Towards a Single ASR Model That Generalizes to Disordered Speech

Jimmy Tobin, Katrin Tomanek, Subhashini Venugopalan

Google Research

ICASSP 2025 HYDERABAD, INDIA — Celebrating Signal Processing

Project Euphonia: Personalized ASR

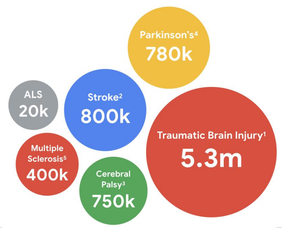
Improve ASR to help people with **speech disorders** who have difficulty being understood by other people and technology.

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

Project Relate is an app that was developed by this team to allow individuals with speech impairments personalize their own on-device model so that they can be understood better.

Condition prevalence (US)

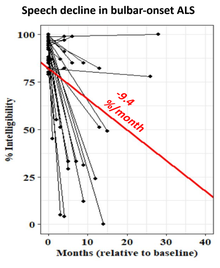
Millions of users have neurological conditions that cause speech impairments, in the US and around the world.



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 prompt can be physically demanding because of muscular weakness and fatigue. Cognitive impairment may also lead to incorrectly recorded prompts.
- **Degenerating speech intelligibility**: Diseases like A 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 sho transactional phrases, and not trained for conversations. Conversations have more varied vocabulary, are longer in length, and contain more named entities / rare words.

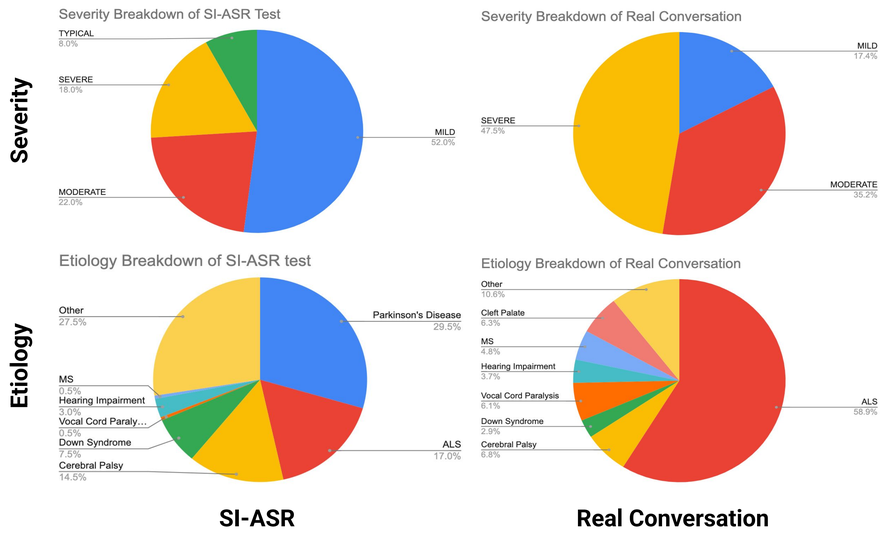


Eshghi et al., 2022

Objective: Speaker Independent ASR

- **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.

Datasets



Speaker Independent (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 both at *the speaker and phrase level* between the training validation and test sets.

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

Real Conversation Test Set

To evaluate generalization to conversational speech, we we compiled data from real world usage of Euphonia's data collection app. Speech-language pathologists

- scrubbed data of any PII,
- transcribed the speech

# Speakers	29
# Utterances	1515

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

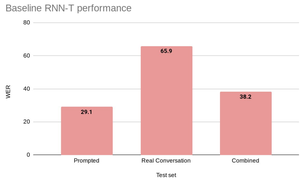
Models

Baseline on-device (RNN-T) model

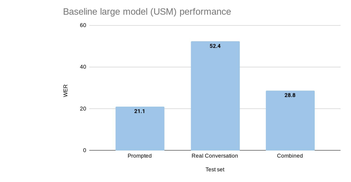
On-device model is an **RNN-T** model that was trained on a wide range of standard speech including conversational data.

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.



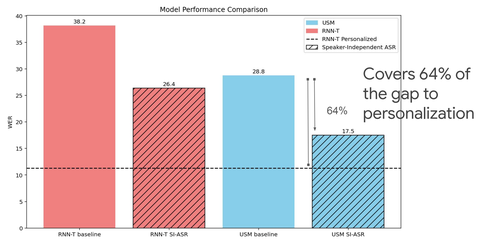
RNN-T Models Fail to Generalize to Out-of-Domain Audio: Causes and Solutions
Chung-Cheng Chiu, Arun Narayanan, Wei Han, Rohit Prabhavalkar, Yu Zhang, Navdeep Jaitly, Ruoming Pang, Tara N. Sainath, Patrick Nguyen, Liangliang Cao, Yonghui Wu
SLT 2020



Google 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

Method: Finetune and adapt on disordered speech with different weights on SI-ASR training data and other speech training datasets.

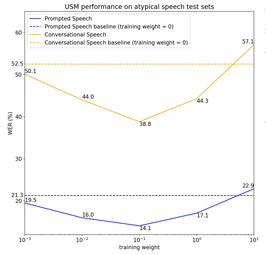
Results



Covers 64% of the gap to personalization

Tuning on the disordered speech dataset produces substantial gains for both the RNN-T and USM models adapted to the speaker independent disordered speech dataset

Weighting the SI-ASR training data



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

Training data weight	Multilingual test en-US	Multilingual test 18 langs.	Librispeech Clean	Librispeech 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

- Ensure training helps improve on all categories of disordered speech
- Observe: No regression on standard and multilingual speech eval sets.

Examples

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.

Conclusion

Given the substantial benefit of adding disordered speech data to standard ASR training datasets to help improve recognition for impaired speakers (without loss on standard eval. benchmarks), incorporating a small fraction of high quality disordered speech data in a training recipe is an easy step that could be done to make speech technology more accessible for users with speech disabilities.

ArXiv link: https://arxiv.org/abs/2412.19315