Helping atypical speakers monitor progression and enable communication

Speech and language technologies for improved healthcare.

ICCCSP '23 Keynote, Jan. 6

Subhashini Venugopalan
Google Research

Subhashini Venugopalan

Research Scientist, Google

2017- pres.

Google

ML applications in healthcare and sciences

Large Language Models for typing suggestions.

Audio classification for monitoring disease.

Disease biomarkers from microscopy images.

Pathology (breast cancer) prediction.

Diabetic Retinopathy severity prediction.

2012-2017

Univ. of Texas at Austin

Language and vision

Video description, Image captioning

20XX

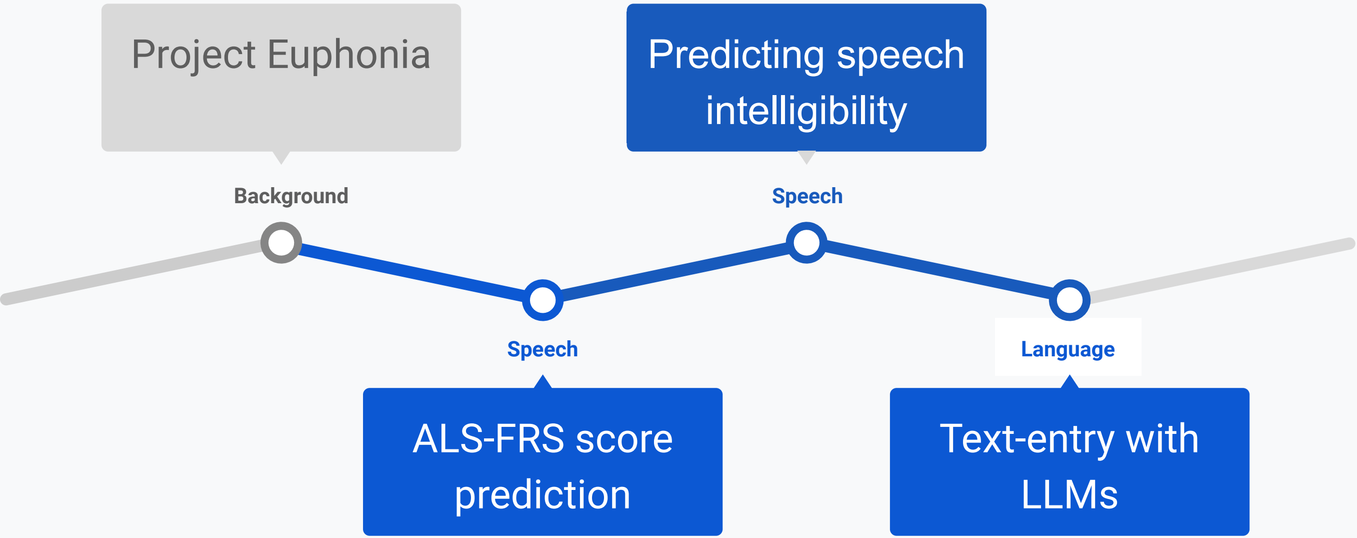
IBM Research

IIT Madras

NITK Surathkal

Outline

Google Research



Speech and language technologies for improved healthcare.

Project Euphonia

focused on helping people with atypical speech be better understood

g.co/euphonia, g.co/projectrelate

Google Research

Why study speech intelligibility?

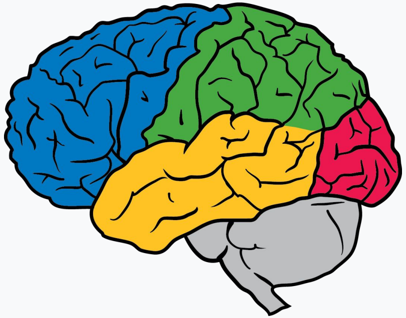
how well speech is understood by a human listener.

Will ASR on device work for you?
Or do you need a custom model?

Can users monitor deterioration?
Across different speaking disorders.

Improve YouTube transcriptions.
Collect disordered speech at scale.

ML models to measure disease severity



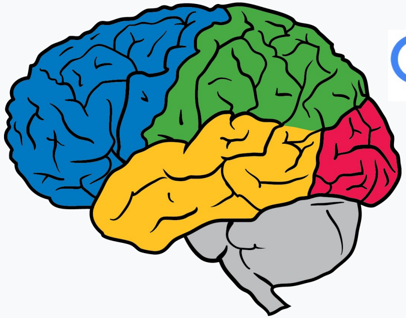
Google Research

ML models to measure disease severity

ALS
THERAPY DEVELOPMENT
INSTITUTE

+

Google Research



ALS severity
(with ALS-TDI)

A Machine-Learning Based Objective Measure For ALS Disease Severity

F. Viera*[1], S. Venugopalan*[2], A. S. Premasiri[1], M. McNally[1], A. Jansen[2], K. McCloskey[2], M. P. Brenner[2], S. Perrin[1]

npj Digital Medicine (Nature), Apr.'22

equal contribution, [1]ALS-TDI, [2]Google

We need an objective measure.

Speech/Neurological disorders have subjective rating scales

Speech (bulbar)

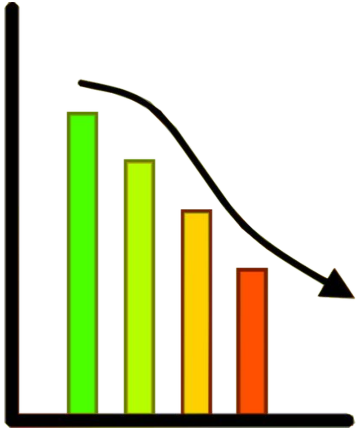
Speech is about more than how your voice sounds. It's how well you feel forming words in your mouth. Problems thinking of the right word shouldn't affect your answer to this question.

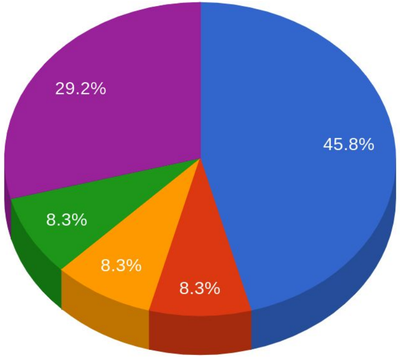
Normal speech processes	Perfectly normal compared to before you had ALS symptoms.
Detectable speech disturbance	You notice a difference in the way your voice sounds or it's harder to make sounds.
Intelligible with repeating	You need to repeat yourself because people cannot understand all of your words.
Speech combined with non-vocal communication	In addition to your voice you use non-vocal communication (writing, machines, etc.)
Loss of useful speech	Most people cannot understand you. You must use non-vocal communciation.

Image credit: https://blog.patientslikeme.com/tag/als-functional-rating-scale/

Why objective measures? monitoring progression

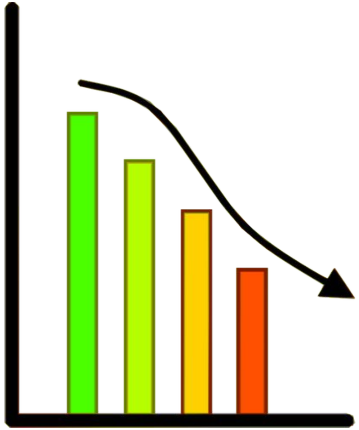
- Monitor disease progression.
- Document response to drug interventions.
- Patient stratification for clinical trials.
- Early detection of neurological disease e.g. stroke, ALS..

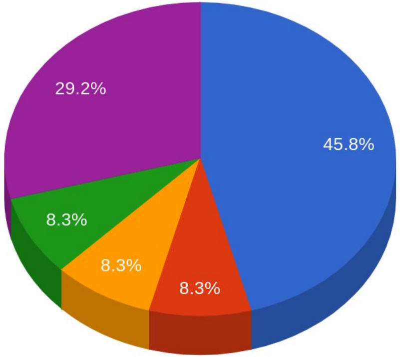




Why objective measures? monitoring progression

- ✓ Monitor disease progression.
- ✓ Document response to drug interventions.
- Patient stratification for clinical trials.
- Early detection of neurological disease e.g. stroke, ALS..





A Machine-Learning Based Objective Measure for ALS disease severity. Viera et. al.

ALS-TDI Precision Medicine Program (PMP)

PMP goal: More accurately diagnose ALS

- Enrolled 600+ people living with ALS

PMP data

- Physiological indicators - voice recordings, accelerometer measurements.
- Biological samples - skin biopsy, genome sequencing, blood-based biomarkers.
- Self-reported ALSFRS-R scores.

ALS-TDI Precision Medicine Program

ALS THERAPY DEVELOPMENT INSTITUTE

Google Research

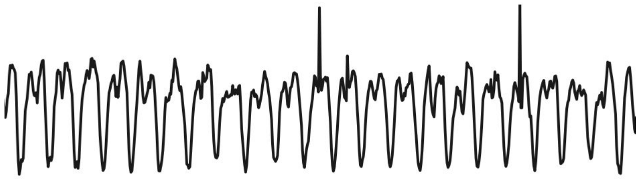
PMP goal: More accurately diagnose ALS

- Enrolled 600+ people living with ALS

PMP data

- Physiological indicators - **voice recordings, accelerometer measurements**.
- Biological samples - skin biopsy, genome sequencing, blood-based biomarkers.
- Self-reported ALSFRS-R scores.

“I owe you a yo-yo today” x 5



5 exercises involving 4 limbs + torso

ALS-TDI Precision Medicine Program (PMP)

PMP goal: More accurately diagnose ALS

- Enrolled 600+ people living with ALS

PMP data

- Physiological indicators - **voice recordings, accelerometer measurements**.
- Biological samples - skin biopsy, genome sequencing, blood-based biomarkers.
- **Self-reported ALSFRS-R scores** (scale of 0-4 for 12 functions).

Speech	Speech	Salivation	Swallowing
Limb	Handwriting	Cutting food	Climbing stairs
	Turning in bed	Walking	Dressing and hygiene
Respiratory	Dyspnea (difficulty breathing)	Orthopnea (shortness of breath while lying down)	Breathing insufficiency

Data statistics

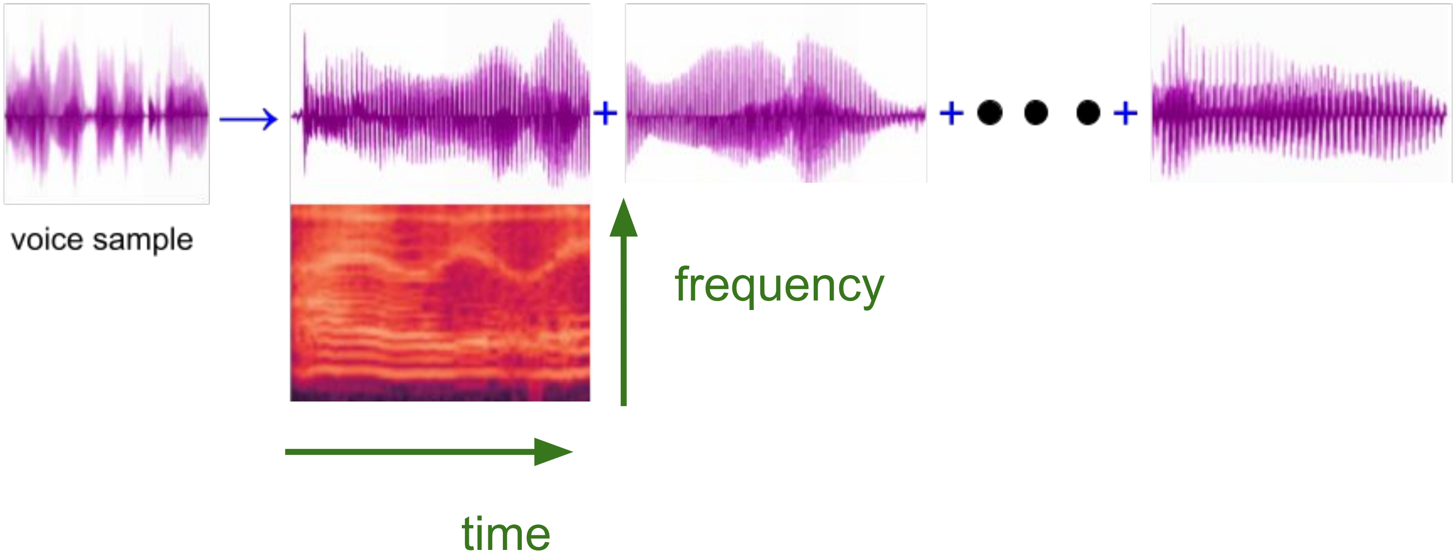
584 participants (Sep. ‘14 - Aug. ‘19)

recordings: voice - 5814, accelerometer - 13009

Split randomly by patient. (with drug participants in test set)

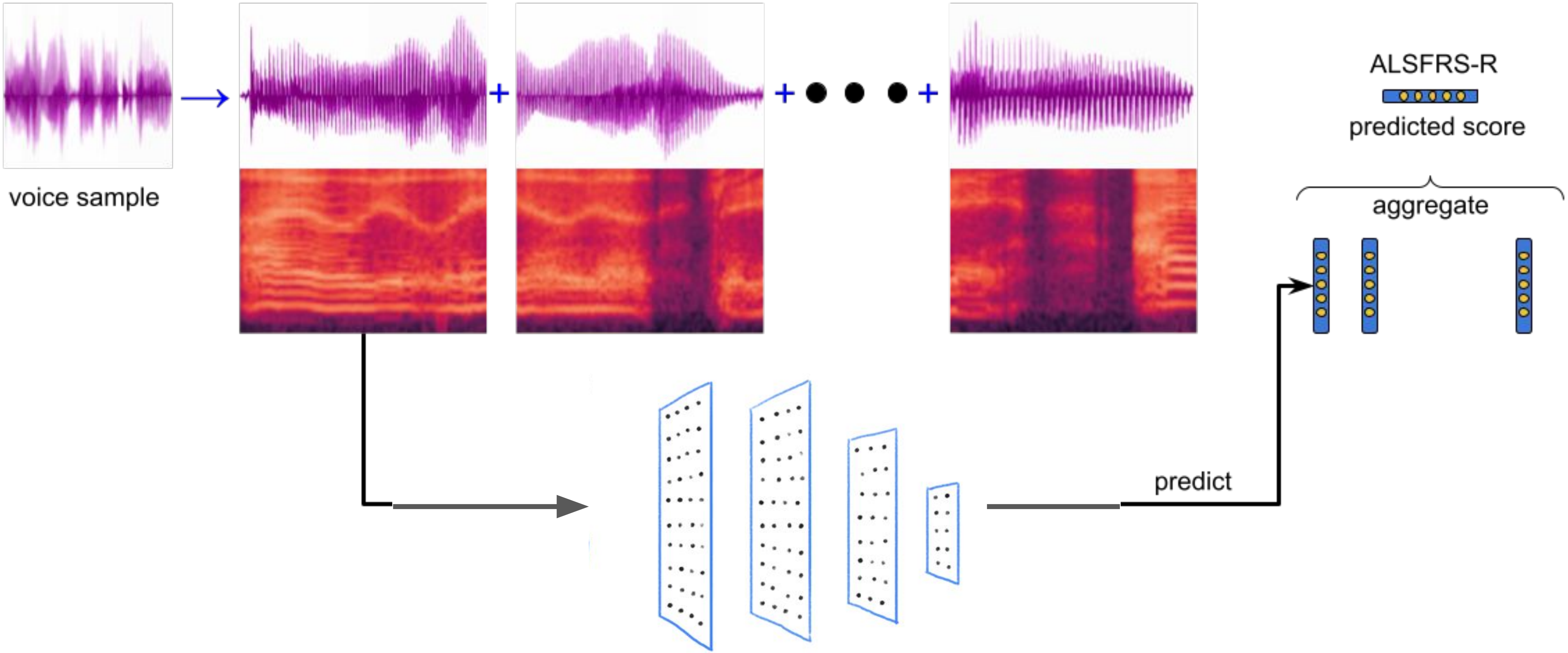
	Train	**Validation**	**Test**	**Drug cohort**
Voice participants (recordings)	389 (3776)	63 (705)	90 (150)	49 (832)
Accelerometer participants (recordings)	209 (7448)	58 (2028)	83 (3533)	44 (2061)
Age in years (standard deviation)	58.69 (11.79)	57.83 (12.15)	59.35 (10.48)	59.41 (10.59)
Sex (Male/Female)	261 / 134	52 / 28	66 / 43	33 / 21

Processing - Convert recordings to spectrograms



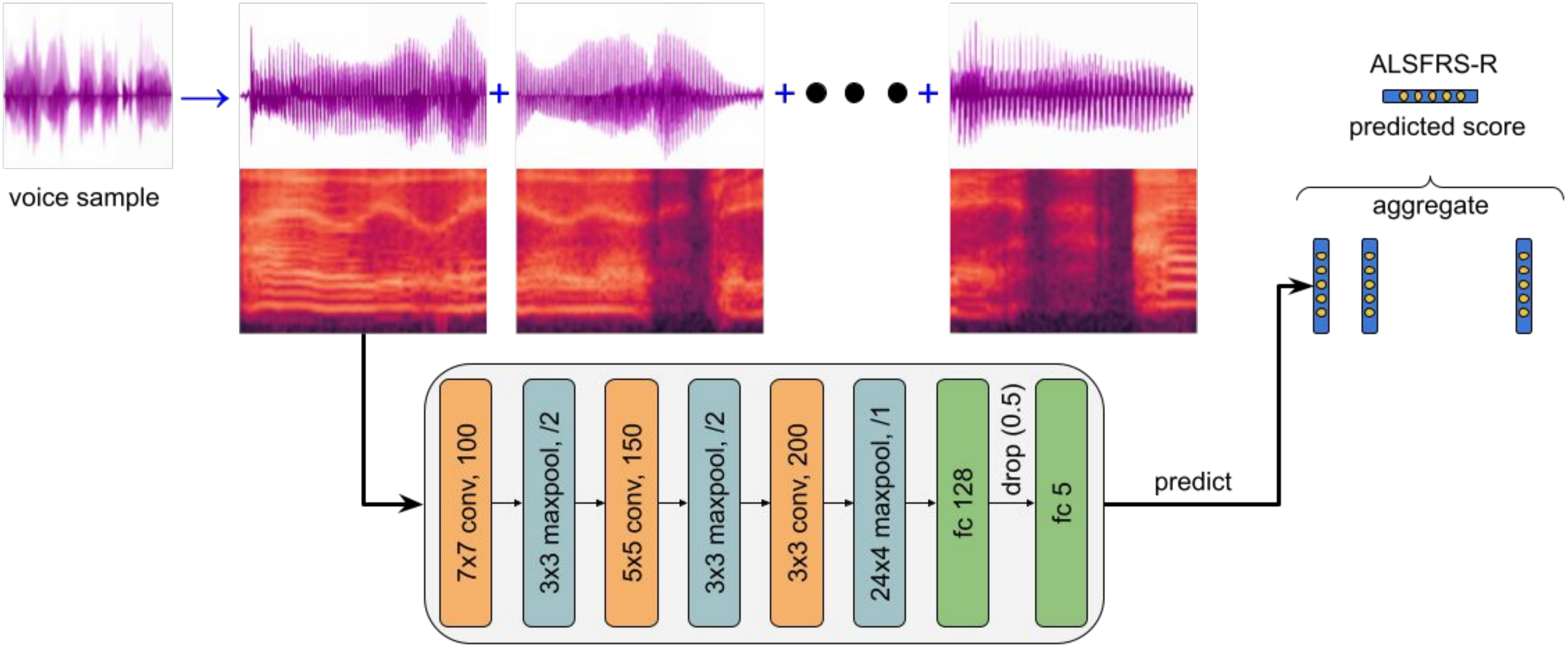
spectrogram for each ~1s non-overlapping window

Model



Google Research

CNN Model makes predictions for each ~1s window



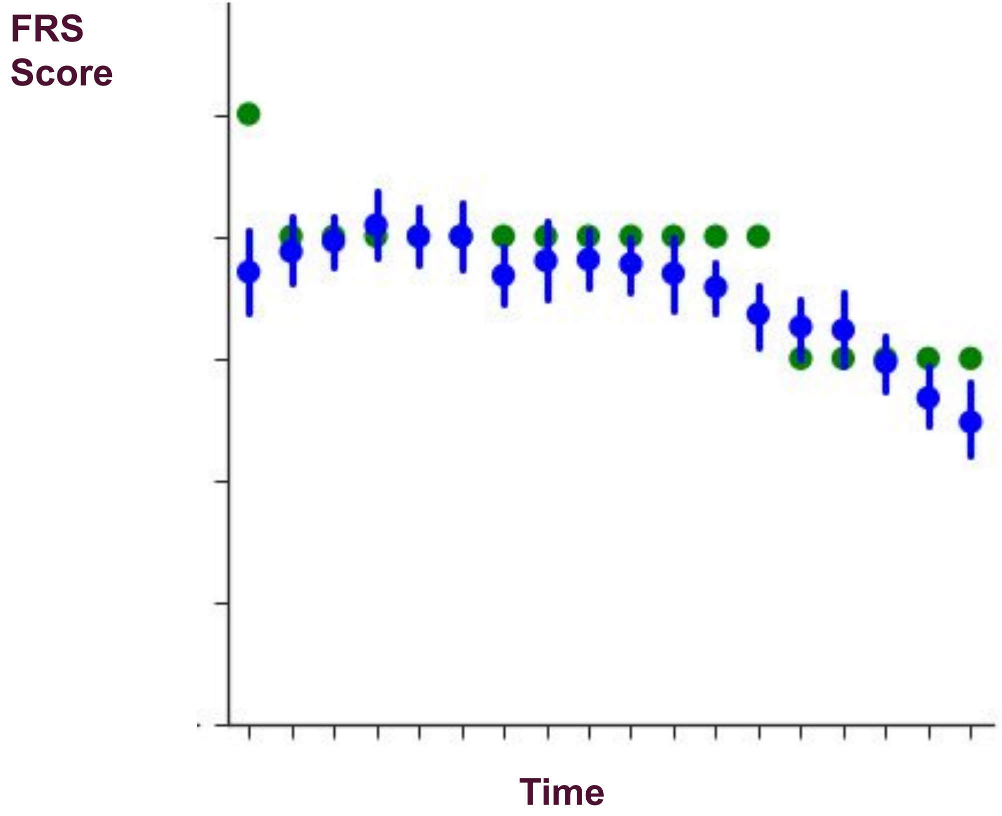
Google Research

Quantitative Results: Predicting 0-4 ALSFRS-R score

	Function	AUC	95%CI
CNN	Speech	0.865	[0.847 - 0.884]
MLP	Climbing_stairs	0.701	[0.691 - 0.712]
	Cutting_food	0.733	[0.723 - 0.743]
	Dressing_hygiene	0.729	[0.719 - 0.742]
	Handwriting	0.645	[0.634 - 0.658]
	Turning_in_bed	0.755	[0.745 - 0.766]
	Walking	0.756	[0.746 - 0.766]

Google Research

Sample test prediction



Green: Groundtruth

Blue: Model

Subjective scores - discrete integers.
Model scores - continuous so capture degradation better

Can we generalize?

ALS severity

- Recordings had 1 phrase ('I owe you a yoyo today")
- 5 point rating scale
- Self reported

ALS severity
(with ALS-TDI)

With Euphonia speakers, we want to generalize to

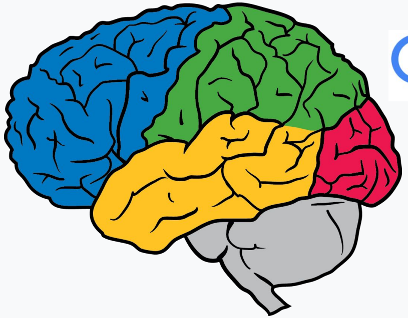
- Different phrases
- Many different underlying speech disorders

Google Research

ML models to measure disease severity

ALS
THERAPY DEVELOPMENT
INSTITUTE

Google Research



ALS severity
(with ALS-TDI)

Speech intelligibility
(with Euphonia)

Comparing Supervised Models And Learned Speech Representations For Classifying Intelligibility Of Disordered Speech On Selected Phrases
S. Venugopalan, J. Shor, M. Plakal,
J. Tobin, K. Tomanek, J. R. Green, M.P. Brenner
INTERSPEECH 2021

Google Research

Why study speech intelligibility?

how well speech is understood by a human listener.

Will ASR on device work for you?
Or do you need a custom model?

Can users monitor deterioration?
Across different speaking disorders.

Improve YouTube transcriptions.
Collect disordered speech at scale.

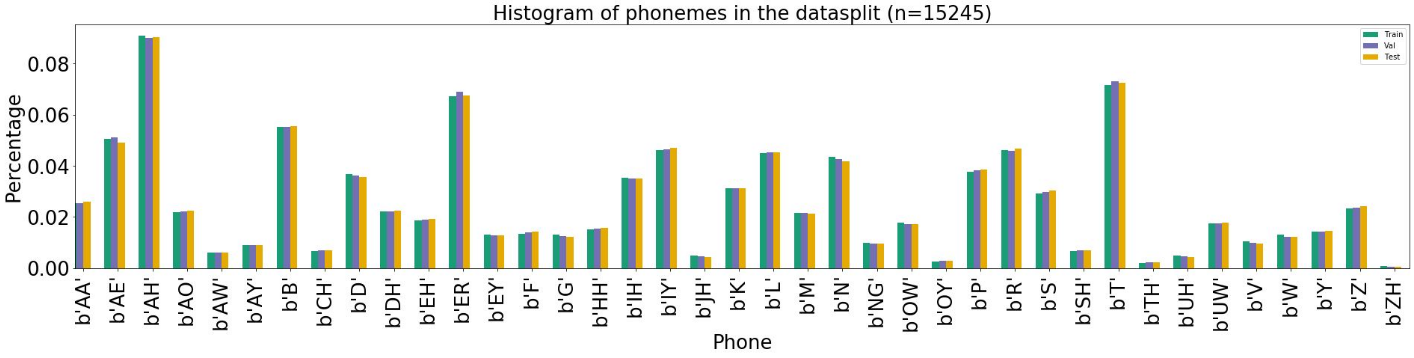
Google Research

Pilot study: Euphonia QC data - Only a tiny portion of Euphonia

```
'Buy Bobby a puppy.'
'I owe you a yo-yo today.'
'The police helped a driver.'
'The boy ran down the path.'
'The fruit came in a box.'
'The shop closes for lunch.'
'Strawberry jam is sweet.'
'Flowers grow in a garden.'
'He really scared his sister.'
'The tub faucet was leaking.'
```

```
'Sadder.'     'Banter.',
'Chatter.'    'Shatter.'
'Batter.'     'Tatter.',
'Meaner.'     'Patter.',
'Eater.'      'Ladder.'
'Manner.'     'Bladder.'
'Platter.'    'Banner.'
'Heater.'
'She looked in her mirror.'
'A match fell on the floor.'
```

```
'He said buttercup, buttercup, buttercup, buttercup all day.'
'Bamboo walls are getting to be very popular because
      they are strong, easy to use, and good-looking.'
```

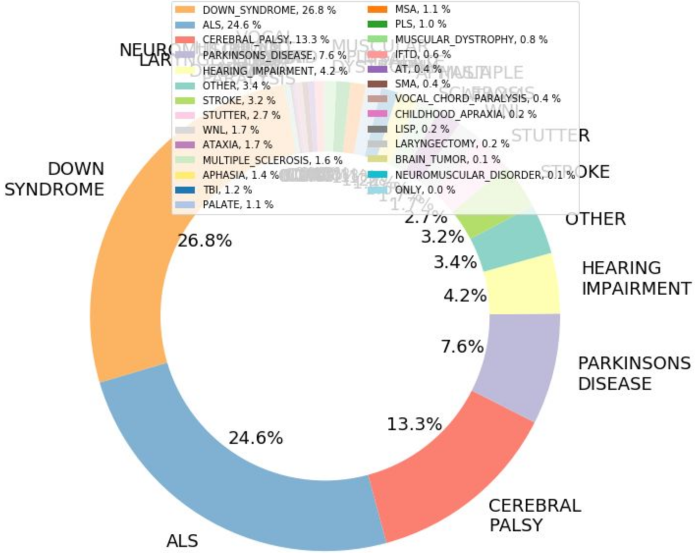


Dataset - Euphonia QC labels

5 Intelligibility classes (rated on a Likert Scale 1-5)

Intelligibility	# speakers Train	# speakers Val.	# speakers Test	# utterances Train	# utterances Val.	# utterances Test
TYPICAL	160	30	23	3,875	734	544
MILD	153	35	36	3,343	817	788
MODERATE	87	25	18	1,969	567	471
SEVERE	54	12	14	1,113	316	388
PROFOUND	10	1	3	224	9	87
OVERALL	464	103	94	10,524	2,443	2,278

Table 1: *Count of speakers and utterances in the data splits.*



Google Research

ALS severity task vs Speech intelligibility

ALS severity

- Recordings had 1 phrase (‘I owe you a yoyo today”)
- 5 point rating scale
- Self reported

ALS severity
(with ALS-TDI)

Euphonia - QC (Quality Control data)

- 29 phrases
  - From over 600 participants
- Scored by Speech and Language Pathologists (SLPs)
- 5 point scale - (severity, intelligibility, speaking rate…)

ALS severity task vs Speech intelligibility

ALS severity

- Recordings had 1 phrase ('I owe you a yoyo today")
- 5 point rating scale
- Self reported

Euphonia - QC (Quality Control data)

- 29 phrases
- Scored by Speech Language Pathologists
- 5 point scale - (severity, intelligibility, speaking rate…)
- Intelligibility
  - Measures how well speech is understood by a human listener.
  - More relevant for ASR (and possibly better correlated with ASR model WERs)

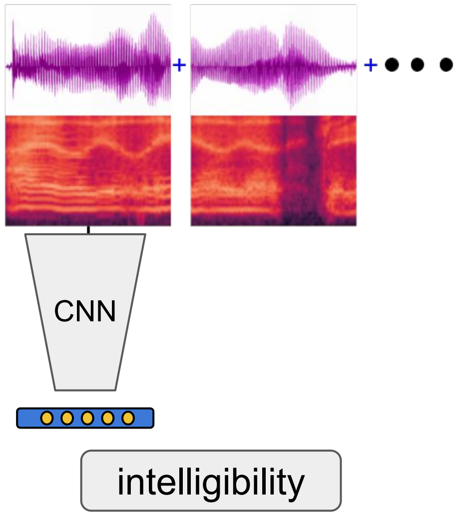
ALS severity
(with ALS-TDI)

Speech intelligibility
(with Euphonia)

… and trained classifiers based on different approaches.

Supervised CNN

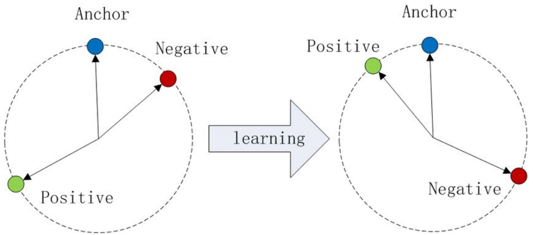
Standard for audio classification [1]



Unsupervised representations

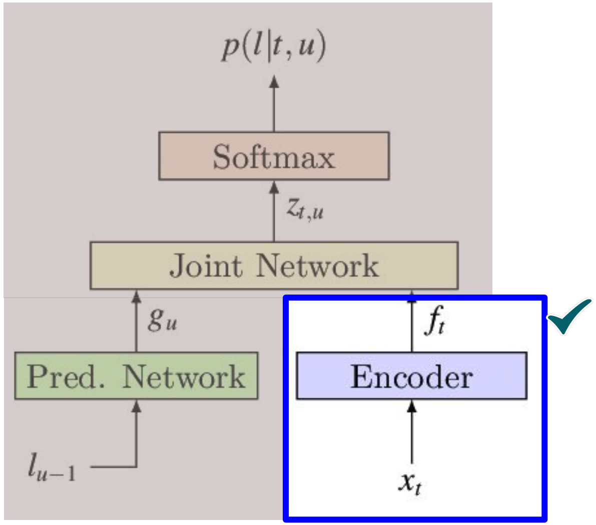
Classifiers on top of non-semantic speech representations (TRILL) [2]

(Pre-training objective)
Triplet Loss



ASR encoder representations

RNN-T model trained on typical speech [3]



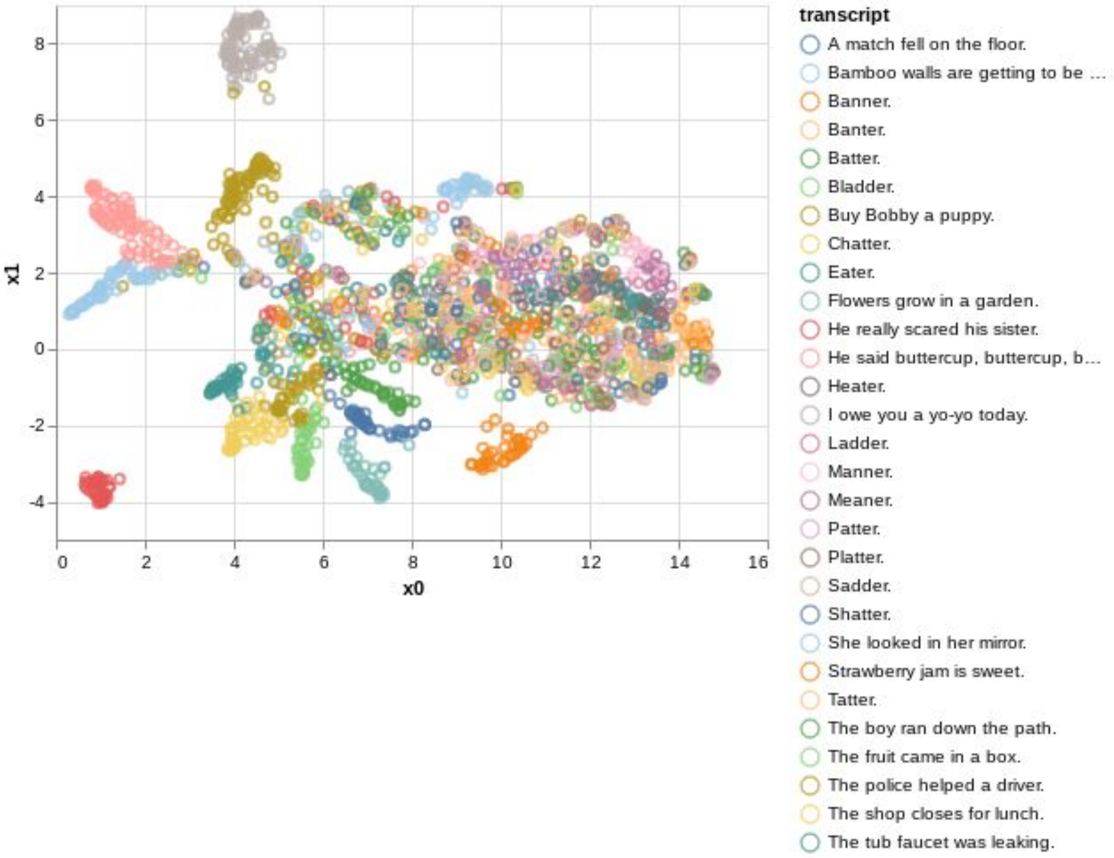
[1] Hershey et. al. CNN Architectures for Large-Scale Audio Classification ICASSP '17

[2] Shor et. al. Towards Learning a Universal Non-Semantic Representation of Speech (TRILL) INTERSPEECH '20

[3] Narayanan et. al. Recognizing longform speech in end-to-end models ASRU '19

Google Research

Pilot study: ASR encoder model generalizes quite well!



Model predicts speech intelligibility ratings.

The embeddings of the model cluster based on content of the transcript.

Key question: Can we generalize to the larger Euphonia dataset?

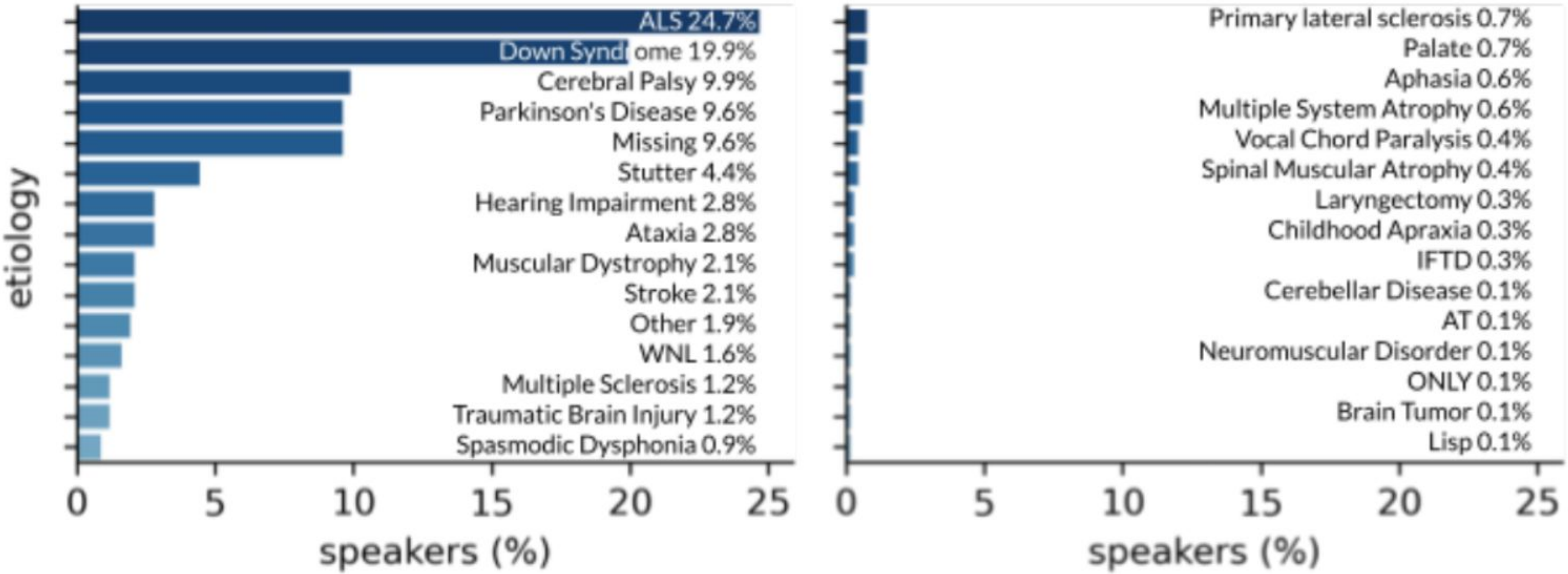
Euphonia-SpICE dataset: >750K utterances, 650+ speakers

Table 1: *Count of speakers and utterances in Euphonia-SpICE.*

Intelligibility	# speakers			# utterances		
	Train	Val.	Test	Train	Val.	Test
TYPICAL	161	41	25	149,941	24,142	10,664
MILD	161	29	37	208,843	22,532	39,007
MODERATE	83	23	19	124,984	48,814	21,214
SEVERE	54	12	15	60,692	13,868	22,397
PROFOUND	9	4	4	6,716	1,691	642
OVERALL	468	109	100	551,176	111,047	93,924

All roughly similar distribution

The Euphonia-SpICE dataset: Diverse etiologies

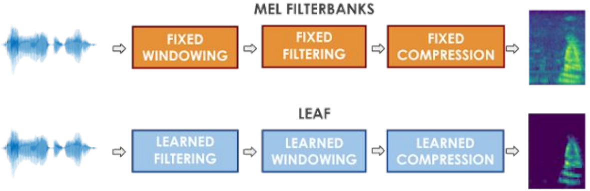


Distribution of etiologies in the SpICE dataset.

We wanted an open-sourceable model competitive to ASR encoder

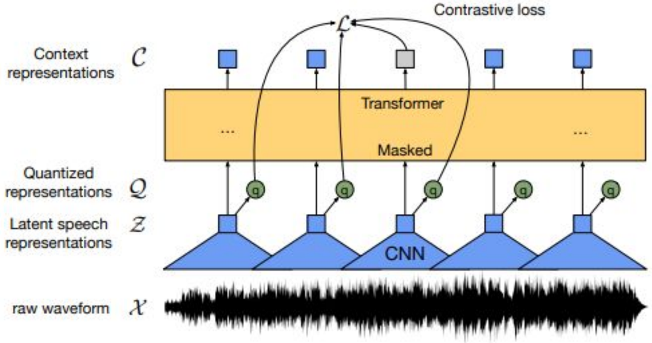
LEAF + CNN

Learnable frontend [4]



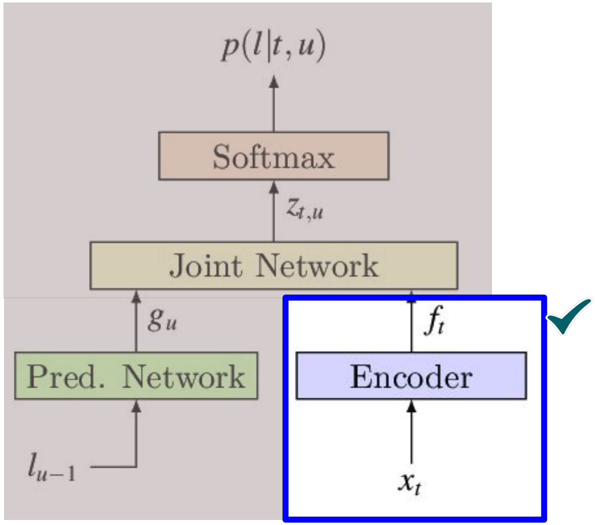
wav2vec2

Transformer+CNN [5] and is open-source and includes model weights.



ASR encoder representations

RNN-T model trained on typical speech [3]



[4] LEAF: A Learnable Frontend for Audio Classification ICLR '21

[5] wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations NeurIPS '20

[3] Narayanan et. al. Recognizing longform speech in end-to-end models ASRU '19

Classification tasks and metrics

2 class MILD+: 0:{TYPICAL}, 1: {MILD, MODERATE, SEVERE, PROFOUND}

5 class classification tasks

AUC, F1 and Acc. as evaluation metrics

		Euphonia-QC (dataset in [20])						Euphonia-SpICE dataset					
	Size	**2-class MILD+**			**5-class**			**2-class MILD+**			**5-class**		
Models	(MB)	AUC	F1	Acc.	AUC	F1	Acc.	AUC	F1	Acc.	AUC	F1	Acc.

ASR-enc does best closely followed by wav2vec2

2 class and 5 class classification tasks

AUC, F1 and Acc. as evaluation metrics

Table 2: *We report the mean 1-vs-rest AUC values, F1 score, and accuracy (Acc.) for the models on the two classification tasks when trained and evaluated on the Euphonia-QC and Euphonia-SpICE datasets. Higher is better.* **bold** *indicates highest value.*

		Euphonia-QC (dataset in [20])						Euphonia-SpICE dataset					
	Size	**2-class MILD+**			**5-class**			**2-class MILD+**			**5-class**		
Models	(MB)	AUC	F1	Acc.	AUC	F1	Acc.	AUC	F1	Acc.	AUC	F1	Acc.
LEAF + CNN	55	0.750	0.751	0.759	0.644	0.413	0.421	0.669	0.833	**0.886**	0.600	0.362	0.378
wav2vec 2.0	360	0.794	0.744	0.739	0.564	0.138	0.300	0.742	0.857	0.863	0.652	0.416	0.423
ASR-enc	122	**0.820**	**0.776**	**0.776**	**0.771**	**0.448**	**0.459**	**0.761**	**0.861**	0.862	**0.714**	**0.422**	**0.432**

SpICE models do well on ALS, CP and PD at speaker level

Table 4: *Performance sliced by etiology. The wav2vec 2.0 and ASR-enc. show identical per-speaker accuracy and similar AUC per-utterance and have low scores on Down Syndrome and PD.*

Etiology	# Utts. (%)	atyp./total # Spkr	per-utterance AUC wav2vec 2.0	per-utterance AUC ASR-enc.	Spkrs. Acc
ALS	22076 (23.7)	14 / 18	0.749	0.763	0.778
CP	14518 (15.6)	11 / 12	0.890	0.916	0.834
Down Syn.	13971 (15.0)	18 / 23	0.544	0.525	0.652
PD	13863 (14.9)	8 / 11	0.489	0.521	0.727
Hearing Imp.	8478 (9.1)	5 / 5	NA	NA	1.000
MS	6272 (6.7)	3 / 4	0.842	0.942	0.750
Musc. Dystr.	2544 (2.7)	1 / 3	0.935	0.958	0.667

Table 5: *Accuracy sliced by intelligibility class.*

5-class	Typical	Mild	Mod.	Severe	Profound
LEAF + CNN	0.386	0.469	0.493	0.118	0.000
wav2vec 2.0	0.366	0.604	0.329	0.236	0.016
ASR-enc	0.459	0.623	0.313	0.223	0.003

Why is evaluating generalization important?

- A review paper, Huang et al., 2021, shows many existing works tested/trained on same speakers; most at best use different speakers within same dataset; a handful train and test across datasets
- Comparison with SOTA ASR-error-rate-based approaches
- Evaluate/demonstrate generalization to realistic setting & etiologies not well represented in the Euphonia-SpICE train dataset

Generalization to TORGO dataset

- 7 speakers with either cerebral palsy (CP) or ALS, ~100 utterances per speaker
- We collected our own SLP intelligibility labels

Speaker	# Utts.	TORGO label	SLP label
FC01	26	Control	typical
FC02	122	Control	typical
FC03	125	Control	typical
MC01	118	Control	typical
MC02	122	Control	typical
MC03	119	Control	typical
MC04	121	Control	typical
F03	100	a	mild
F04	97	a	typical
M03	92	a	typical
F01	20	d/e	moderate
M02	92	d/e	moderate
M04	86	d/e	severe
M05	17	c	severe

Utterance prompts:

F03 *yet he still thinks as swiftly as ever.*

F04 *Both figures would go higher in later years.*

F01 *A long, flowing beard clings to his chin,*

M05 *This was easy for us.*

ASR-enc and wav2vec2 generalize out-of-the-box.

Table 3: *Generalization (only inference) on the TORGO database. Per-speaker predictions and (binarized accuracy %).*

Speaker	# Utts.	TORGO label	SLP label	SpICE 5-cls models LEAF+CNN	wav2vec 2.0	ASR-enc
FC01	26	Control	typical	typ. (34.6)	typ. (96.2)	typ. (96.2)
FC02	122	Control	typical	typ. (68.9)	typ. (95.9)	typ. (100)
FC03	125	Control	typical	typ. (65.6)	typ. (83.2)	typ. (78.4)
MC01	118	Control	typical	typ. (55.1)	typ. (96.6)	typ. (92.4)
MC02	122	Control	typical	sev. (22.1)	typ. (94.3)	typ. (92.6)
MC03	119	Control	typical	typ. (75.6)	typ. (98.3)	typ. (98.3)
MC04	121	Control	typical	mod. (5)	typ. (98.3)	typ. (99.2)
F03	100	a	mild	typ. (63)	mild (87.0)	mild (88.0)
F04	97	a	typical	mod. (8.2)	typ. (91.8)	typ. (74.2)
M03	92	a	typical	mod. (15.2)	typ. (98.9)	typ. (100)
F01	20	d/e	moderate	mod. (85)	mod. (100)	mod. (100)
M02	92	d/e	moderate	mod. (92.4)	mild (100)	mild (100)
M04	86	d/e	severe	mod. (59.3)	sev. (100)	mod. (100)
M05	17	c	severe	typ. (41.2)	sev. (100)	mod. (100)

- both wav2vec 2.0 and ASR-enc generalize well on all 14 speakers in TORGO

Utterance prompts:

F03 *yet he still thinks as swiftly as ever.*

F04 *Both figures would go higher in later years.*

F01 *A long, flowing beard clings to his chin,*

M05 *This was easy for us.*

Generalization to ALS-TDI test set

- speakers with ALS
- "I owe you a yoyo today" 5x
- 90 test spkrs, ~1330 recordings, ~4yrs
- Self-reported speech severity scores
- CNN trained on ~400 speakers
- AUC: 0.86

Speech		Predicted Scores					
		0	1	2	3	4	Total
Ground-truth scores	0	13	7	0	0	0	20
	1	29	41	14	5	14	103
	2	5	20	80	33	3	141
	3	6	12	35	213	110	376
	4	8	0	6	58	620	692

npj | digital medicine

Explore content ⌄ About the journal ⌄ Publish with us ⌄

nature > npj digital medicine > articles > article

Article | Open Access | Published: 08 April 2022

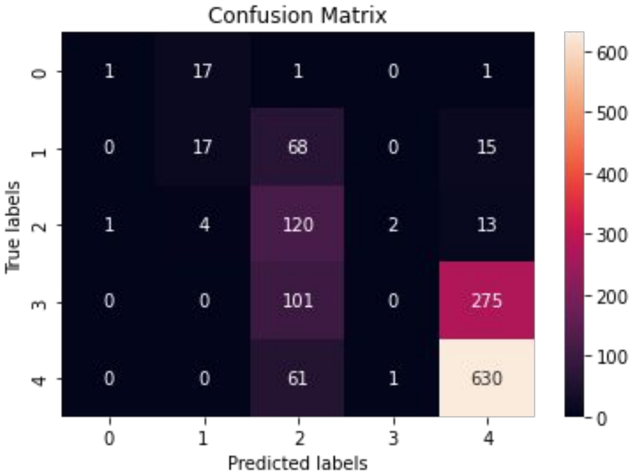
A machine-learning based objective measure for ALS disease severity

Fernando G. Vieira, Subhashini Venugopalan, Alan S. Premasiri, Maeve McNally, Aren Jansen, Kevin McCloskey, Michael P. Brenner & Steven Perrin

ASR-enc and wav2vec2 generalize for typical vs atypical.

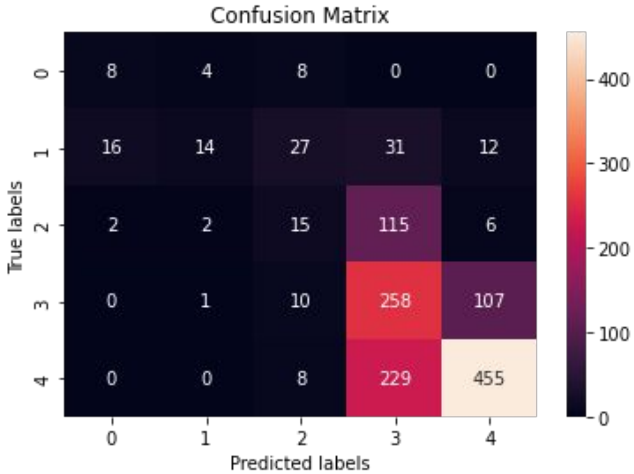
Both come close to existing model performance (0.86 AUC) but require no additional training

ASR-enc (AUC 0.82)



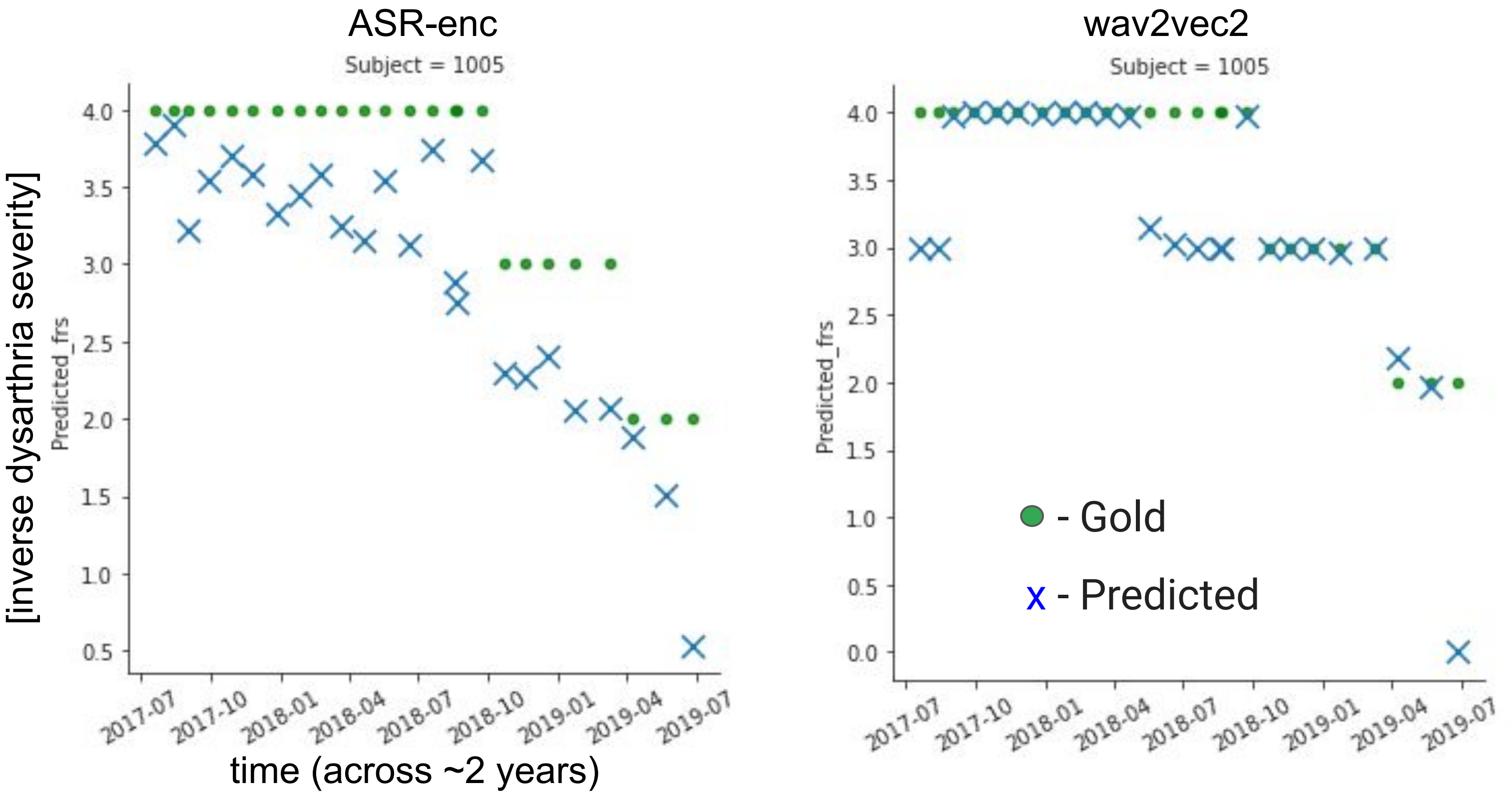

```
{'accuracy': 0.39750949628406274,
 'auc': 0.8242097327174441,
 'dprime': 1.317379457849476,
 'eer': 0.2325214845069323,
 'map': 0.4919124773837029}
```

W2V2 (AUC 0.81)




```
{'accuracy': 0.3981655041551734,
 'auc': 0.8128091281156599,
 'dprime': 1.256239743061756,
 'eer': 0.2513251956935751,
 'map': 0.4323433106527146}
```

ASR-enc and wav2vec2 generalize to longitudinal data



Generalization to UASpeech dataset

- 28 consented speakers: 15 CP, 13 controls
- 765 isolated words per speaker
- percentage intelligibility label
- ASR-error-rate-based model:
  - SOTA model, no training required
  - 0.98 correlation

Table 7. Performance of the proposed and state-of-the-art measures on the UA Speech corpus.

Method	Pearson Correlation
Martinez et. al.[23]	0.91
Hummel [24]	0.92
Janbakhshi et. al.[25]	0.95
Paja et. al.[26]	0.96
DS + I_{sm} (Proposed)	0.96
DS + I_{ld} (Proposed)	**0.98**

Tripathi et al., 2020

Table 1. Summary of speaker information (UA Speech Corpus [21]).

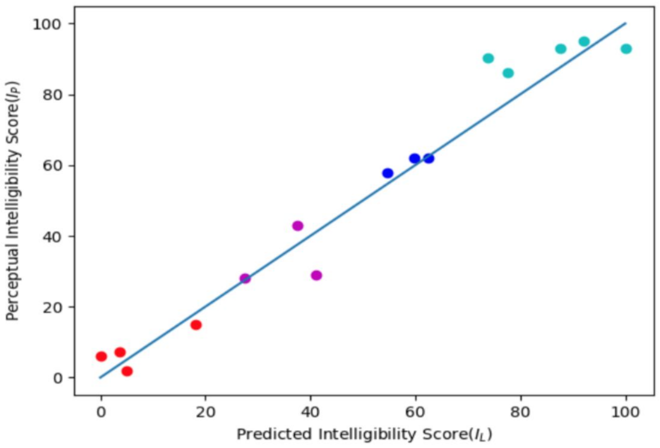
Spk ID	Age	Speech Intelli (I_p)	Dysarthria Diag
M04	>18	Very low (2%)	Spastic
F03	51	Very low (6%)	Spastic
M12	19	Very low (7.4%)	Mixed
M01	>18	Very low (15%)	Spastic
M07	58	Low (28%)	Spastic
F02	30	Low (29%)	Spastic
M16	-	Low (43%)	Spastic
M05	21	Medium (58%)	Spastic
M11	48	Medium (62%)	Athetoid
F04	18	Medium (62%)	Athetoid
M09	18	High (86%)	Spastic
M14	40	High (90.4%)	Spastic
M08	28	High (93%)	Spastic
M10	21	High (93%)	Mixed
F05	22	High (95%)	Spastic

wav2vec2 (somewhat) generalizes UASpeech

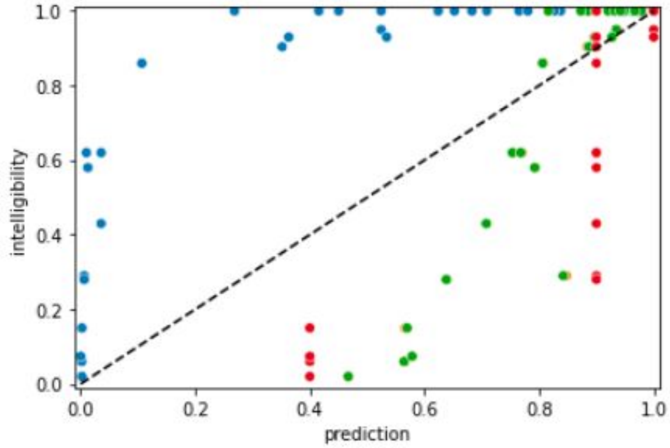
- UASpeech access - only to academia
- wav2vec2 5-class prediction's "typical" class prob. taken as predicted % intelligibility
- Simple map from 5-class prediction to percentage
  - {typical: 100, mild: 90, moderate: 60, severe: 40, profound: 20}

wav2vec2

Tripathi et al., 2020 **0.98**



15 non-controls and 13 controls: Green dots **0.93**



Why is evaluating generalization important?

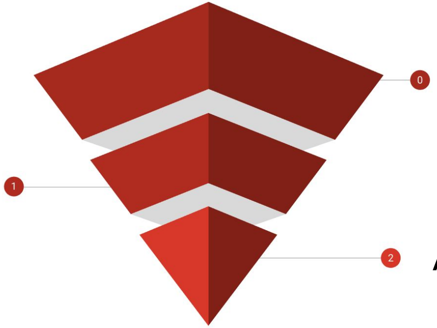
- A review paper, Huang et al., 2021, shows many existing works tested/trained on same speakers; most at best use different speakers within same dataset; a handful train and test across datasets
- Comparison with SOTA ASR-error-rate-based approaches
- Evaluate/demonstrate generalization to realistic setting & etiologies not well represented in the Euphonia-SpICE train dataset

SpICE-V benchmark dataset

SpICE-V data collection : 106 Dysarthric videos

2 Run *a different* binary classifier to tag “regions of interest” (ROIs)

ASR-enc trained additionally on Audio Set (0.5M non-speech and 0.6M typical speech utterances)



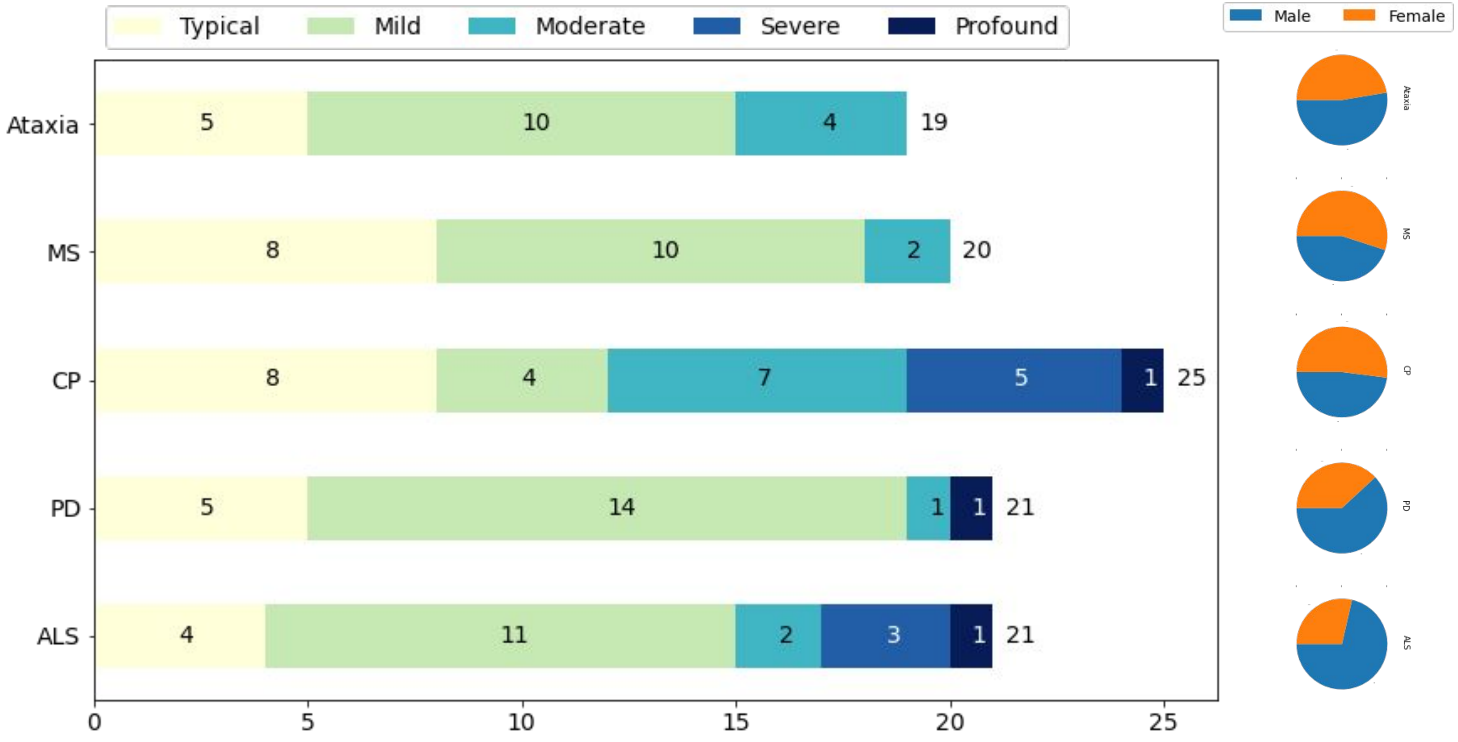
1 Search to filter videos based on relevant topics.

3 Further manual filtering. And SLPs tag/edit “regions of interest” (ROIs)

SLPs label

- ROI - time segments when dysarthric speaker is speaking
- severity and intelligibility - 5-point Likert
- inferred gender (to help balance)

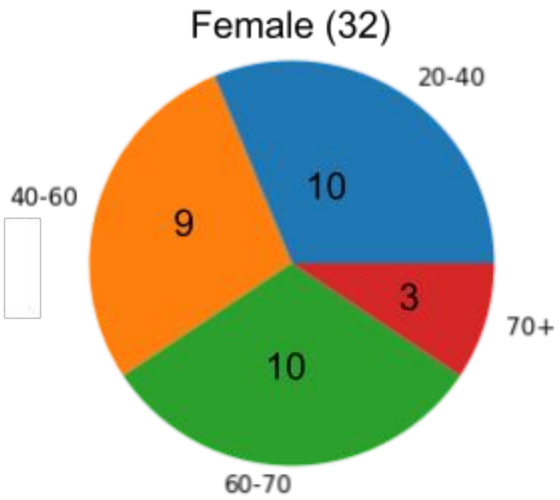
SpICE-V distribution

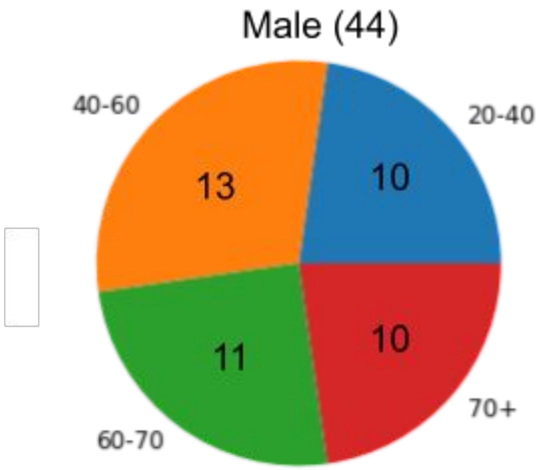


SpICE-V Controls: 76 speakers/videos

1. Select videos from AudioSet specifically the category tagged as “Speech”
2. We select from the unlabelled training set of 1M+ videos. Specifically only videos with tag
    a. Male speech, man speaking
    b. Female speech, woman speaking
    c. Optionally allowing for the tags “Narration, monologue” (and the tag speech)
    d. [detail] We looked at thumbnails of videos to determine - existence of video, confirmation of male/female speaker.
3. We watched the videos to infer age.
    a. We used the title and information tags in the video to look up speaker information as many of the speakers are somewhat public personalities e.g. sports persons, politicians featured heavily.
4. We tried to find as many videos of older people as we could.
    a. Intention to reduce bias of young adults and skew towards older age group and match gender.

SpICE-V Controls: 76 speakers/videos





Spice-V Results

Comparing accuracy of identifying atypical speech

Group	w. Typ. non-ctrl	# Utts.	Total (Atyp.) # Spkr	wav2vec 2.0 Acc. (%) spkr	wav2vec 2.0 Acc. (%) utt.	ASR-enc Acc. (%) spkr	ASR-enc Acc. (%) utt.
Controls	×	76	76 (0)	76.32	76.32	**96.42**	**96.42**
Dysarthric (-Typ.)	×	1489	76 (76)	**93.42**	**94.83**	63.16	66.92
Dysarthric (all)	✓	2221	106 (76)	**77.36**	**75.64**	68.65	67.92
All (-Typ.& Dys.)	×	1565	152 (76)	**84.87**	**93.93**	78.29	68.21
All	✓	2297	182 (76)	76.92	**75.66**	**78.57**	69.47

Sliced by Etiology

Etiology	# Utt.	# Spkr Total (Typ.)	wav2vec 2.0 Acc. (%) spkr	utt.	ASR-enc Acc. (%) spkr	utt.
ALS	443	21 (4)	90.5	87.6	76.2	76.0
PD	498	21 (5)	85.7	84.9	61.9	73.0
CP	620	25 (8)	72.0	69.8	72.0	74.5
MS	352	20 (8)	55.0	57.5	60.0	48.6
Ataxia	308	19 (5)	84.2	75.6	68.4	62.1

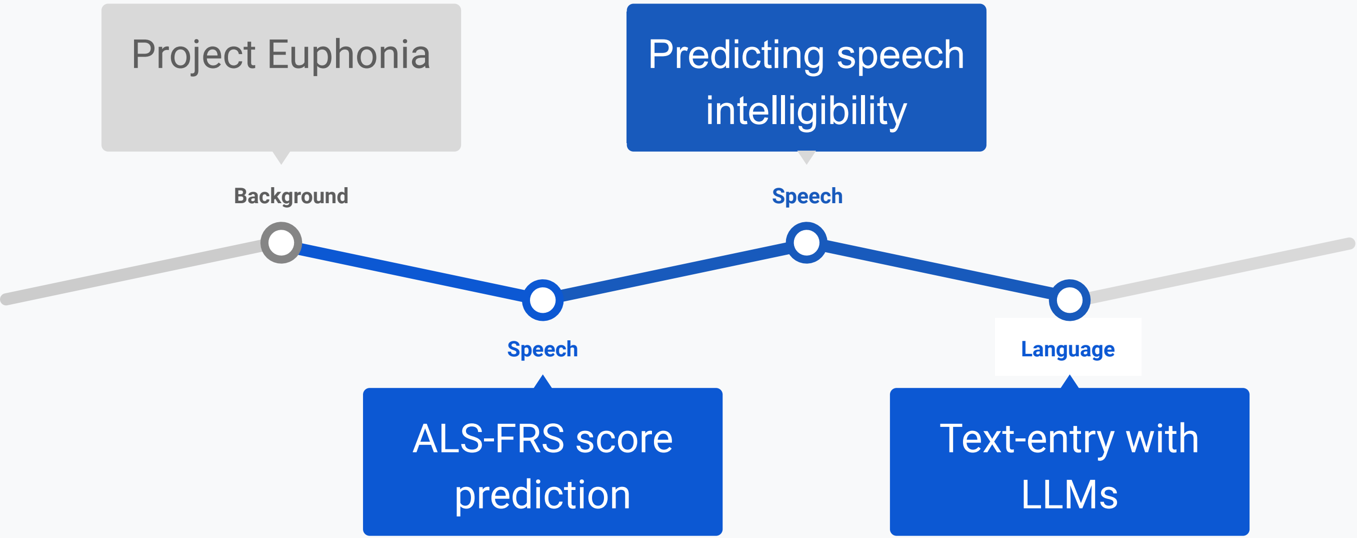
Findings

- Models do well on ALS, PD, CP and Ataxia.
- Dysarthric speakers with typical speech are harder to classify.
- No observed age or gender bias.
- Not good enough for clinical use - need accuracies in the high 90s.

Soon to release

- Open source version of the model

Outline


Project Euphonia
Background
Predicting speech intelligibility
Speech
Speech
ALS-FRS score prediction
Language
Text-entry with LLMs

Google Research

Context-Aware Abbreviation Expansion Using Large Language Models

S. Cai*, S. Venugopalan*, K. Tomanek, A. Narayanan, M. R. Morris, M. P. Brenner
NAACL'22

Motivation: AAC & Eye gaze typing

People who have difficulty communicating with speech use Augmentative and Alternative Communication (AAC) systems such as gaze based typing and speech synthesis to enter text and communicate.

E.g. people with conditions such as amyotrophic lateral sclerosis (ALS), multiple sclerosis (MS), and others.

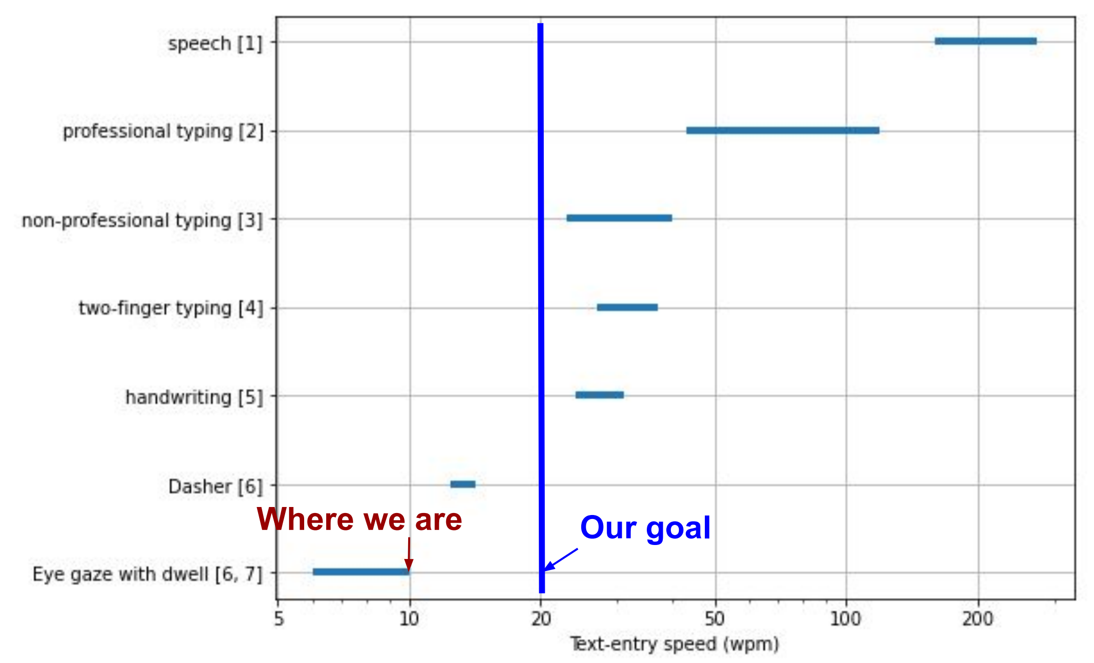
https://archinect.com/news/article/149956776/steve-saling-retired-landscape-architect-with-als-designs-residence-he-can-control-by-blinking

Goal: Faster typing

Gaze typing is very slow.

- Single-threaded
- There is only one point of gaze.
- Saccades and dwelling take time.

Can we find ways to save as many keystrokes as possible?

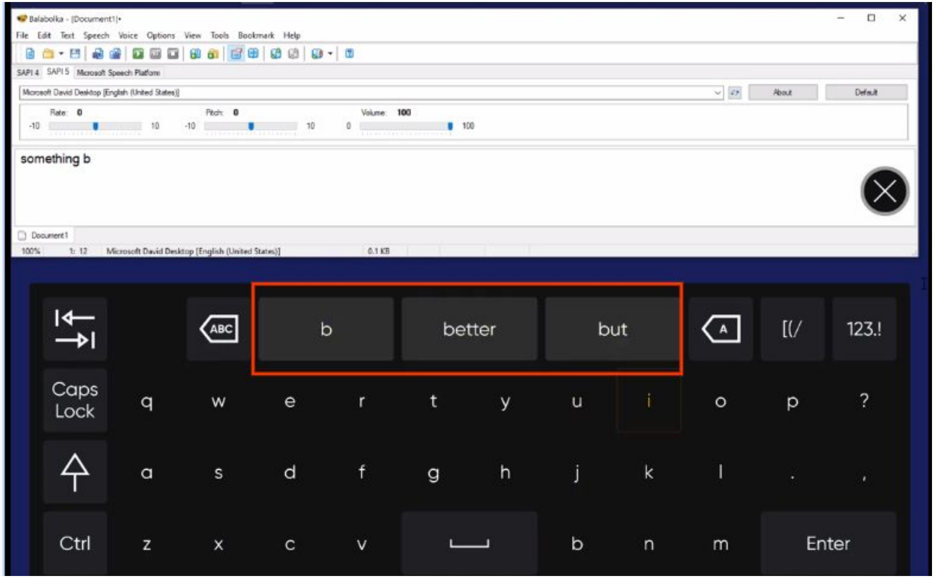


Google Research

n-gram predictions are a double-edged sword for AAC

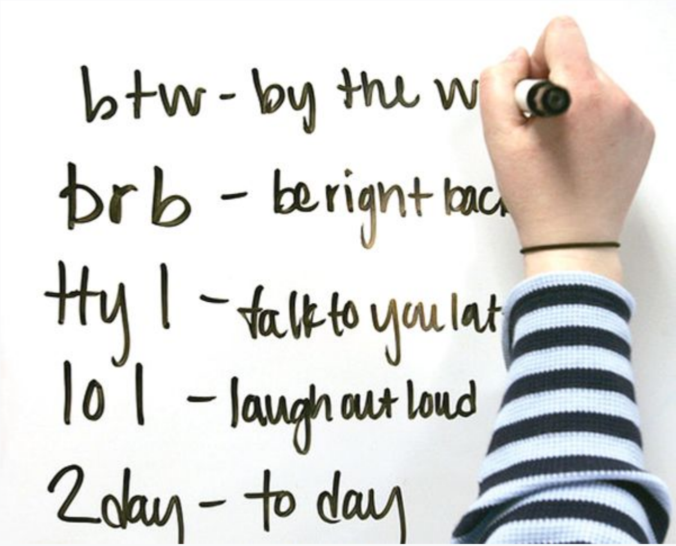
Repeated scanning of predictions is itself an overhead.

- The eyes play the dual roles of clicking keys and scanning predictions.
- A significant portion of the time involves no match, leading to wasted scanning time.
- ⇒ Can we devise a new paradigm of text entry to minimize scanning of predictions while achieving high Keystroke Savings?



Abbreviation Expansion (AE): Save as many keystrokes as possible.

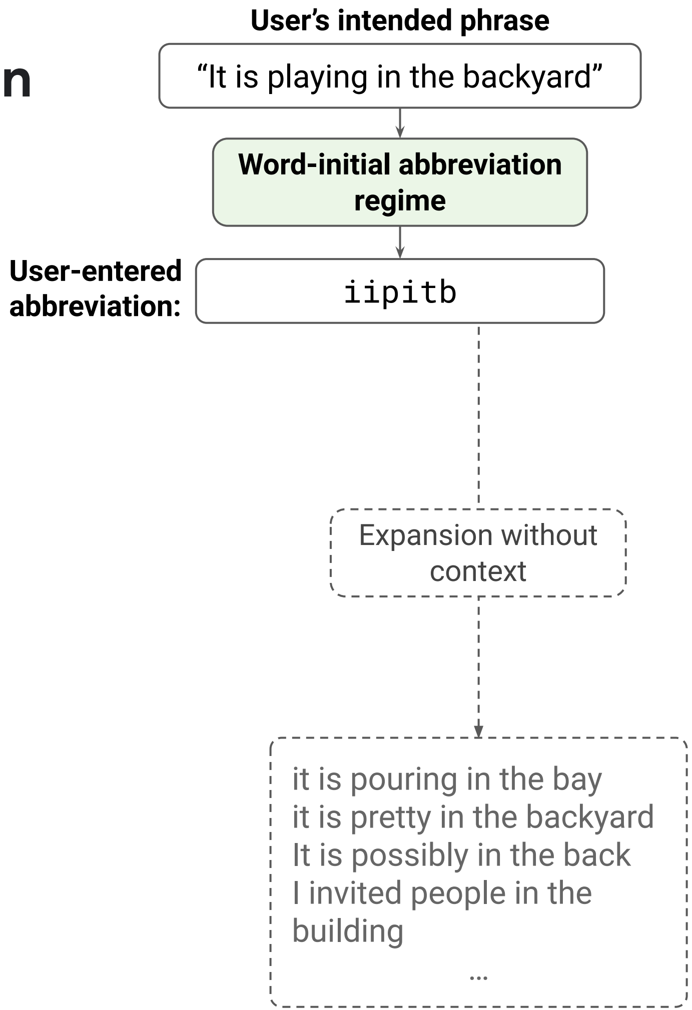
- Abbreviation expansion (AE): Partly inspired by “SMS language”, but extended to an open set
- Average word length in English is 4.7 + 1 = 5.7
- Theoretically 82% Keystroke Savings Rate (KSR) >> 40-50% from n-gram LM



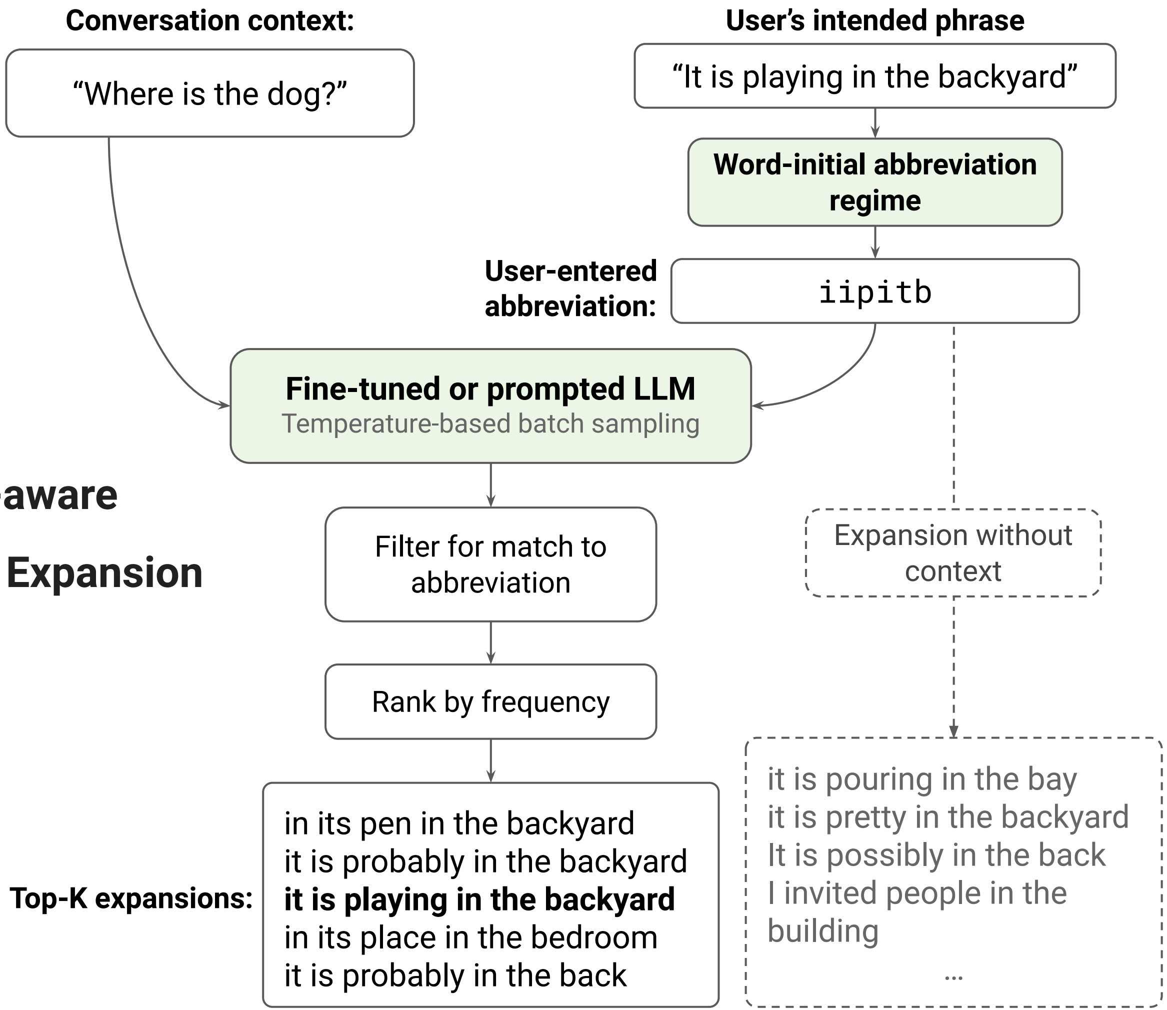
https://www.bms.co.in/is-text-sms-language-destroying-english-yes-or-no/

Task: Abbreviation Expansion

Expanding words from initial characters can be hard and ambiguous.



Context-aware Abbreviation Expansion



Abbreviation Expansion Prompt

Context:	{Content of the contextual turn}
Shorthand:	{Abbreviation of *next turn*}
Full:	{Expanded content of *next turn*}
Context:	{Would you like to sit down?}
Shorthand:	{n,imfsu}
Full:	{No, I'm fine standing up}

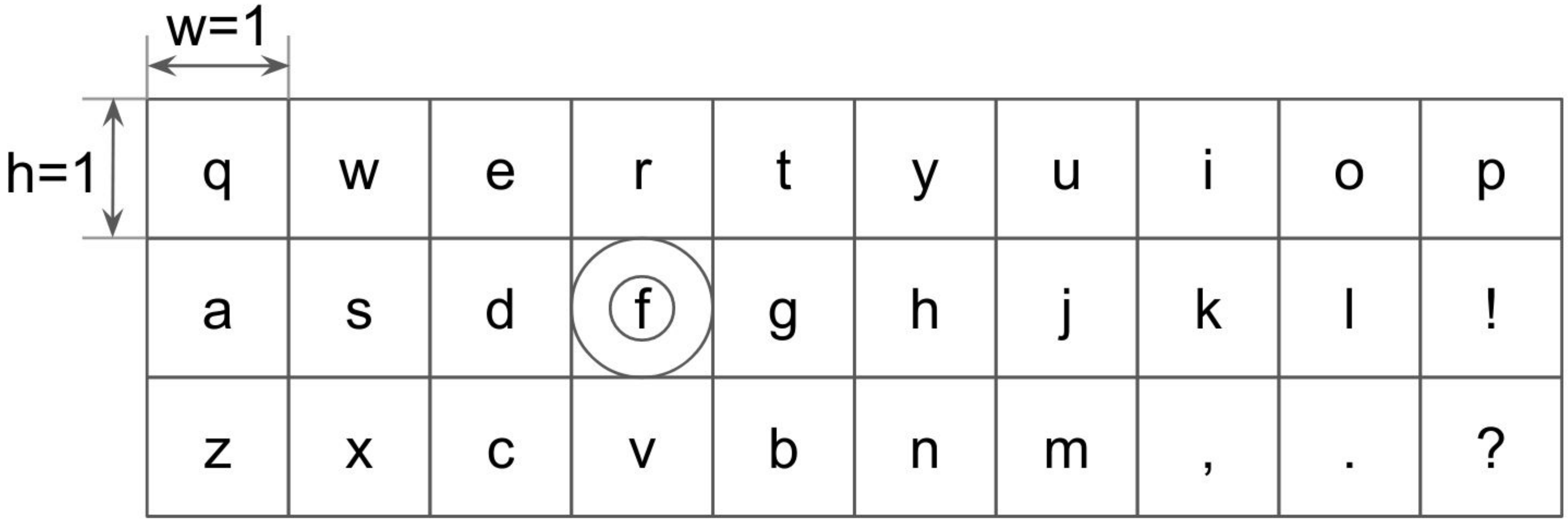
Generate abbreviation expansion examples from dialogs

	Original dialog	AE example
turn-1 turn-2 turn-3 turn-4 turn-5 turn-6	Would you like to sit down? No, I'm fine standing up Are you sure you don't want to sit down? Been sitting all day. Work was just one meeting after another. Oh, I'm sorry. I don't enjoy work days like that. It feels good to stretch my legs a bit.	**0-turn context:** Shorthand: {wyltsd}. Full: {Would you like to sit down?} **1-turn context:** Context: {Would you like to sit down?}. Shorthand: {n,imfsu}. Full: {No, I'm fine standing up} . . . **5-turn context:** Context: {Would you like to sit down?} {No, I'm fine standing up} {Are you sure you don't want to sit down?} {Been sitting all day. Work was just one meeting after another.} {Oh, I'm sorry. I don't enjoy work days like that.}. Shorthand: {ifgtsmlab}. Full: {It feels good to stretch my legs a bit.}

You can also imagine adding some typo noise

Original dialog	**AE example**	**AE example (noise σ=0.3)**
Would you like to sit down? No, I'm fine standing up Are you sure you don't want to sit down? Been sitting all day. Work was just one meeting after another. Oh, I'm sorry. I don't enjoy work days like that. It feels good to stretch my legs a bit.	**0-turn context:** Shorthand: {wyltsd}. Full: {Would you like to sit down?} **1-turn context:** Context: {Would you like to sit down?}. Shorthand: {n,imfsu}. Full: {No, I'm fine standing up} . . . **5-turn context:** Context: {Would you like to sit down?} {No, I'm fine standing up} {Are you sure you don't want to sit down?} {Been sitting all day. Work was just one meeting after another.} {Oh, I'm sorry. I don't enjoy work days like that.}. Shorthand: {ifgtsmlab}. Full: {It feels good to stretch my legs a bit.}	**0-turn context:** Shorthand: {wy!tsd}. Full: {Would you like to sit down?} **1-turn context:** Context: {Would you like to sit down?}. Shorthand: {n,infsu}. Full: {No, I'm fine standing up} . . . **5-turn context:** Context: {Would you like to sit down?} {No, I'm fine standing up} {Are you sure you don't want to sit down?} {Been sitting all day. Work was just one meeting after another.} {Oh, I'm sorry. I don't enjoy work days like that.}. Shorthand: {ifgtsmoab}. Full: {It feels good to stretch my legs a bit.}

Simulating gaze typo noise



Keyboard layout for simulating noise in AE key-presses. The circles on the f key show 1σ around the mean for $\sigma \in \{0.3, 0.5\}$ in the 2D Gaussian distributions used to model typing noise.

Train and evaluate on multiple datasets.

Select existing dialog datasets

- mostly everyday conversations
- one with dialogs from movies

Turk Dialogues* Corrected (TDC)

- 6 turns consistency
- Clean and diverse
- dev set - used for all param tuning.

* Vertanen K. Towards improving predictive aac using crowd sourced dialogues and partner context. SIGACCESS 2017

	train	dev.	test
Dataset	#conv.	#conv.	#conv.
Turk Dialogues Corrected (TDC)	859	280	280
Turk AAC (TAC)	5,019	559	565
DailyDialog Corrected (DDC)	11,188	823	772
Cornell Movie Dialog (CMD)	66,848	8,645	7,444
Task Master Self-Dialog (TMSD)			770

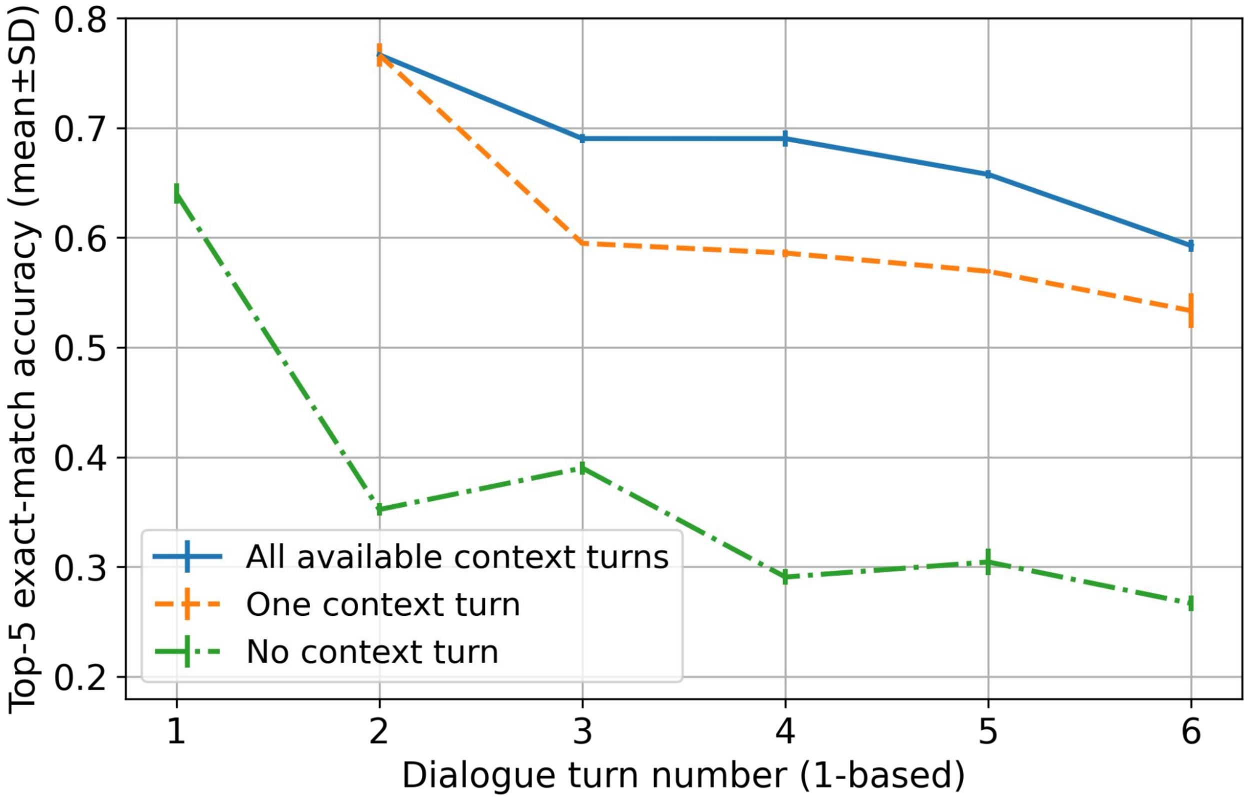
Evaluation metrics

- Accuracy (in top-5)
  - Generates/predicts the **exact** desired expansion.
- BLEU score
  - Partial credit. Looks for n-gram (1-, 2-, 3-, 4- word) matches and computes score.
- Keystrokes Savings Rate (KSR)
  - keystrokes saved compared to the full set of characters typed.
  - KSR-all → compute saved keystrokes if you have an **exact** match otherwise **penalize** for the user having to type the entire phrase over.
  - KSR-success → optimistic, only compute when you have a match.

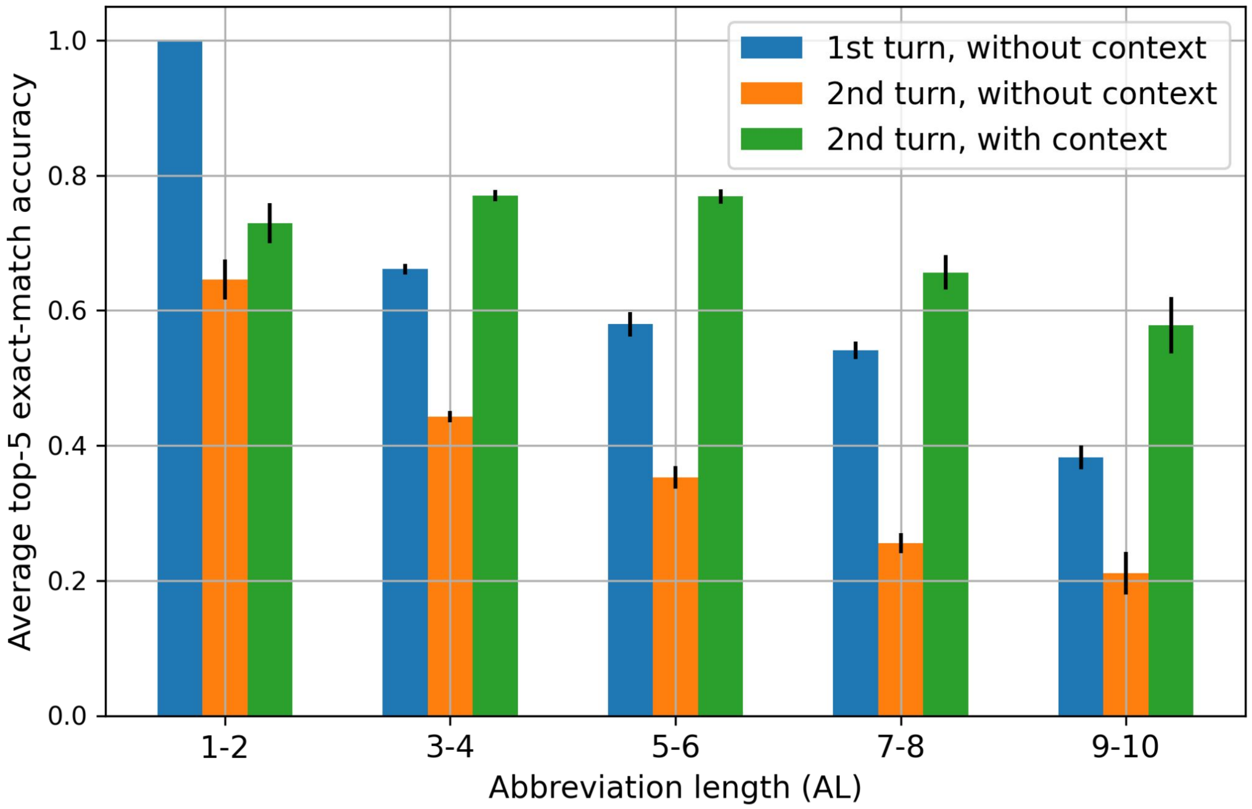
$$KSR_{all} = \begin{cases} \left(1 - \frac{L_{abbrev}}{L_{full}}\right) \times 100, & \text{if in top-5.} \\ \left(1 - \frac{L_{abbrev} + L_{full}}{L_{full}}\right) \times 100, & \text{otherwise.} \end{cases}$$

Google Research

Context reduces ambiguity and helps considerably.



Effect of context is more pronounced in longer sentences.



Error analysis: Examples of near misses

Google Research

#	Context	Abbreviation	Ground truth	Non-matching expansion options
1	Awesome! My favorite weather!	*swhottwp*	Shall we head over to the water park?	**shall we head out to the water park**
2	Can we go out for a drive?	*ygstc*	Yeah go start the car	**yes go start the car** yes go straight to church yes go settle the children yeah get some tunes cranked yes go straight to chicago
3	i took a lot of courses, such as philosophy, logic, ethics, aesthetics, etc	*wcdylb*	which course did you like best	what courses do you like best what courses did you like best what course do you like best what course did you like best **which courses did you like best**
4	it's hard to be optimistic about things with the way the economy's headed... the trade deficit is getting larger, consumption's down, i really think we're headed for a recession	*tehbsfawn*	the economy has been stagnant for a while now	the economy has been slowing for a while now **the economy has been sluggish for a while now** the economy has been strong for a while now the economy has been slow for a while now the economy has been suffering for a while now

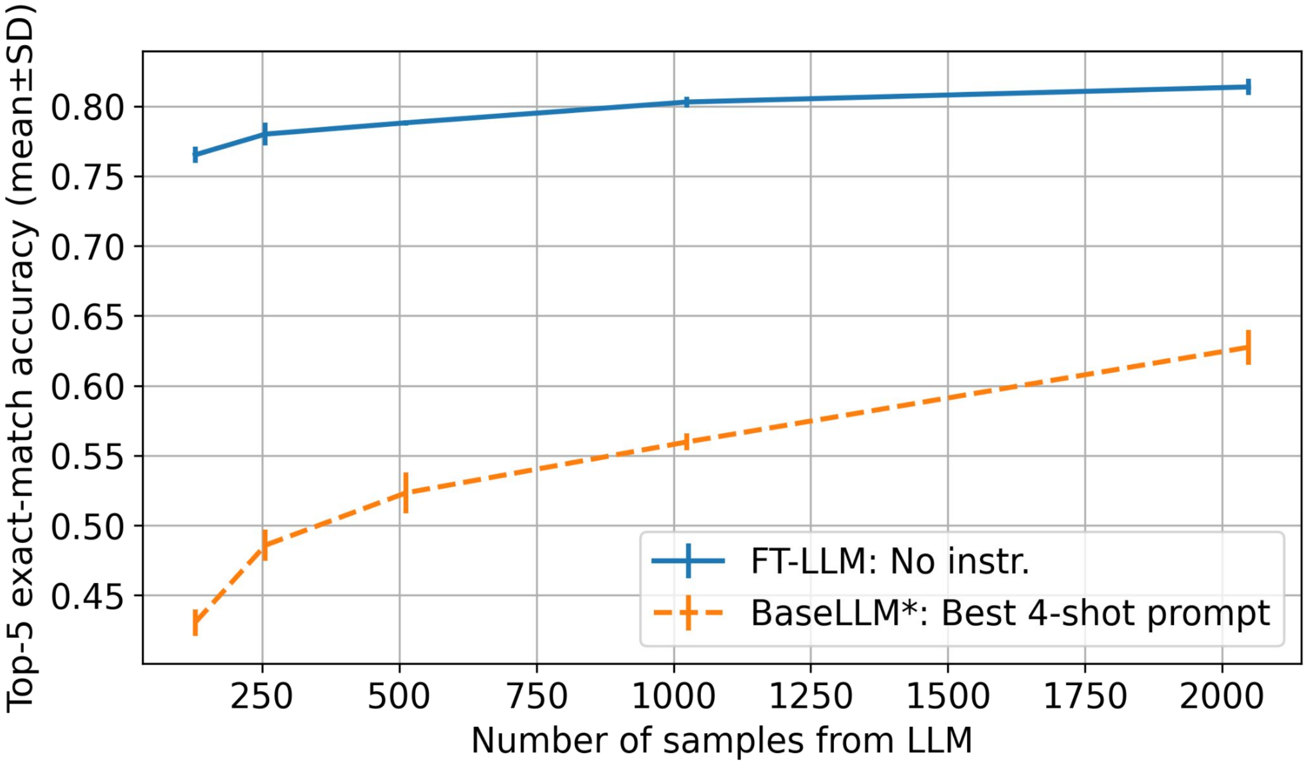
Error analysis: Fails to predict proper nouns

Google Research

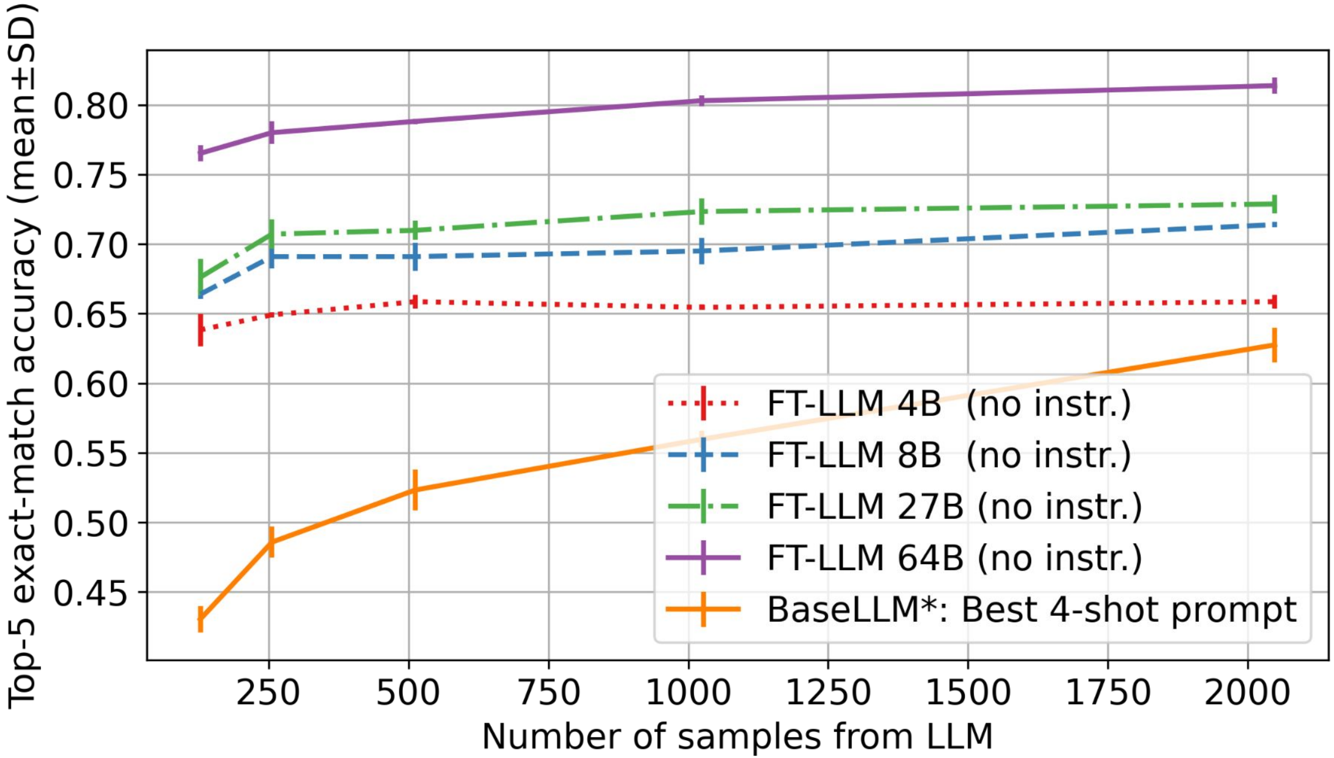
5	What is your name?	*mnir*	My name is Rey	my name is robert my name is rebecca my name is richard my name is rose my name is roy
6	hey, isabelle...	*l*	Logan	lisa linda look lillian liz
7	so, paula, where are you from	*imfc,o*	i'm from canada, originally	i'm from china, ok i'm from california, originally i'm from california, ok i'm from california, okay i'm from california, obviously

Google Research

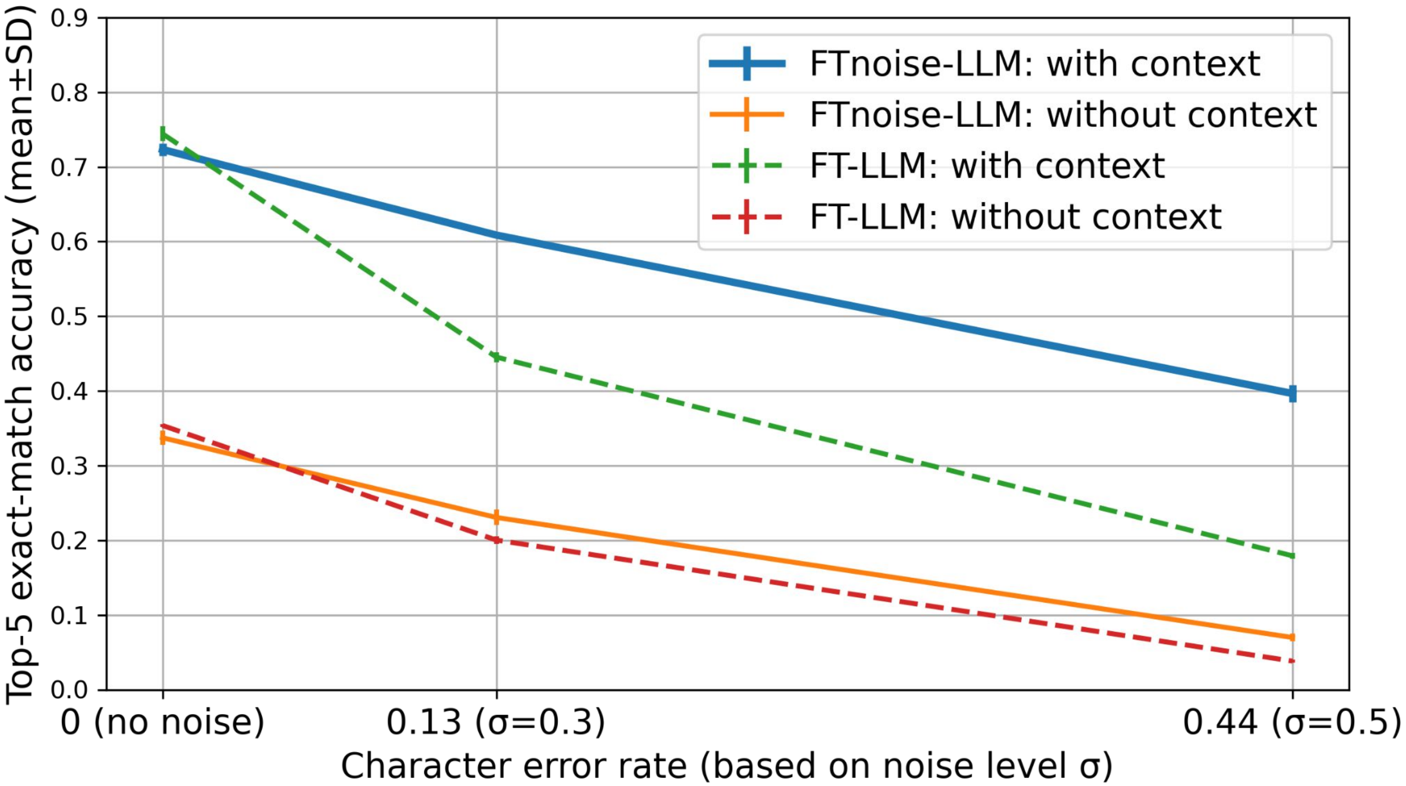
Fine tuning far outperforms few-shot even with low samples.



Size matters. Decode fewer samples from the largest model.

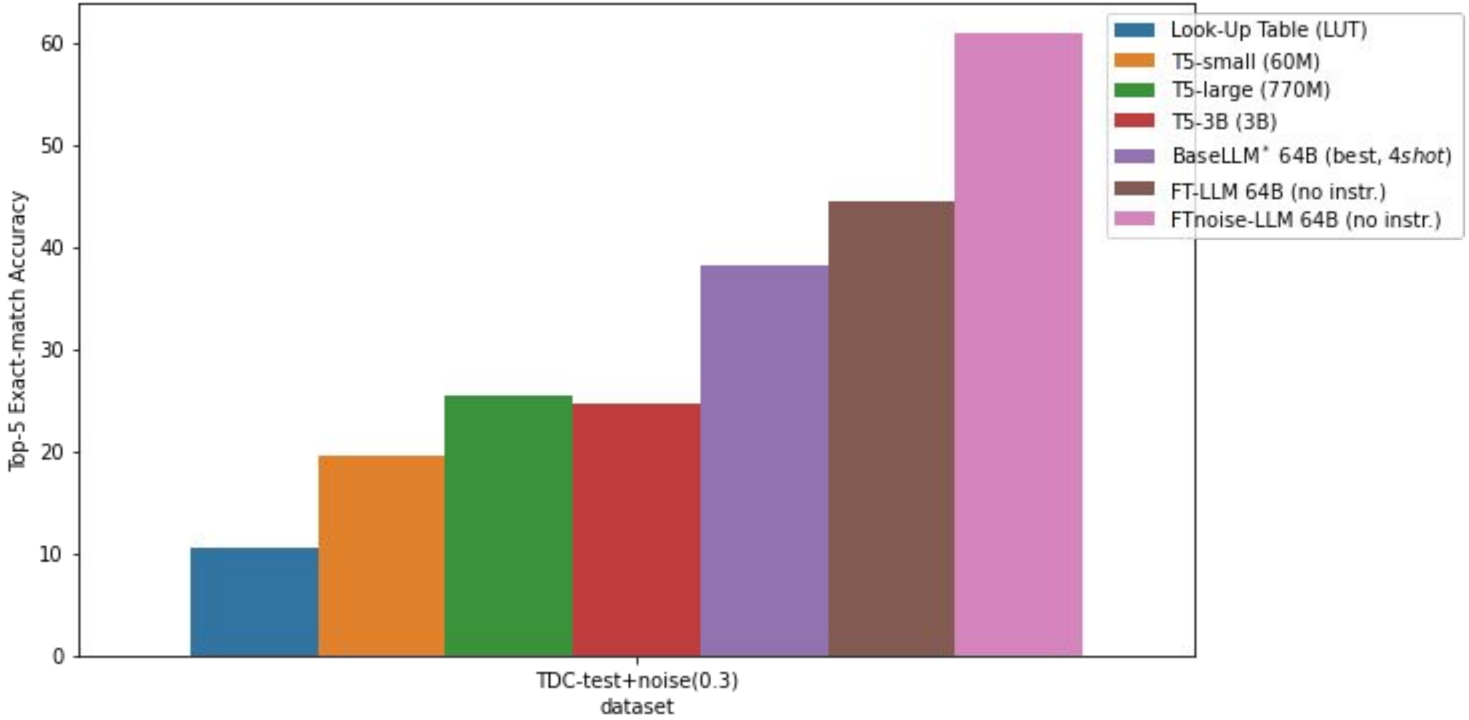


Context and fine-tuning with noise improves typo tolerance.



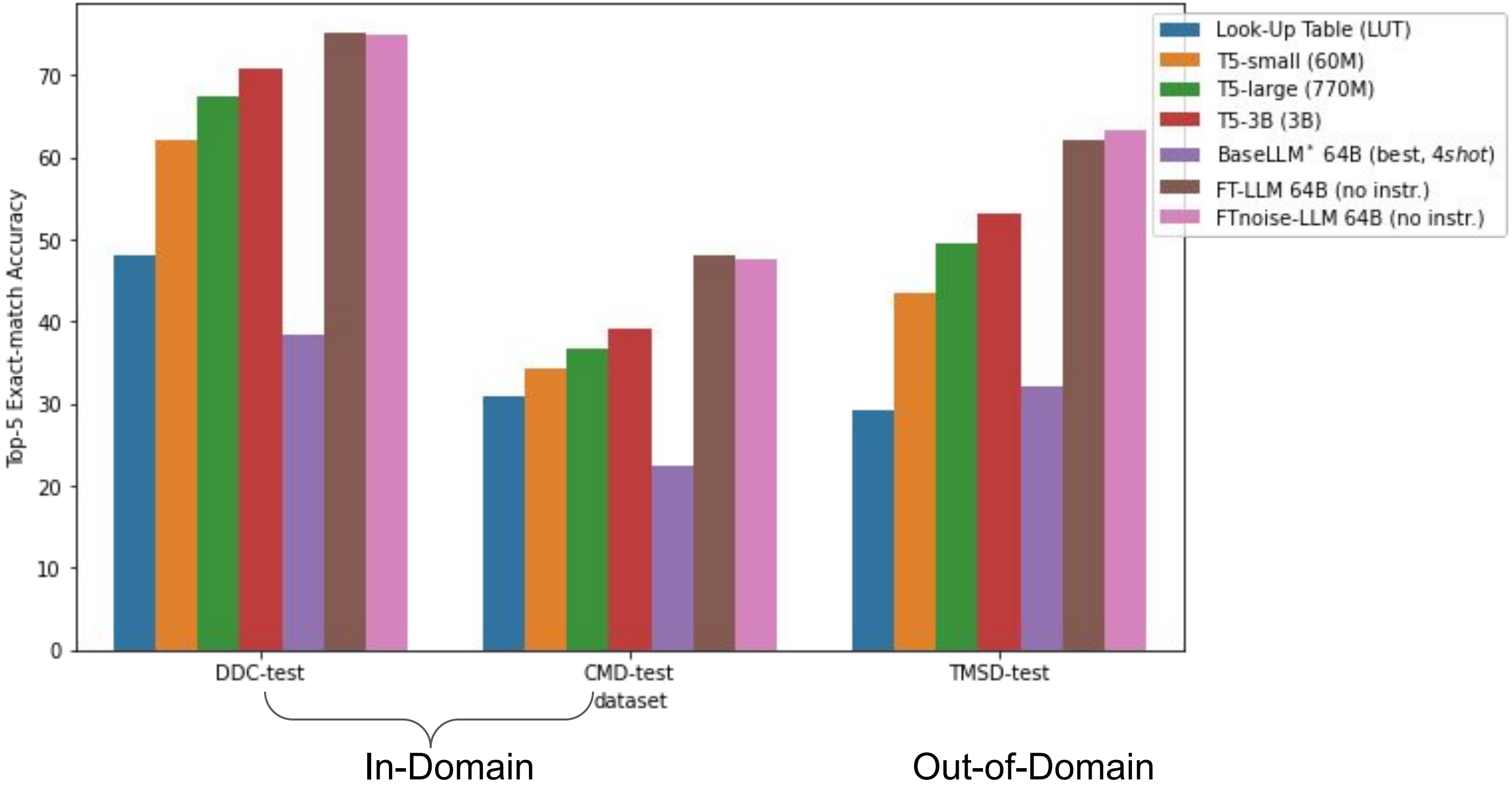
Google Research

Larger model is more tolerant to noise.



Google Research

Generalizes to different datasets and OOD.



LLMs show promise with huge keystrokes savings.

Dataset-split	AE task	KSR_{all}	$KSR_{success}$
TDC-test	1st turn (no context)	37.1 ± 0.19	76.8 ± 0.04
	2st turn (with context)	49.0 ± 0.99	73.5 ± 0.03
DDC-test	1st turn (no context)	20.0 ± 1.15	74.6 ± 0.04
	2st turn (with context)	49.0 ± 0.60	72.9 ± 0.04

In most previous typing scenarios (e.g. n-gram completions on mobile phone keyboards) theoretical keystrokes savings is close to 50%, and effective savings on studies turns out to be 20%-30%. Here, improving accuracy can result in huge keystrokes savings.

So, LLMs show promise in enabling a much harder regime.

Google Research

Summary

- We aim to speed up eye-gaze AAC text entry speed 2x by using ML.
- We are using LLM to perform context-dependent abbreviation expansion (AE)
- Under certain testing conditions, context-dependent AE can show up to 76% keystrokes savings, but needs to be validated through real-world user testing.
    - Plan to measure through user study - Will the overall system result in close to 2x speed-up?