Google Research

# Assessing ASR Model Quality on Disordered Speech using BERTScore

Jimmy Tobin, Qisheng Li, Subhashini Venugopalan, Katie Seaver, Richard Jonathan Noel Cave, Katrin Tomanek

http://arxiv.org/abs/2209.10591

Speech for Social Good 2022

# Outline

01 Overview of Project Euphonia

02 Personalized ASR

03 Measuring model performance

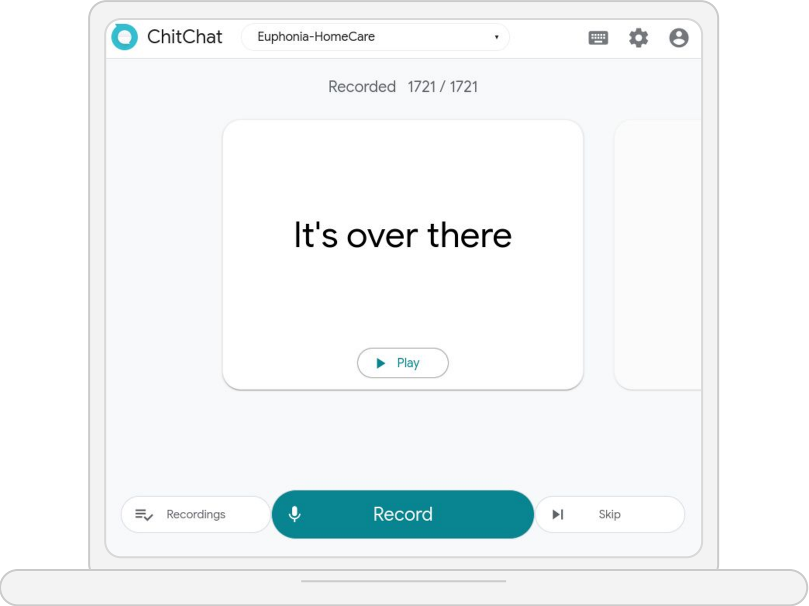
**250M+** people worldwide who have non-standard speech can’t rely on speech technology today.

...So our research over the past several years has been centered around recording as much audio data as possible from participants who have non-standard speech.

# Research and data collection began in 2018

- Research and Speech teams within Google began to work on this problem in 2018

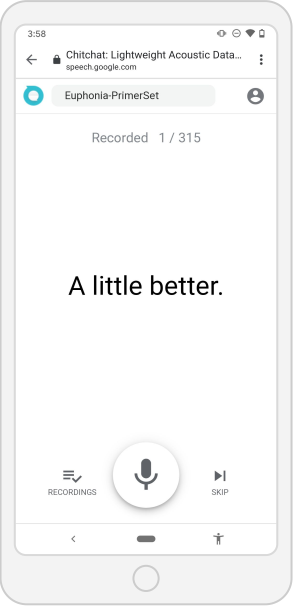
Computer



Recorded 1721 / 1721

It's over there

Mobile



Recorded 1 / 315

A little better.

# Each phrase we ask participants to record corresponds to a key use case

## 1 Interacting with technology

**Home Automation**

*Stop the music.*
*Call Mom mobile.*
*Turn the bedroom lights on.*
*Close to front door.*

**Voice Access**

*A, B, C, D, E...*
*1, 2, 3, 4, 5....*
*Up, down...*

## 2 Interacting with others

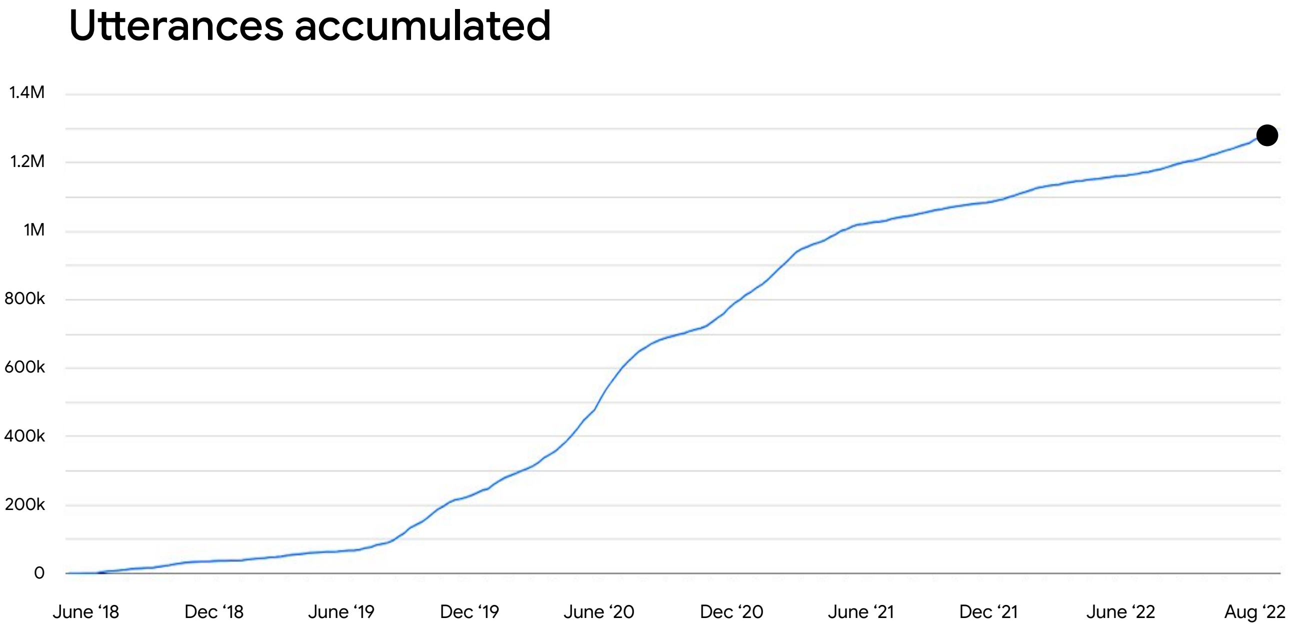
**Talking to a caregiver**

*I need to move.*
*Can you turn the TV on?*
*I want to go to bed.*
*I’m hungry.*

**Conversations**

*That food will never go bad.*
*I like reading books more than watching TV.*
*He told his mother a long story.*

# To date, we have over 1 million recordings from over 1,000 people



**>1.3m**
utterances

**>1800**
hours of audio

**>2000**
speakers w/ recordings

Google Research

ALS-TDI

Canadian Down Syndrome Society

CureDuchenne

LSVT Global

MND Association

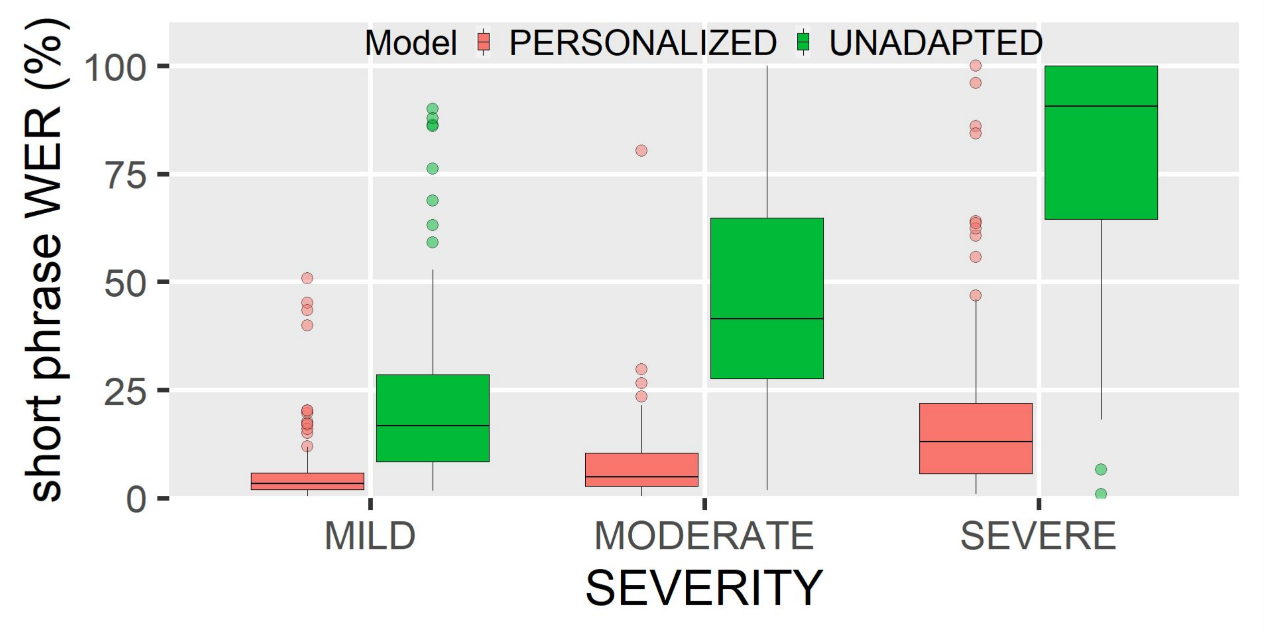
Team Gleason

# Personalized ASR Models

Personalized models reduced WER (word error rate) by over 75%

Users used personalized models for:

- Home automation
- Face to face conversation
- Dictation and emails
- Transactional interactions (asking for things)

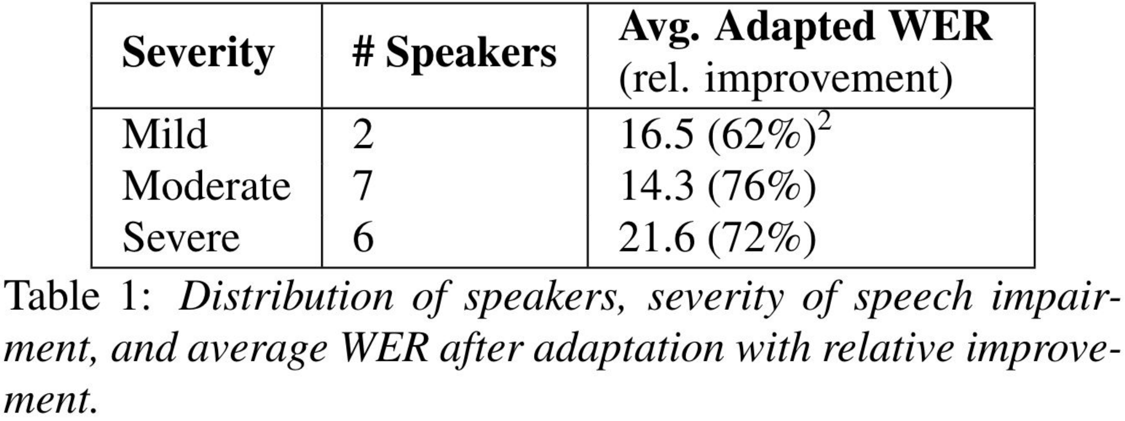


# Focusing on 15 speakers

Etiologies include:

- ALS
- Cerebral palsy
- Down Syndrome
- Multiple sclerosis

WER* is an order of magnitude higher than models evaluated on typical speech.



| Severity | # Speakers | Avg. Adapted WER (rel. improvement) |
|---|---|---|
| Mild | 2 | 16.5 (62%)[2] |
| Moderate | 7 | 14.3 (76%) |
| Severe | 6 | 21.6 (72%) |

Table 1: *Distribution of speakers, severity of speech impairment, and average WER after adaptation with relative improvement.*

***WER = (Deletions + Insertions + Substitutions) / Total tokens**
**Word Accuracy = 1 - WER**

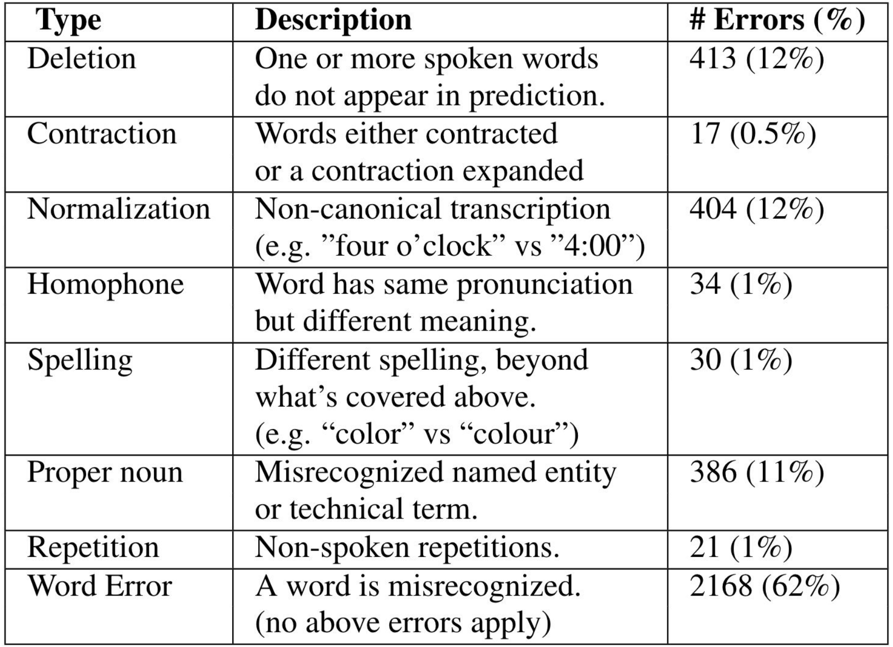
# Error Analysis

2 Speech Language Pathologists labeled 3473 model transcriptions errors for:

- Error Type
- Error Severity Assessment

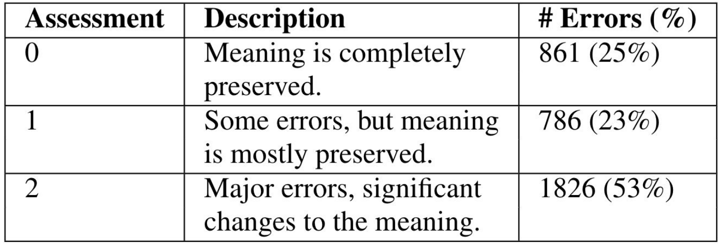
“Substantial” inter-annotator agreement as measured by Cohen’s kappa:

- Error Type $\kappa$=0.64
- Error Severity $\kappa$=0.69



| Type | Description | # Errors (%) |
|---|---|---|
| Deletion | One or more spoken words do not appear in prediction. | 413 (12%) |
| Contraction | Words either contracted or a contraction expanded | 17 (0.5%) |
| Normalization | Non-canonical transcription (e.g. ”four o’clock” vs ”4:00”) | 404 (12%) |
| Homophone | Word has same pronunciation but different meaning. | 34 (1%) |
| Spelling | Different spelling, beyond what’s covered above. (e.g. “color” vs “colour”) | 30 (1%) |
| Proper noun | Misrecognized named entity or technical term. | 386 (11%) |
| Repetition | Non-spoken repetitions. | 21 (1%) |
| Word Error | A word is misrecognized. (no above errors apply) | 2168 (62%) |

Table 2: *Description of error types with counts and proportion of 3473 errors.*



| Assessment | Description | # Errors (%) |
|---|---|---|
| 0 | Meaning is completely preserved. | 861 (25%) |
| 1 | Some errors, but meaning is mostly preserved. | 786 (23%) |
| 2 | Major errors, significant changes to the meaning. | 1826 (53%) |

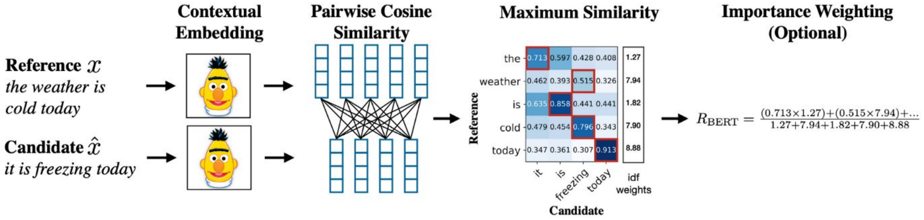
Table 3: *Error severity assessment response scale, descriptions, counts and proportion of total 3473 errors.*

# BERTScore

**BERT** is a contextual embedding model.

Comparing two sentences' token embeddings, **BERTScore** is the maximization of cosine similarity.

We report **F**$_{\text{BERT}}$, which is the F1 measure combining precision and recall from BERTScore.



| Reference | it | is | freezing | today | idf weights |
|---|---|---|---|---|---|
| the | 0.713 | 0.597 | 0.428 | 0.408 | 1.27 |
| weather | 0.462 | 0.393 | 0.515 | 0.326 | 7.94 |
| is | 0.635 | 0.858 | 0.441 | 0.441 | 1.82 |
| cold | 0.479 | 0.454 | 0.796 | 0.343 | 7.90 |
| today | 0.347 | 0.361 | 0.307 | 0.913 | 8.88 |

$$R_{\text{BERT}} = \frac{(0.713\times1.27)+(0.515\times7.94)+\ldots}{1.27+7.94+1.82+7.90+8.88}$$

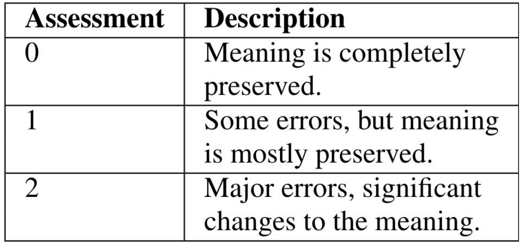
Source: Bertscore: Evaluating text generation with bert

Code for Bertscore is available at https://github.com/Tiiiger/bert score
Image from Bertscore paper https://arxiv.org/abs/1904.09675

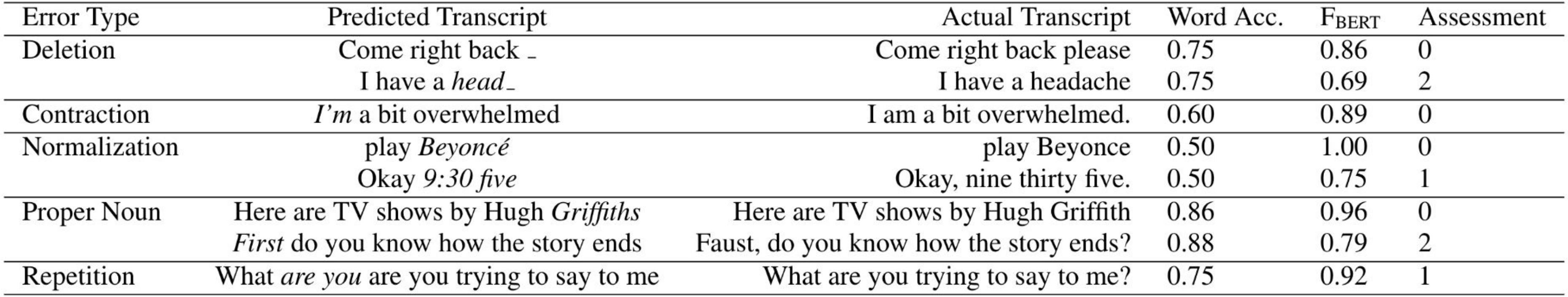
$$R = \frac{1}{|x|}\sum_{x_i \in x} \max_{y_j \in y}(x_i \cdot y_j) \quad P = \frac{1}{|y|}\sum_{y_j \in y} \max_{x_i \in x}(x_i \cdot y_j)$$

$$F_{BERT} = 2\frac{P \times R}{P + R} \tag{1}$$

# Comparing metrics



| Assessment | Description |
|---|---|
| 0 | Meaning is completely preserved. |
| 1 | Some errors, but meaning is mostly preserved. |
| 2 | Major errors, significant changes to the meaning. |



| Error Type | Predicted Transcript | Actual Transcript | Word Acc. | $F_{BERT}$ | Assessment |
|---|---|---|---|---|---|
| Deletion | Come right back _ | Come right back please | 0.75 | 0.86 | 0 |
| | I have a *head*_ | I have a headache | 0.75 | 0.69 | 2 |
| Contraction | *I'm* a bit overwhelmed | I am a bit overwhelmed. | 0.60 | 0.89 | 0 |
| Normalization | play *Beyoncé* | play Beyonce | 0.50 | 1.00 | 0 |
| | Okay *9:30 five* | Okay, nine thirty five. | 0.50 | 0.75 | 1 |
| Proper Noun | Here are TV shows by Hugh *Griffiths* | Here are TV shows by Hugh Griffith | 0.86 | 0.96 | 0 |
| | *First* do you know how the story ends | Faust, do you know how the story ends? | 0.88 | 0.79 | 2 |
| Repetition | What *are you* are you trying to say to me | What are you trying to say to me? | 0.75 | 0.92 | 1 |

Table 4: *Examples of errors with associated Word Accuracy,* $F_{BERT}$ *and Error Assessment metrics.*

# BERTScore distinguishes Error Severity better

| Metric | BERTScore | Word Accuracy |
|---|---|---|
| Std Dev | 0.142 | 0.274 |
| One-way ANOVA | F=684*** | F=209*** |

***(p<0.001)

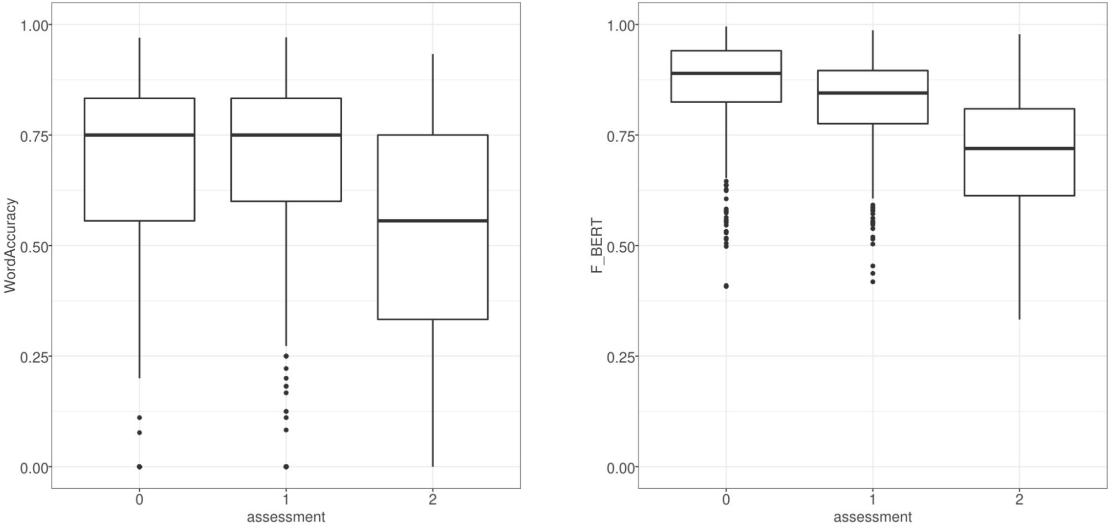


Figure 1: *Distribution of Word Accuracy (left) and* $F_{BERT}$ *(right) broken out by error assessment.*

Google Research

# BERTScore distinguishes Error Type better

BERTScore is more robust to normalization and contraction errors that do not change semantic meaning.

| Metric | BERTScore | Word Accuracy |
|---|---|---|
| One-way ANOVA | F=41.8*** | F=9.45*** |

***(p<0.001)

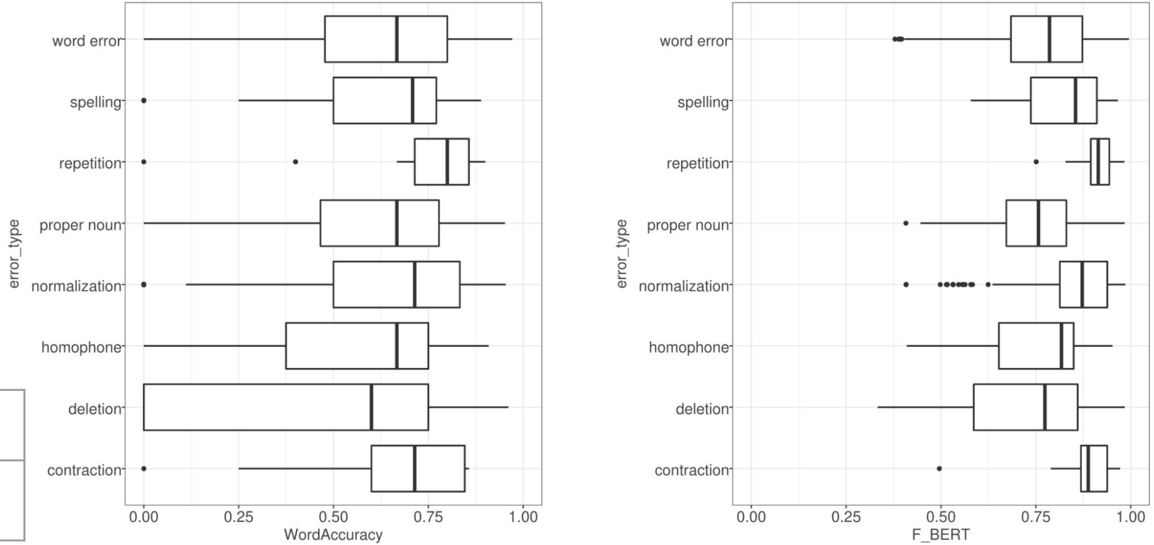


Figure 2: *Distribution of Word Accuracy (left) and $F_{BERT}$ (right) broken out by error type.*

# BERTScore fits SLP error severity assessments better

Though both are Word Accuracy and $F_{BERT}$ are significant predictors or error severity assessment, $F_{BERT}$ is more predictive. This is evidenced by higher magnitude of coefficient and lower Akaike Information Criterion (AIC).

The Akaike information criterion (AIC) is an estimator of prediction error and thereby relative quality of statistical models for a given set of data.

## Ordinal Logistic Regression Analysis

| Metric | Coeff | Std Err | t value | p value | AIC |
|---|---|---|---|---|---|
| Word accuracy | -2.52 | 0.144 | -17.4 | $p<0.001$ | 6733 |
| $F_{BERT}$ | **-10.87** | 0.380 | -28.6 | $p<0.001$ | **5854** |

# Conclusions

- When creating ASR models for individuals if atypical speech, conveying semantic meaning is the most important metric.
- Both metrics are significantly correlated with Error Type and Assessment, but BERTScore is a stronger predictor of Error Assessment.
- BERTScore can be used in conjunction with WER to measure ASR models for speakers with disordered speech.

# Thank you so much to our many participants and testers!!

# Thank you.

**More info**

g.co/Euphonia