Large Language Models as a Proxy For Human Evaluation in Assessing the Comprehensibility of Disordered Speech Transcription

Katrin Tomanek[1], Jimmy Tobin[1], Subhashini Venugopalan[1], Richard J.N. Cave[1], Katie Seaver[1,2], Rus Heywood[1], Jordan Green[1,2], [1] *Google Research,* [2] *MGH Institute of Health Professions,*
{katrintomanek, jtobin}@google.com

Introduction

WER treats all errors the same

- Word Error Rate (WER) is a measure of the syntactic accuracy of an automatic speech recognition (ASR) model. It is not meant to measure **comprehensibility**.
- When working with low resource languages like disordered speech, WER is often >20 and sometimes >60 for certain etiologies and severities.
- Individuals with disordered speech may still find benefit from an ASR model with relatively high WER, so we aim to create a system that will automatically assess the ability of an ASR model to convey the user's intended message.

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

Dataset and Method

Transcript Comprehensibility Data Set

- Created a dataset mapping transcription errors to assessments from Speech-Language Pathologists (SLPs) of meaning preservation.
- Significant inter-annotator agreement when assessing meaning preservation, Cohen's κ = 0.7

Error Severity	Meaning Preserved	Description	# Examples (%)	Example
0	yes	Meaning is completely preserved	900 (19%)	**G:** I would be fascinated to know your answers. **T:** I *will* be fascinated to know your answers.
1	yes	Some errors, but meaning is mostly preserved.	1145 (24%)	**G:** Yeah I have one basically every day. **T:** Yeah I have *I'm* basically every day.
2	no	Major errors, significant loss of intended meaning.	2686 (57%)	**G:** How large is that file? **T:** How large is a *funnel*?

Table 1. Error severity assessment response scale, descriptions, counts and proportion of the total 4,731 erroneous transcripts with representative examples (**G** for ground truth and **T** for transcript).

Severity	Mild	Moderate	Severe
# Speakers	13	15	14

Table 2. Distribution of speakers and impairment severity.

Split	train	test	dev
# Examples	2840	940	951
Meaning Preserved	43.6%	40%	45.2%

Table 3. Data splits and percentage of transcripts categorized as *meaning preserved*.

LATTEScore

LLMS to Assess TranscripTion Errors Score

$$\text{LATTEScore} = \frac{\text{\# Predicted Meaning Preserved}}{\text{\# Total Examples}} \times 100 \quad (1)$$

LLM Prompt Tuning to Predict Transcript Comprehensibility

Example 1
Input Sequence
```
Ground truth: {no no there are fifteen hundred total}.
Transcription: {no no there are 50 energy total}.
Transcript preserves the meaning of the ground truth: {
```
Target Sequence
```
no}
```

Example 2
Input Sequence
```
Ground truth: {He's huggable and lovable and a good with people.}.
Transcription: {He's huggable and laughable and a good with people}.
Transcript preserves the meaning of the ground truth: {
```
Target Sequence
```
yes}
```

Figure 1. Representation of task for LLM-based classifiers.

Results

Comparison with other approaches

- As models get bigger, the AUC is increases.
- The LLM approach ranks higher than sentence embeddings of the same model size.

Approach	approx. # params	full test set (940)	speech type: prompted (391)	speech type: unprompted (549)	severity: severe (467)	severity: moderate (302)	severity: mild (149)
BERTScore + WER	350M	0.791	0.788	0.794	0.753	0.791	0.856
SentT5 Embedding Sim	11B	0.857	0.894	0.831	0.813	0.879	0.899
FLAN-T5 XXL	11B	0.878	0.903	0.860	0.836	0.923	0.890
LLM62B	62B	**0.900**	**0.918**	**0.886**	**0.863**	**0.944**	**0.903**

Table 4. AUC-ROC scores on full and subsets of the test set for the different approaches to predict meaning preservation of erroneous transcripts (numbers in brackets represent # examples in specific subset).

Using LATTEScore to Assess Personalized ASR Models

- Personalized ASR models were evaluated on real conversational test phrases.
- The Word Accuracy metric overstated how useful some models were.
- LATTEScore more closely matched ground truth annotations

Speaker	Etiology	Severity	Utterances	Word Accuracy	True Meaning Preservation Percentage	LATTEScore
S1	MS	Moderate	72	59.2	48.6	47.2
S2	Cleft Palate	Severe	94	60.0	35.1	34.0
S3	ALS	Severe	152	60.9	48.7	31.6
S4	PLS	Mild	61	66.7	44.3	55.7
S5	ALS	Severe	262	71.9	46.9	42.7
S6	VCP	Severe	50	72.6	**74.0**	64.0
S7	ALS	Moderate	179	**80.0**	52.0	52.0
S8	ALS	Moderate	76	**80.5**	57.9	55.3
S9	VCP	Severe	41	**86.5**	68.3	**70.7**
S10	DS	Mild	44	**90.8**	**77.3**	**77.3**

Table 5. Word Accuracy, Percentage Meaning Change (based on SLP assessment) and LATTEScore on real conversation test. Bolded numbers show which models would have been accepted based on our deployment decision thresholds (Word Accuracy >= 80 and Meaning Preserved >= 70).

Conclusion

Takeaways

- We define a new approach to measuring ASR model performance using LLMs.
- LATTEScore measures comprehensibility rather than syntax preservation in ASR model performance.
- When compared to WER, LATTEScore can better assess how useful an ASR model will be to the end user.