Towards Automatic Transcription of Post-Stroke Disordered Speech

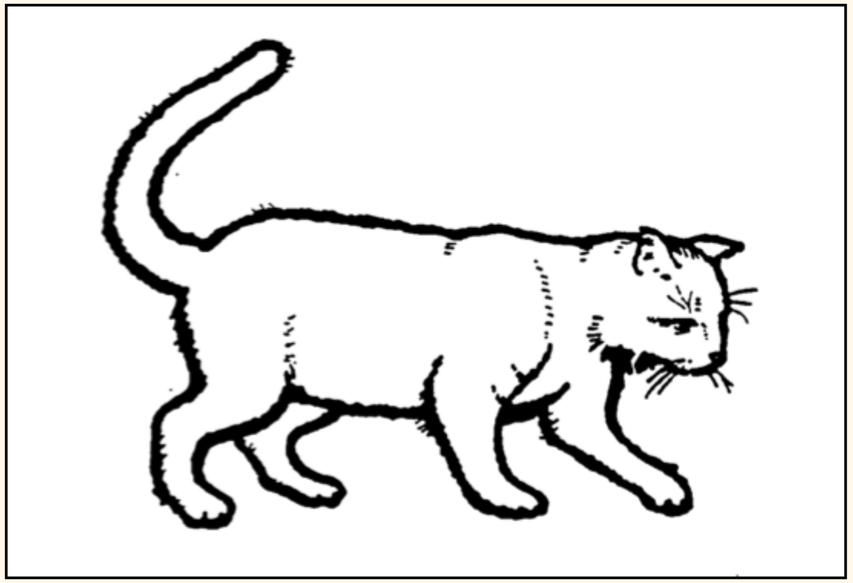
Supplement to R01DC015999, "Algorithmic Classification of Paraphasias"

PIs Gerasimos Fergadiotis and Steven Bedrick

Heroic research staff: Robert Gale and Mikala Fleegle

Portland State
UNIVERSITY

Confrontation Naming Tests are key tools in diagnosing and characterizing anomia...



"Cat"	Correct response!
"Cap"	Phonemically related!
"Dog"	Semantically related!

Key outcome: ***How many*** and ***what type*** of errors are produced?

Key measures of interest : ***How many*** and ***what types*** of errors are produced?

Using this information, we gain insight into the nature of a patient's underlying deficit...

But...

Actually *scoring* a naming test can be time-consuming and error-prone:

1. Transcribe productions (phonemically) EGAD, ALL 175 OF THEM!
2. Determine lexicality WERE THEY REALLY SAYING "FISSILE"?
3. Determine phonemic similarity

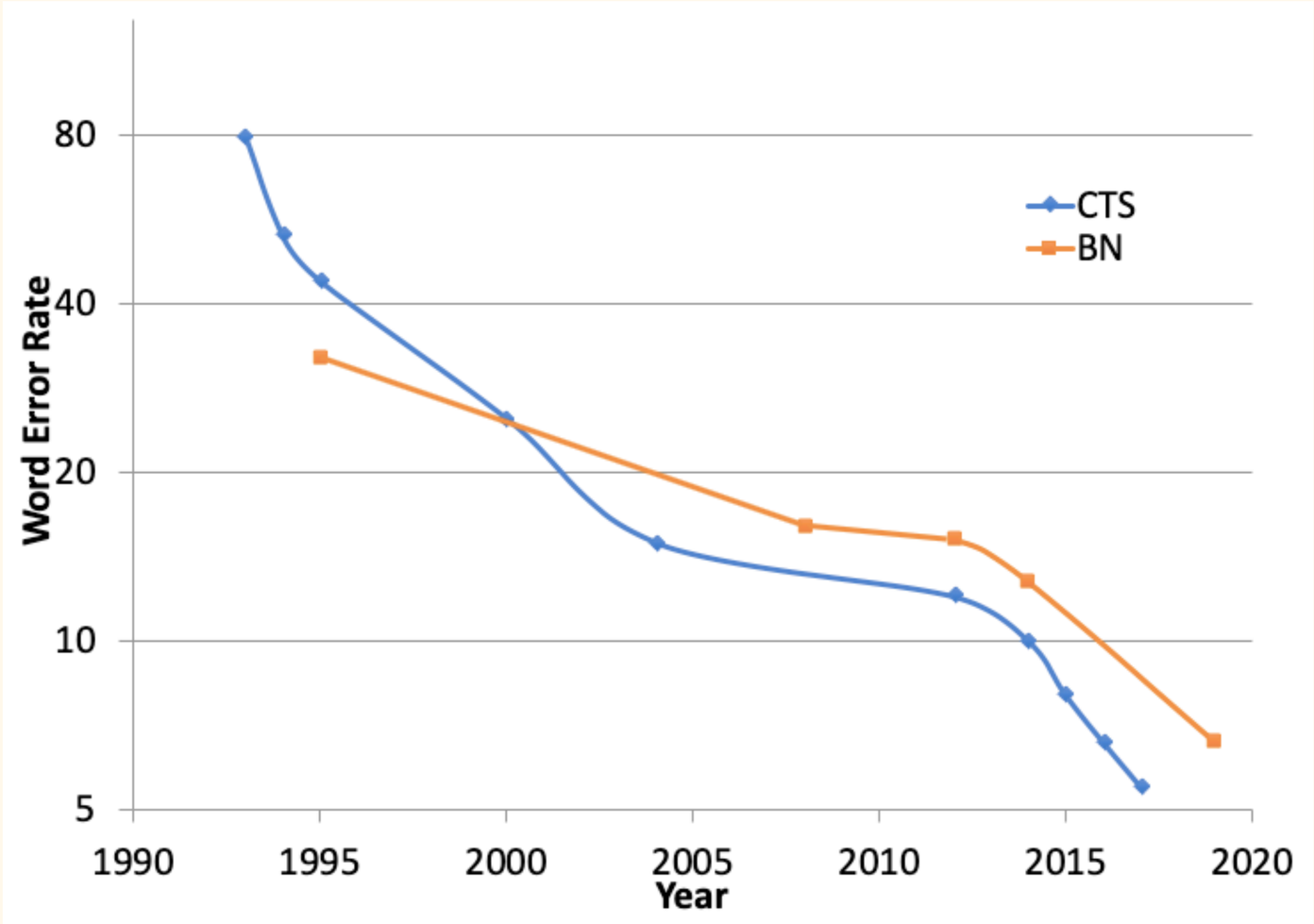
    ALIGN SYLLABLE AND WORD POSITIONS OF EACH PHONEME, BUT DON'T FORGET THE CONSONANT CLUSTERS, AND WHAT ABOUT...

4. Assess semantic similarity ARE "AMBULANCE" AND "FIREMAN" RELATED?

Actually *scoring* a naming test can be time-consuming and error-prone:

1. Transcribe productions (phonemically) EGAD, ALL 175 OF THEM!
2. Determine lexicality
3. Determine phonemic similarity
4. Assess semantic similarity

Automatic speech recognition (ASR) has improved dramatically in recent years...



https://www.ibm.com/blogs/research/2019/05/automatic-broadcast-news-captioning/

Automatic speech recognition (ASR) has improved dramatically in recent years... but:

"Mainstream" ASR	**ASR + aphasiology**
"Standard" speech characteristics	Disordered speech
Pronunciation variation = "noise" to be ignored	Pronunciation variation = vital clinical clues
Word-level output	Phoneme-level output
Unlimited* data	Very limited data

The goals of our admin supplement:

1. Produce a "shovel-ready" dataset of aphasic speech for use in ASR research
2. Establish an SOTA baseline, publish resulting model
3. Organize community shared evaluation task

Produce a "shovel-ready" dataset of aphasic speech for use in ASR research

Based on data from AphasiaBank

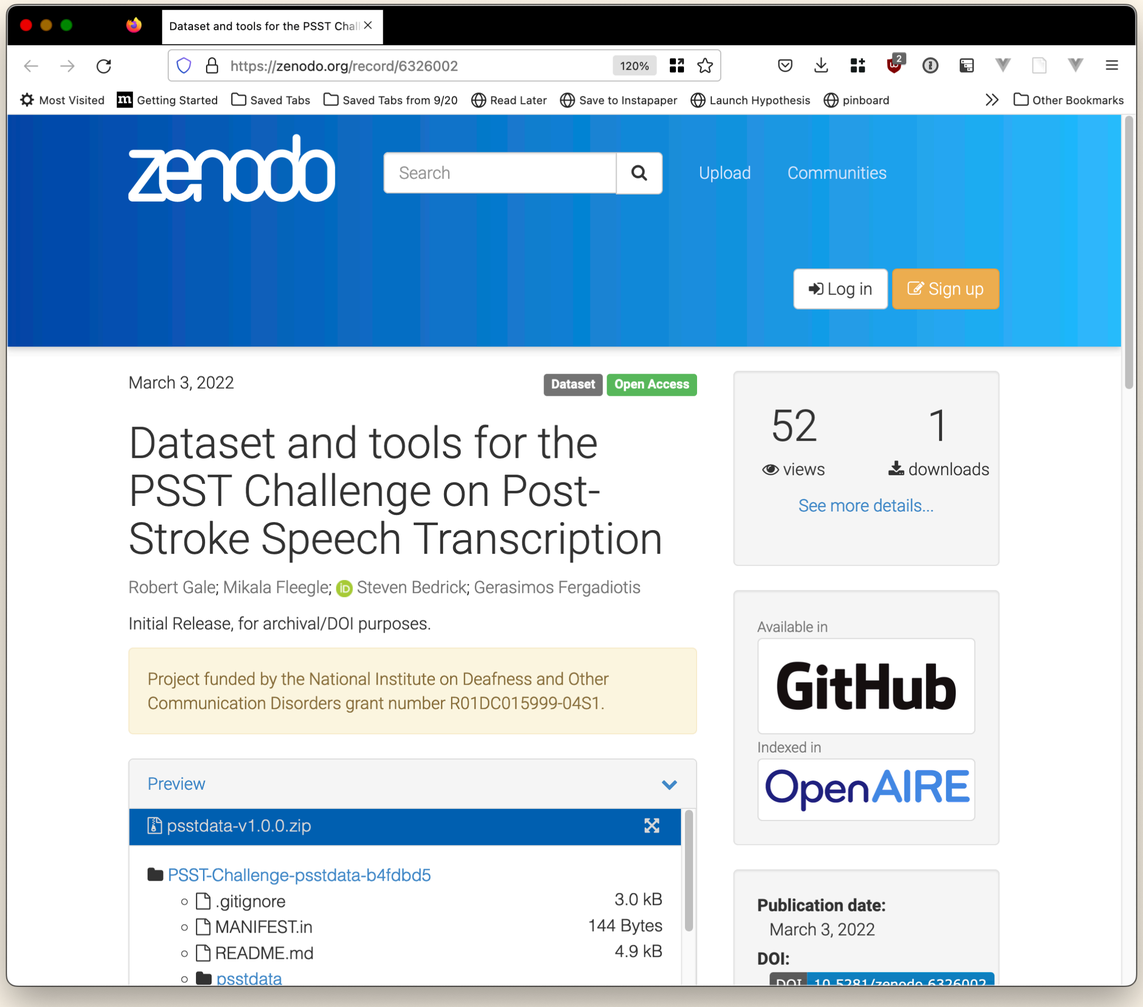
Clinical metadata & labels

Administrations of the Boston Naming Test (Short Form) and Verb Naming Test

Fine-grained, time-aligned phonemic transcription

	Train	Validation	Test
Hours	2.59 (73%)	0.36 (10%)	0.59 (17%)
Segments	2173 (70%)	325 (10%)	624 (20%)
Speakers	74 (69%)	11 (10%)	22 (21%)

Table 1: Quantities of data for each split of the PSST dataset in terms of hours of audio, number of segments, and number of speakers.



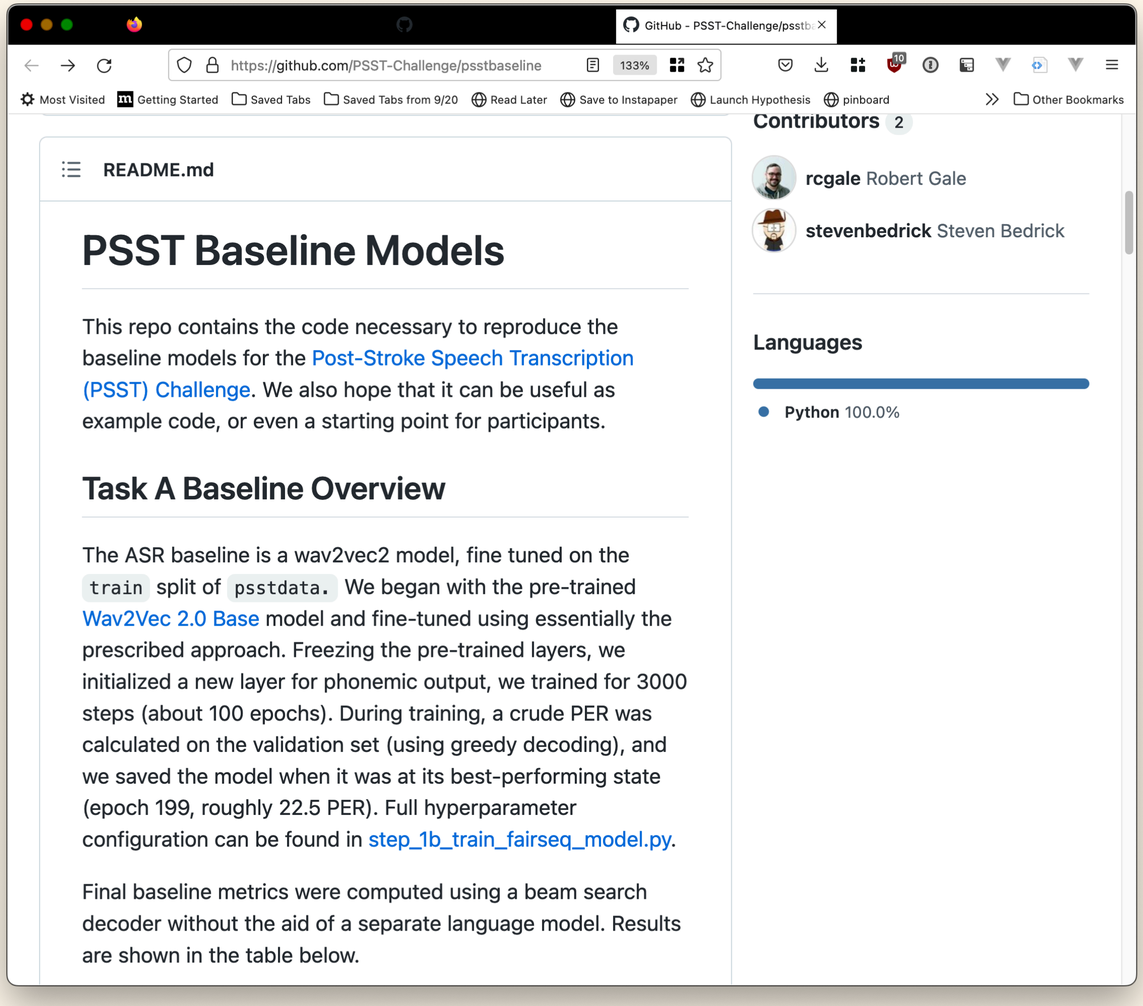
Establish an SOTA baseline, publish resulting model

Built on top of Wav2Vec2.0 architecture

Pre-trained on 960 hrs. of Librispeech, fine-tuned on our data using CTC loss

Evaluated using phoneme error rate:

Split	FER	PER
Valid	.102	.222
Test	.121	.264



Organize community shared evaluation task

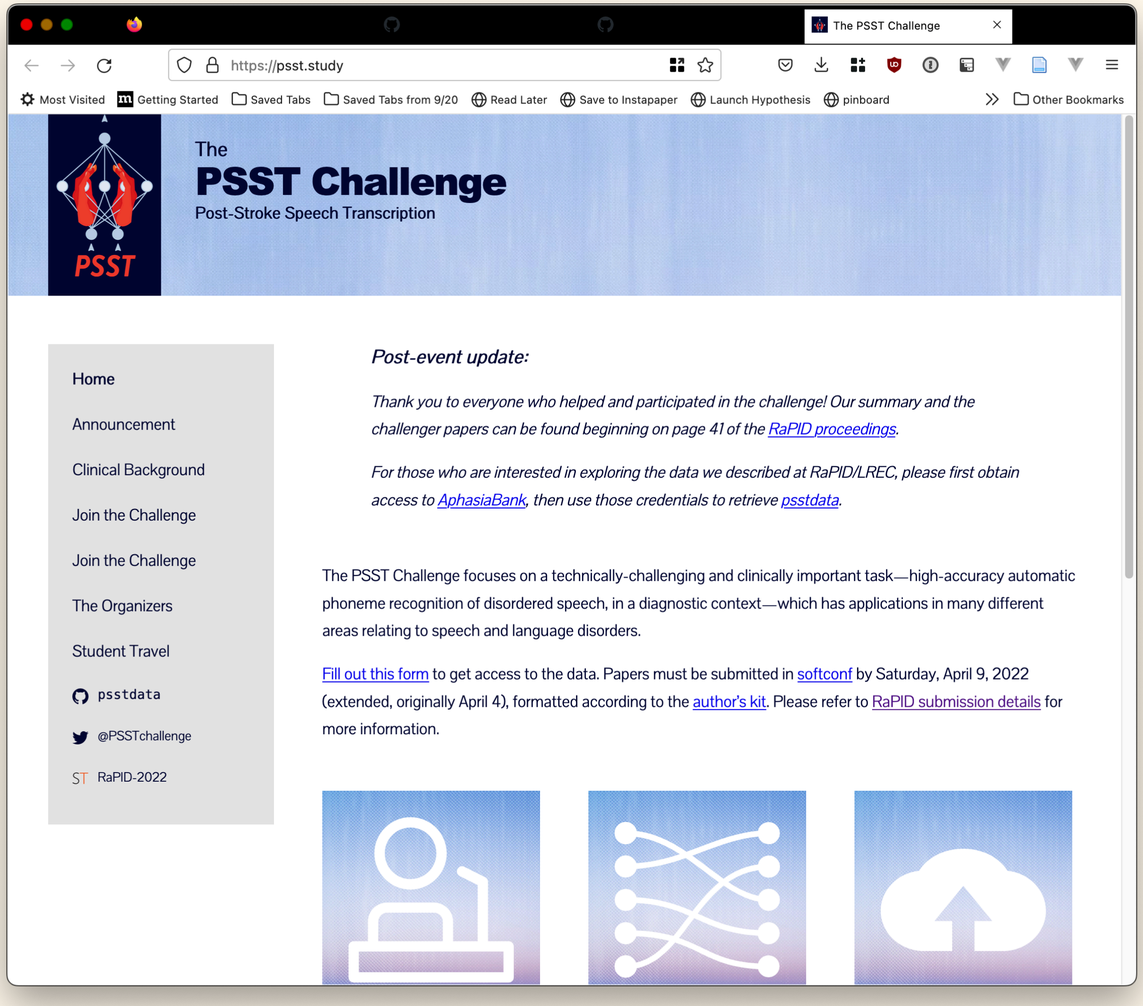
PSST Challenge: Post-Stroke Speech Transcription

Held as part of the RAPID-4 workshop at LREC 2022

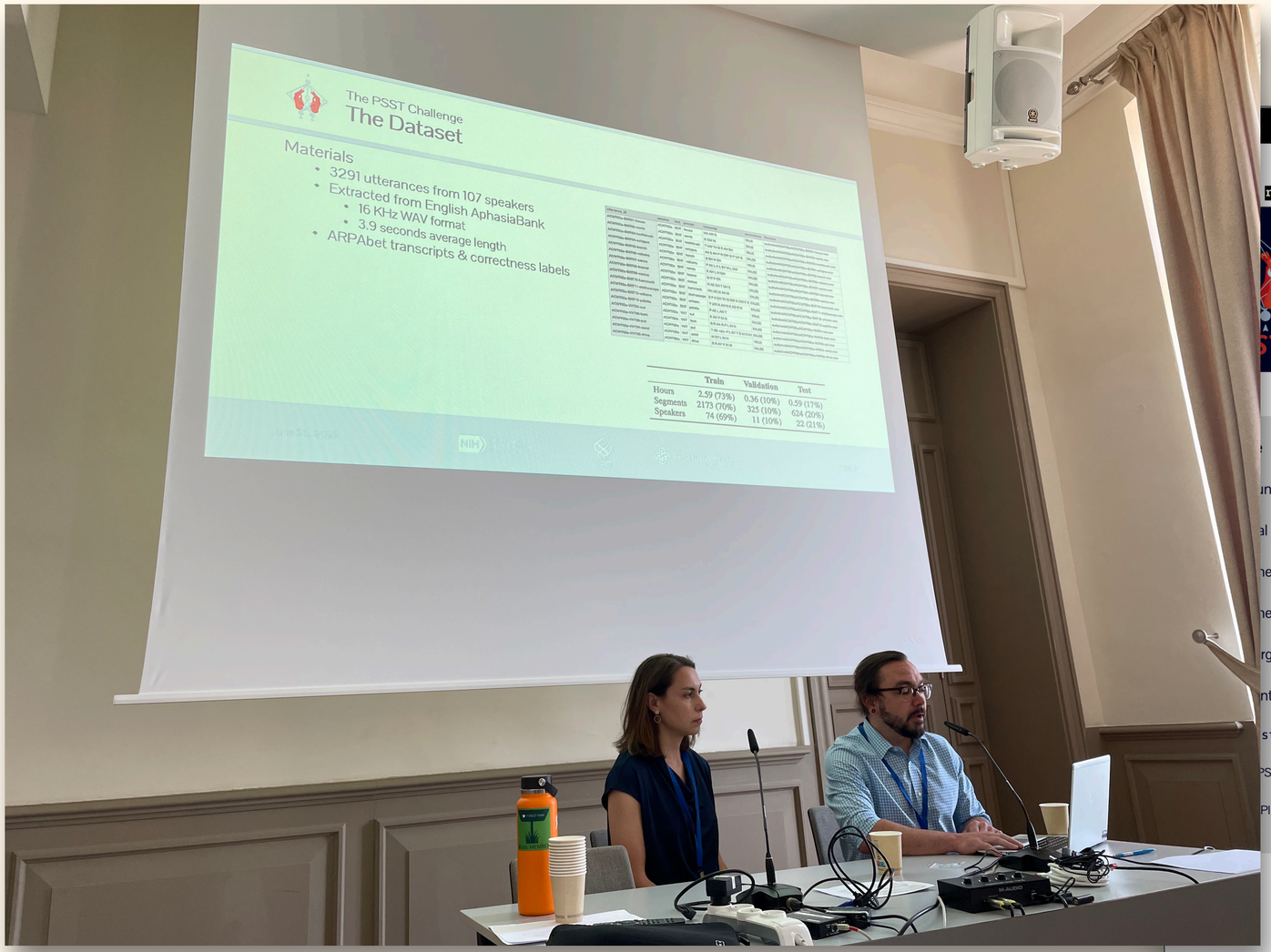
Two tasks:

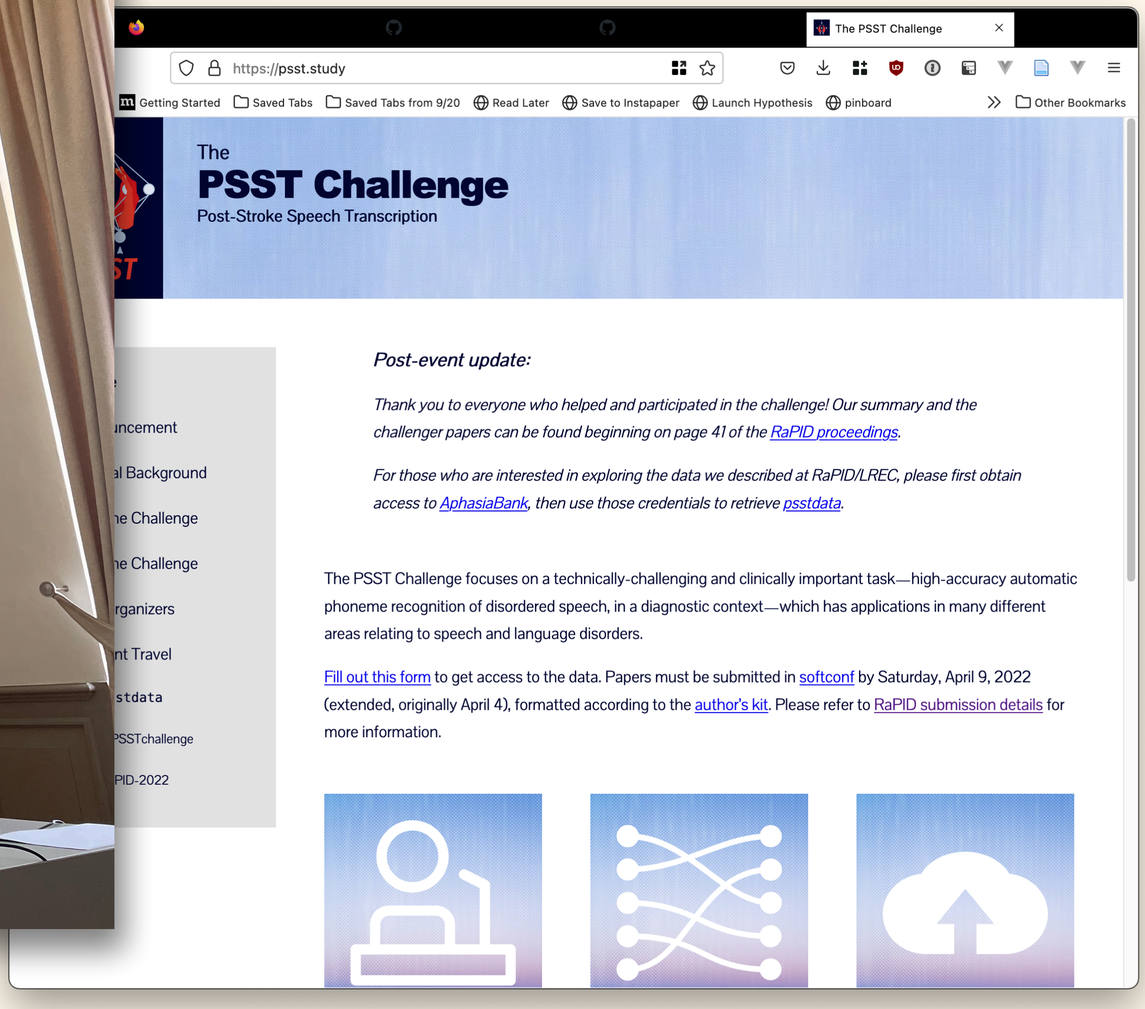
Task A: ASR

Task B: "Correctness"



Organize community shared evaluation task





Task A: Automated phoneme recognition of naming test utterances

		Data (hours of audio)					ASR	
Model	**Arch**	**Pretrain**	**PSST**	**TIMIT**	**AphasiaBank**	**Other**	**FER**	**PER**
Y1	LARGE	60,000	2.8		33.3^{U}		**9.9%**	**20.0%**
Y2	LARGE	60,000	2.8	3.9			10.3%	21.1%
Y3	LARGE	60,000	2.8		44.0^{W}		10.4%	21.5%
Y4	LARGE	60,000	2.8			3.9^{L}	10.6%	22.2%
Y5	LARGE	60,000	2.8				10.9%	22.3%
MO1	LARGE	960	2.8	1.1^{r}			11.3%	25.5%
MO2	LARGE	960	5.6^{p}				11.4%	25.1%
MO3	BASE	960	2.8	1.1^{r}			11.7%	26.3%
MO4	LARGE	960	5.6^{t}				11.7%	25.4%
MO5	LARGE	960	5.6^{p}	1.1^{r}			11.9%	26.0%
MO6	LARGE	960	2.8				12.0%	25.9%
MO7	BASE	960	5.6^{n}				12.0%	26.1%
PSST–A	BASE	960	2.8				12.1%	26.4%
Y6	LARGE	60,000	2.8			100^{L}	12.5%	26.0%
Y7	LARGE	60,000	2.8			960^{L}	16.7%	38.0%

^{L} Librispeech, pseudo-labeled with G2P
^{U} iteratively pseudo-labeled (unweighted)
^{W} iteratively pseudo-labeled (weighted)
^{p} with pitch-shifted variants
^{t} with time-shifted variants
^{n} with Gaussian noise augmentation
^{r} RIR reverb applied

Table 2: ASR results for Test set. Results are show in terms of feature error rate (FER), phoneme error rate (PER). Values in gray did not improve on *PSST–A*.

Task B: Automated determination of "correctness"

Transcripts	F1	Precision	Recall	Accuracy	FER	PER
PSST-Gold	0.984	0.968	1.000	0.985	0%	0%
Y2	**0.921**	0.941	0.901	**0.928**	10.3%	21.1%
Y5	0.920	0.926	**0.914**	0.926	10.9%	22.3%
Y1	0.917	0.941	0.894	0.925	**9.9%**	**20.0%**
Y3	0.903	**0.949**	0.861	0.914	10.4%	21.5%
Y4	0.899	0.930	0.871	0.910	10.6%	22.2%
PSST-Baseline	0.892	0.929	0.858	0.903	12.1%	26.4%
MO7	0.888	0.928	0.851	0.900	12.0%	26.1%
MO4	0.887	0.910	0.865	0.897	11.7%	25.4%
MO6	0.885	0.934	0.842	0.899	12.0%	25.9%
MO1	0.884	0.912	0.858	0.896	11.3%	25.5%
MO3	0.884	0.931	0.842	0.897	11.7%	26.3%
MO2	0.883	0.921	0.848	0.896	11.4%	25.1%
MO5	0.878	0.930	0.832	0.893	11.9%	26.0%
Y6	0.798	0.934	0.696	0.836	12.5%	26.0%
Y7	0.593	0.942	0.432	0.724	16.6%	38.0%

Table 3: Correctness results using the *PSST–B* model, using Test transcripts generated by Task A models Y1-Y7 and MO1-MO7. F1, precision, recall, and accuracy scores are shown, alongside the FER and PER shown in Task A. The first row, *PSST-Gold*, used the gold standard transcripts. Values in gray did not improve on *PSST–A*.

Future work:

Extend ASR model to additional naming tests

Integrate with the automated scoring system from parent award

Move beyond naming tests to include connected speech

PSST-2?

Other languages?

Thank you!

Robert Gale, Mikala Fleegle

Alexandra Salem

Brian MacWhinney and the rest of the AphasiaBank crew

Dimitrios Kokkinakis and the rest of the RaPID-4 organizers

Our intrepid PSST participants!

The NIH ODSS and the "AI Readiness" program!

The NIDCD!