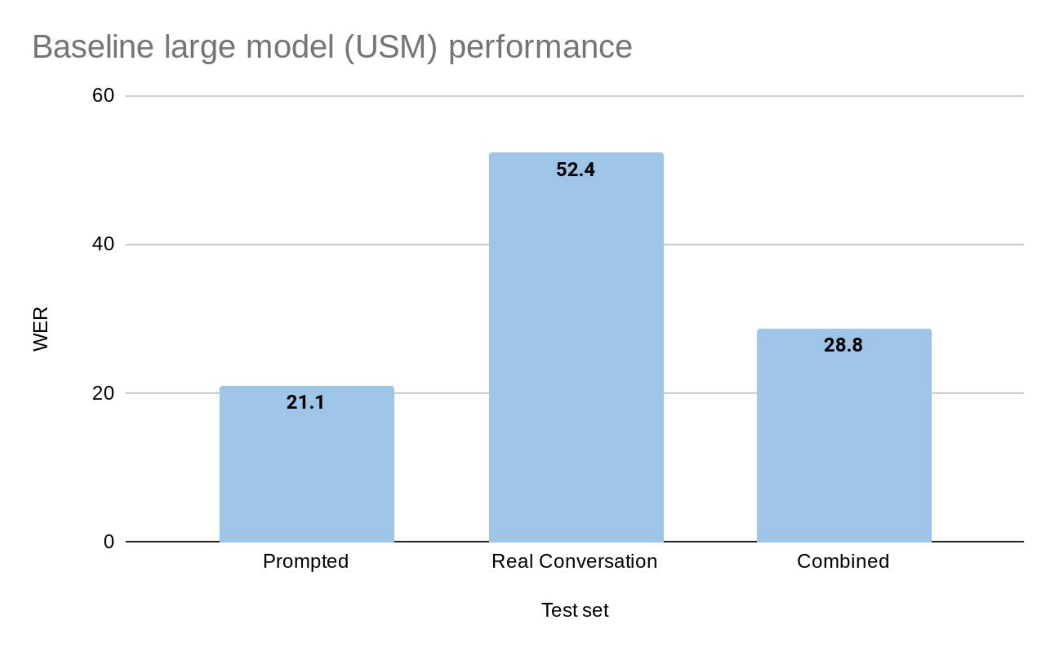


RNN-T Models Fail to Generalize to Out-of-Domain Audio: Causes and Solutions
Chung-Cheng Chiu, Arun Narayanan, Wei Han, Rohit Prabhavalkar, Yu Zhang, Navdeep Jaitly, Ruoming Pang, Tara N. Sainath, Patrick Nguyen, Liangliang Cao, Yonghui Wu
SLT 2020

Google

Baseline large model (USM-CTC)

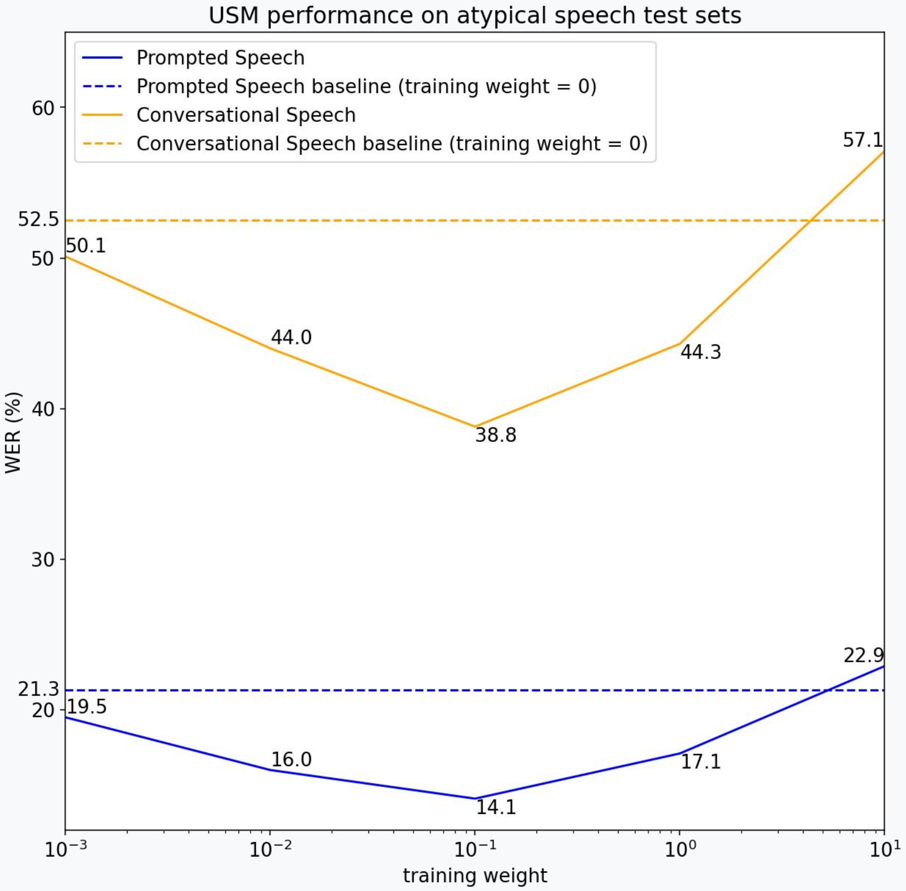
Large model is a Universal Speech Model (USM 2B) It is a CTC conformer model that was trained for multilingual ASR use case on unlabeled and labeled speech in 100+ languages.



Google USM: Scaling Automatic Speech Recognition Beyond 100 Languages
Yu Zhang, Wei Han, James Qin, Yongqiang Wang, Ankur Bapna, Zhehuai Chen, Nanxin Chen, Bo Li, Vera Axelrod, Gary Wang, Zhong Meng, Ke Hu, Andrew Rosenberg, Rohit Prabhavalkar, Daniel S. Park, Parisa Haghani, Jason Riesa, Ginger Perng, Hagen Soltau, Trevor Strohman, Bhuvana Ramabhadran, Tara Sainath, Pedro Moreno, Chung-Cheng Chiu, Johan Schalkwyk, Françoise Beaufays, Yonghui Wu

Finetune and adapt on disordered speech with different weights

Perf. on val under different weights



Google

Monitor perf. on disordered speech dataset

Training data weight	**Mild**	**Moderate**	**Severe**
0 (Baseline)	10.7	29.5	48.1
0.001	9.7	26.5	44.5
0.01	8.2	21.9	35.2
0.1	**7.3**	**19.6**	**31.3**
1.0	10.3	23.1	33.8
10.0	15.1	29.8	42.6

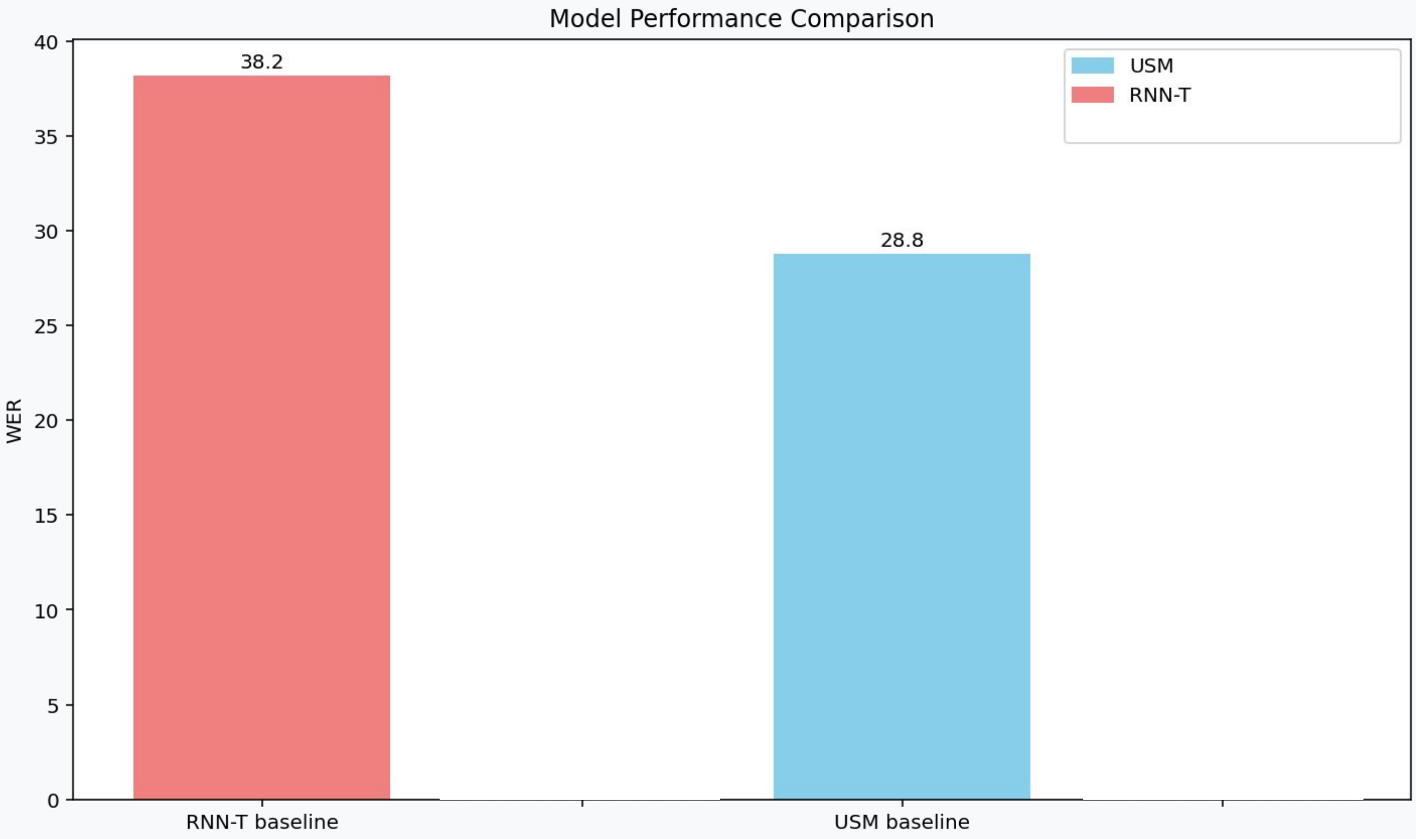
Verify perf. on standard speech tests

Ensure no regression

Training data weight	Multilingual test		Librispeech	
	en-US	18 langs.	Clean	Other
0 (Baseline)	**13.5**	**19.4**	**2.4**	4.6
0.001	**13.5**	**19.4**	**2.4**	**4.5**
0.01	**13.5**	**19.4**	**2.4**	**4.5**
0.1	**13.5**	**19.4**	**2.4**	4.6
1.0	**13.5**	**19.4**	**2.4**	4.6
10.0	13.6	19.5	2.6	4.8

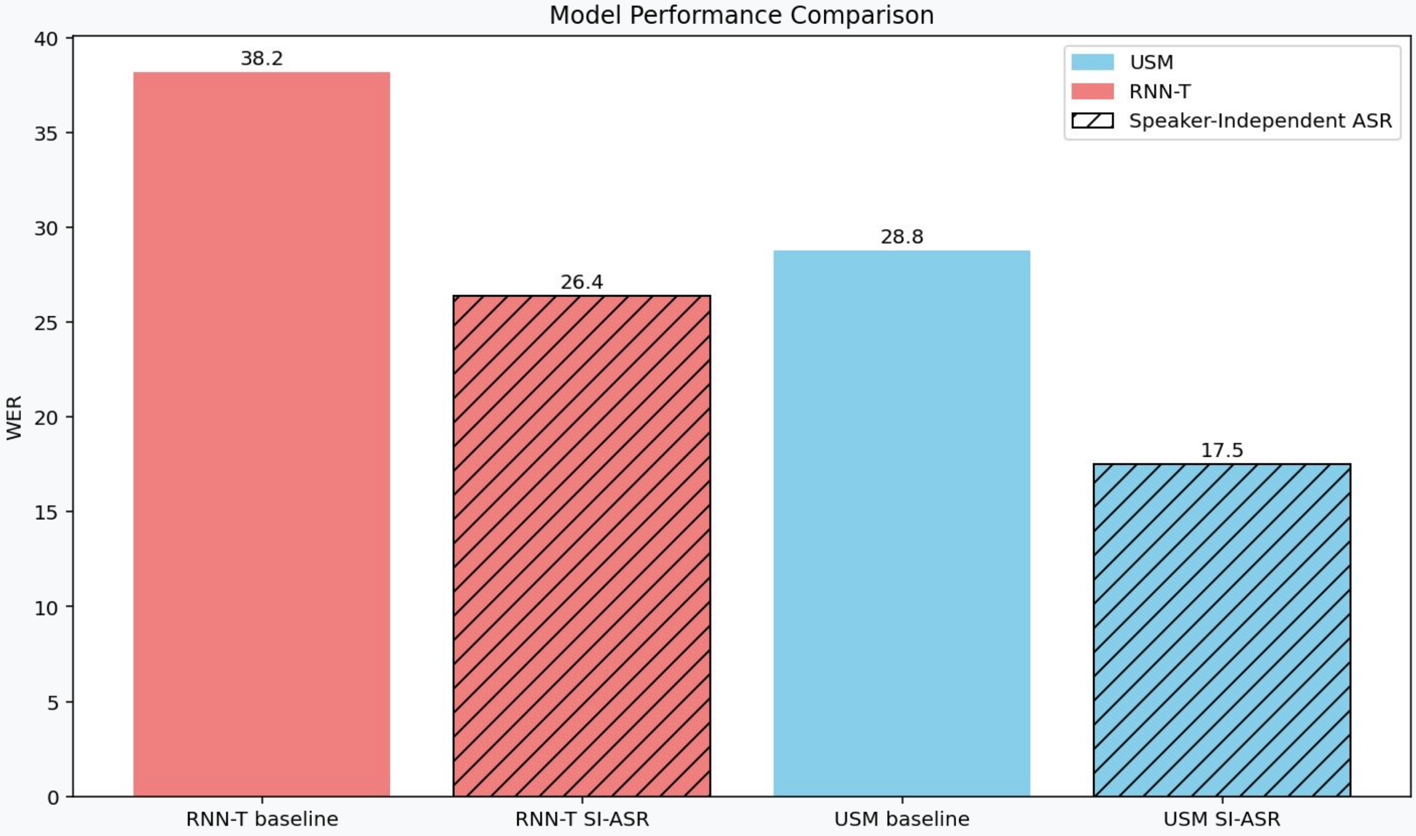
Results

Performance at baseline



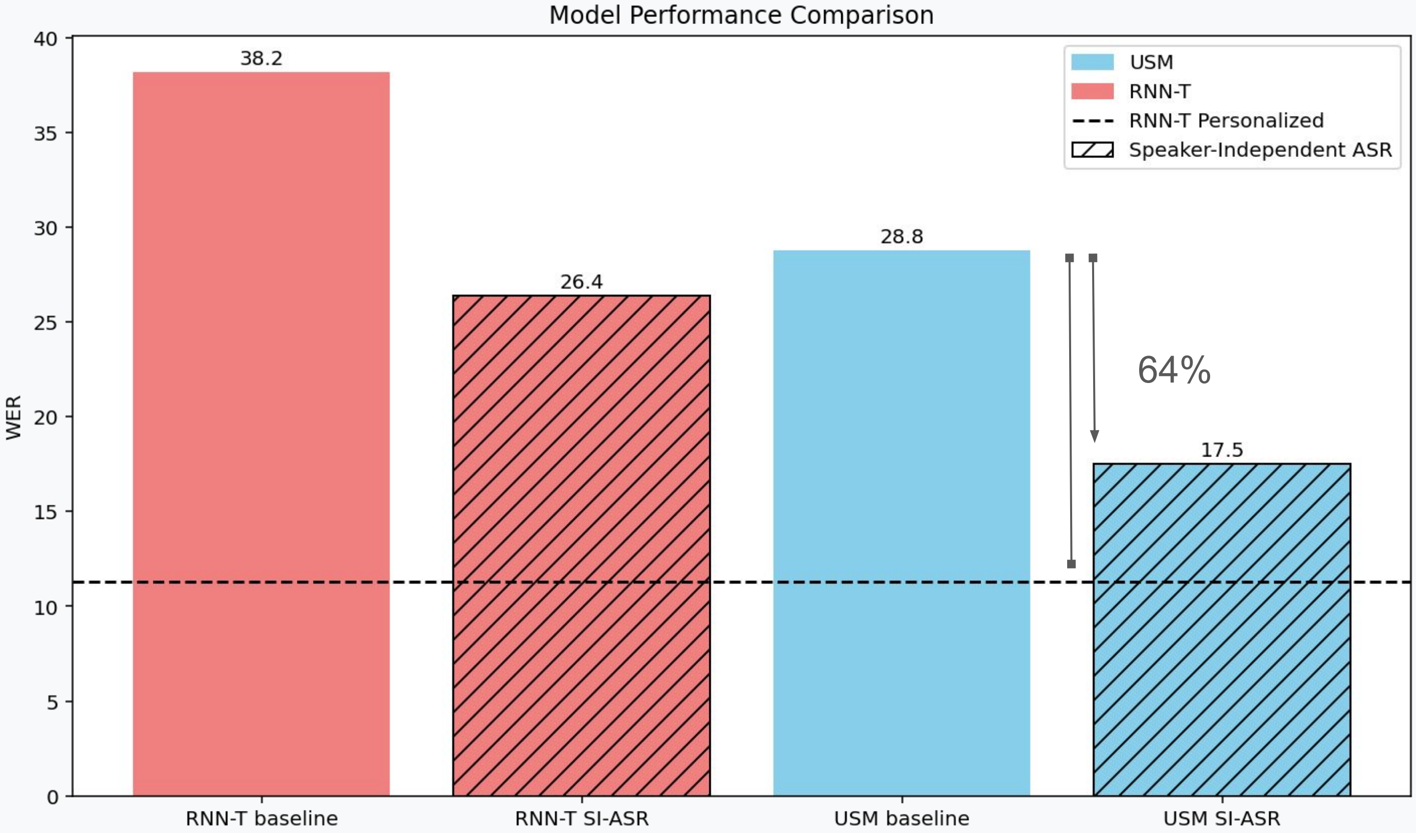
Google

With tuning on the SI-ASR train set



Google

Covers 64% of the gap to personalization



Google

Examples

Clip*	Ground Truth	Baseline (trained with YouTube)	USM	Finetuned USM
	and it's going to go back to like it was before. where you trip and think you know, that could have happened to anyone. There are a lot of things I now look back and notice	is	or that could happen anyway i and	is kind of report back to like it was before or you trip and and you think you know that could've happened anyway and a lot of things i look back and notic.
	How many people call it a day before they yet get to that point	yeah	point	how many people call a day before they get to that point.
	I now have an Xbox adaptive controller on my lap.	i now have a lot and that consultant on my mouth	i now have an xbox adapted controller on my map	i now had an xbox adapter controller on my lamp.
	I've been talking for quite a while now. Let's see.	quite a while now	quite now	i've been talking for quite a while now.

*Silence removed and gain added for presentation purposes

Summary

- **Speaker Independent ASR (SI-ASR)**: high-performing ASR model targeted at typical speech can be trained to generalize to recognize disordered speech.
- **Weight disordered speech appropriately**: It is important to weight the disordered speech data and monitor performance on several test sets.
- **No regression on standard speech evals**: We want to ensure there is no regression on standard ASR benchmarks so the same model can be used for all users to provide a good experience.
- **Covers 64% of the gap to personalization**

Thanks