1st Workshop on Speech for Social Good (S4SG)
24-25 September 2022, Incheon, Korea

Assessing ASR Model Quality on Disordered Speech using BERTScore

Jimmy Tobin[1], Qisheng Li[2], Subhashini Venugopalan[1], Katie Seaver[3], Richard Cave[4], Katrin Tomanek[1]

[1]Google Research, USA;
[2]University of Washington, USA;
[3]Leonard Florence Center for Living, USA;
[4]MND Association, UK

{jtobin,katrintomanek}@google.com

Abstract

Word Error Rate (WER) is the primary metric used to assess automatic speech recognition (ASR) model quality. It has been shown that ASR models tend to have much higher WER on speakers with speech impairments than typical English speakers. It is hard to determine if models can be be useful at such high error rates. This study investigates the use of BERTScore, an evaluation metric for text generation, to provide a more informative measure of ASR model quality and usefulness. Both BERTScore and WER were compared to prediction errors manually annotated by Speech Language Pathologists for error type and assessment. BERTScore was found to be more correlated with human assessment of error type and assessment. BERTScore was specifically more robust to orthographic changes (contraction and normalization errors) where meaning was preserved. Furthermore, BERTScore was a better fit of error assessment than WER, as measured using an ordinal logistic regression and the Akaike's Information Criterion (AIC). Overall, our findings suggest that BERTScore can complement WER when assessing ASR model performance from a practical perspective, especially for accessibility applications where models are useful even at lower accuracy than for typical speech.

1. Introduction

Automatic Speech Recognition (ASR) model quality is most often measured by Word Error Rate (WER), the aggregate score of word deletions, substitutions and insertions. WER is the *de facto* metric for benchmarking models' improvements and regressions. Most state-of-the-art (SOTA) ASR models trained and evaluated on typical English speech report single digit WERs on standard test corpora [1]. However, models trained only on typical speech perform worse on disordered speech, reporting a median WER of 31.5 on short phrases [2]. In these situations where WER is high, it is difficult to assess how useful a model will be for the end user. In particular, WER doesn't capture semantic closeness or the assessment of the error. For instance, misrecognizing the suffix 's' in a plural and missing the negation of a word can have the same error rate but missing the negation can fail to capture the user's intent (see examples in Table 4). Thus, looking at word error rate alone may not convey the utility of the model in some practical settings where errors are tolerated. An example being video call captioning for a person with a speech impairment, which can ensure that the conversation partner will understand the speaker better.

Green et al. show that, for individuals with disordered speech, personalizing ASR models with the individual's speech can lead to on average 75% relative WER improvements over unadapted models trained for typical speech [2]. The personalized models, although having WERs in the range of 15 or 20 (which would be considered high for typical speech), provided utility to the users in certain domains such as home automation, spoken transactions (asking for something), and conversations [3]. Their work demonstrates that transcriptions do not need to be perfect to communicate a speaker's intent or meaning. Errors such as normalization (one hundred percent vs. 100%) or contraction expansion (I'm vs I am) can be considered a combination of deletions, insertions and substitutions, but to a human or a SOTA natural language understanding (NLU) system, the meaning is conveyed the same.

Alternatives to WER have been investigated, aiming to address the limitations when applied to conversations [4, 5]. There are several metrics in natural language understanding applications that aim to measure similarity of generated text with a human truth e.g. BLEU [6] and METEOR [7] are popular in machine translation, ROUGE [8] is used the context of summarization, SPICE [9] is used in image-captioning, and more recently BLEURT [10] and BERTScore [11] which are based on pre-trained language models have been proposed for text generation. Even for speech applications, recent work [12, 13] proposed metrics utilizing pre-trained language models. In Kim et al. [12], the proposed alternate metric is based on semantic distance between the predicted and ground truth transcripts. Our work is motivated similarly, though we utilize BERTScore[1], an existing semantic distance based metric popular in language generation and adapt it to our task.

In this work, we specifically investigate BERTScore and WER for measuring the utility of personalized ASR models. We worked with Speech Language Pathologists (SLPs) to perform a detailed analysis of the errors, including error type and assessment assessments, based on the model predictions from personalized ASR models. We then compared both BERTScore and WER in terms of how well they aligned with the SLP assessments. Our findings showed that while both BERTScore and WER are well correlated with SLP judgements, BERTScore was a better fit to the assessments overall. Further, BERTScore was a better indicator of utility in the presence of simple contraction and normalization errors than WER.

2. Data and Approach

We first describe the model and data used in our analysis and then describe the metric and comparison approach.

[1]BERTScore is available open source at https://github.com/Tiiiger/bert_score.

10.21437/S4SG.2022-6

2.1. Personalized ASR Models and utility

The ASR models used in our study are based on work from Green et. al. [2] which developed personalized models for 432 speakers with impairments who contributed to Project Euphonia [14]. This corpus consists of over 1 million samples (over 1300 hours) of more than 1000 anonymized speakers with different types and severity levels of speech impairments. Following the same approach in [2], we fine-tuned models on a subset of 15 individuals from the Euphonia corpus [14] by tuning most of the initial encoder layers while keeping the decoder frozen. This included speakers having varying degrees of speech impairment described in Table 1. Speaker etiologies include Amyotrophic Lateral Sclerosis (ALS), Cerebral Palsy, Down Syndrome, Multiple Sclerosis, and other long tail etiologies that have less representation in the overall Euphonia dataset, including Brain Injury, Vocal Chord Paralysis and Cleft Palate. This approach showed huge improvements in WER, from average WER of 66.3 $\pm$ 10.0 before adaptation to a WERs in the range of 15-20 after personalization. These WERs though, are still much higher than the original model's WER of 6-8 on typical speech [15].

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.*

To validate the utility of the models beyond the WER improvements, particularly for overall user experience, we solicited feedback from 15 speakers who used personalized ASR models to dictate their speech through a mobile app for several months. The feedback suggested that in several scenarios, the models were indeed useful. Specifically, users were able to (1) perform home automation through voice controlled assistants, (2) use it for face to face and video conversations with friends and family, (3) long-form dictation and email, and (4) even for short transactional interactions with individuals unfamiliar with the impaired speaker's voice. These user success stories suggested that despite high WERs, users experienced benefits across several domains and situations.

2.2. Error Analysis

To understand and measure the utility of the models and track improvements and regressions, we sought more detailed analysis of the errors from SLPs on the per-speaker test set produced by each speaker's personalized model. SLPs assessed a total of 3473 transcription errors and determined (1) error type and (2) assigned an assessment rating of the error severity. They categorized the errors into 8 types including deletion, contraction, normalization, homophone, spelling, proper noun, repetition, and other word errors. These are described in Table 2. The error severity assessment was rated on a scale of 0, 1 or 2 depending on whether or not an error preserved meaning. This is described in Table 3. An error severity assessed at levels 0 and 1 are considered to be "recoverable," i.e., a communication partner would be able to understand what a speaker meant to say. An error level 2 represents a communication breakdown, where the perceived meaning is lost or drastically altered from a speaker's intent.

[2]Mild speech severity models having a higher average WER than moderate speech severity is an artifact of the small sample size. Green et al. (2021) showed that *Mild* WER was significantly lower than *Moderate* in the Euphonia dataset.

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.*

Dataset. We consolidated the errors and assessments into a dataset. Table 1 shows the distribution of the speakers and the severity of their speech impairment, Table 2 shows distribution over error types and Table 3 includes the assessment at different severity levels. As shown in Table 3, 53% the errors were assessed as Level 2 errors, 25% of the transcription errors were marked as Level 0 (meaning completely preserved) and another 23% as Level 1 (minor errors): in almost half of the cases, the meaning was preserved. The majority of errors were the most general category *Word Error*. There were, however, a significant amount of normalization errors, which are typically meaning preserving (Levels 0 or 1). Examples of different error types and the associated SLP assessments are shown in Table 4.

2.2.1. *Overview of transcript error assessment task*

For completeness, we also detail the specific instructions provided to the SLPs here. SLPs were given the following instructions when assessing an error. To prevent bias and better simulate being a conversational partner, SLPs were asked to look at ground truth transcripts only after making a mental guess at the phrase meaning.

1. Read the model prediction without looking at the original transcript.
2. Make a mental guess at the intended meaning of the text.

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.*

3. Read the ground-truth transcript and compare to the prediction; assign an error assessment according to the scale in table.

To assess Inter-Annotator Agreement, measured with Cohen's kappa [16], 5% of the errors were annotated by two SLPs. Raters were in substantial agreement for both error assessment ($\kappa = 0.64$) and error types ($\kappa = 0.69$) based on the scale in [17].

2.3. Word Accuracy and F_{BERT}

BERTScore [11] is a text evaluation metric based on a pre-trained BERT model's [18] contextual embedding. It measures the similarity of two sentences as the sum of the cosine similarities of the token embeddings of the sentences as follows. Let the ground truth transcript (x) with pre-normalized token embeddings (x_i), be represented as $x = \langle x_1, \cdots, x_k \rangle$, and a similarly tokenized predicted sentence be represented as $y = \langle y_1, \cdots, y_l \rangle$, and the cosine similarity between two tokens x_i and y_j be denoted as $x_i \cdot y_j$. We use the version of BertScore where the cosine similarity is weighted by the inverse document frequency (idf) computed on the test set of ground truth transcripts. Then, BERTScore maximizes the cosine similarity of each token in x with each token in y to compute the recall, precision, and F1 measure as:

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

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

We use the BERT-base model [18] for the token embeddings to compute cosine similarity. BERTScore[11] was evaluated on several hundred machine translation and captioning systems and was shown to be robust to syntax changes and provide a metric of semantic similarity that aligned well with human judgements. In this analysis we use the F1 measure, denoted F_{BERT} (Eqn. 1), combining the sentence level precision and recall returned from BERTScore. The ranges of F_{BERT} and WER[3] are [0,1] and [0,100] respectively. A transcript with an exact match would have F_{BERT} = 1 and WER = 0. In order to make analysis more comparable, we match ranges by reporting Word Accuracy (Eqn. 2) instead of WER in the subsequent sections.

$$\text{Word Accuracy} = 1 - \frac{WER}{100} \tag{2}$$

2.4. Ordinal Logistic Regression

We examine how well WER and F_{BERT} as metrics are aligned with expert-annotated error assessments. Since error assessments were graded on a Likert Scale from 0 to 2 and are inherently ordinal, we ran a series of ordinal logistic regression (OLR) models with error assessments as the dependent variable, and Word Accuracy and/or F_{BERT} as the independent variables.

To further determine model quality with Word Accuracy and F_{BERT} as parameters, we rank each models' associated Akaike Information Criterion (AIC)[19], a lower AIC indicating a better model for the data.

[3]WER is capped at 100 in this analysis.

3. Results

BERTScore distinguishes error severity better. Figure 1 shows the summary statistics of Word Accuracy and F_{BERT} for the three levels of error assessment. We can clearly see that 1) F_{BERT} distinguishes different error assessment levels better than Word Accuracy, especially between 0 and 1; and 2) F_{BERT} has overall smaller variance ($SD = 0.142$) than Word Accuracy ($SD = 0.274$). A one-way ANOVA test also confirms that F_{BERT} ($F = 684.38, p < 0.001$) could differentiate different assessment levels better than Word Accuracy ($F = 209.01, p < 0.001$).

BERTScore distinguishes error types better. Leveraging the expert-annotated error types for all data, we then compare how Word Accuracy and F_{BERT} differentiate between different error types, with results shown in Figure 2. For instance, F_{BERT} differentiates contraction errors better than Word Accuracy (i.e. they should have higher scores because these errors typically do not impact the semantic meanings of the speech), similarly for normalization errors. One-way ANOVA tests again confirm that F_{BERT} ($F = 41.75, p < 0.001$) outperforms Word Accuracy ($F = 9.45, p < 0.001$) when distinguishing different error types on impaired speech.

BERTScore fits SLP assessments better. Our results of ordinal logistic regressions show that both Word Accuracy and F_{BERT} are significant predictors for error assessment, but F_{BERT} has a larger absolute value of coefficient (coeff $= -2.52, t = -17.4, p < 0.001$) than Word Accuracy (coeff $= -10.87, t = -28.6, p < 0.001$), which means F_{BERT} is more predictive of error assessment: a decrease in F_{BERT} is more likely to result in a higher error assessment score (i.e. more severe error). This is also confirmed by AIC scores where $AIC = 5854$ for OLR with F_{BERT} as the parameter and $AIC = 6733$ with Word Accuracy. The OLR model with F_{BERT} as the parameter fits better with the error assessment data, than with Word Accuracy as the parameter.

BERTScore is more useful when measuring semantic similarity. Models based on WER and BERTScore both significantly fit the data, but there are situations where BERTScore is more robust to transcription errors that do not affect the semantic meanings. Table 4 shows transcription error examples with their associated Word Accuracy and F_{BERT}. For instance, both "Deletion" examples (see row 1 in Table 4) have a word

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

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

accuracy of 0.75. However, one (*Come right back _*) has an assessment score of 0 because the original meaning of the utterance is persevered, while the other example (*I have a head_*) has an assessment score of 2 because of the significant change to the intended meaning. Using Word Accuracy in this case fails to distinguish the quality of the predicted transcripts. F_{BERT}, instead, performs better in measuring the quality of the predictions (0.86 vs. 0.69). Similarly, normalization errors often result in low Word Accuracy, especially for short phrases, even though the meaning is not affected. E.g. *play Beyoncé* vs. *play Beyonce* has an Word Accuracy of 0.50 while an F_{BERT} of 1.00.

BERTScore is however more expensive and less flexible than WER. WER is useful in many applications regardless of language or phrase domain. BERTScore on the other hand is more computationally expensive and relies on a pre-trained BERT model that can run into issues with multilingual data and out of vocabulary words.

4. Conclusion

ASR models exhibit higher word error rates when measured on non-typical speech, relative to rates typically reported in SOTA research. When making product decisions, it can be difficult to assess the quality of a model with a large WER. Models that could provide utility to people with impaired speech may be gated from release unnecessarily. To our knowledge, this is the largest user study validating the usefulness of personalized models despite models having WERs larger than products made for typical speakers.

In this study we discussed a separate metric for determining model quality based on loss of semantic meaning in transcription errors using BERTScore, a semantic embedding metric. We found that BERTScore was more robust to several types of transcription errors including normalization and contraction errors.

In future studies, we aim to investigate further metrics, such as Character Error Rate (CER) and BLEURT, that can give insight into when a model will be useful for speakers with disordered speech, especially when a speaker may have a condition that causes their voice to become more severely impaired over time. WER and BERTScore, when used together as parameters, fit the data better than either parameter alone. Future studies should also look at utilizing both signals together to assess model usefulness. It would also be good to a domain specific threshold of WER and F_{BERT} that can help predict success rate, proportion of speakers that will have a usable ASR model after personalization, similar to the home automation domain WER of 15 discussed in Tobin and Tomanek(2022).

Future studies can also investigate applications of BERTScore for estimating ASR model quality beyond disordered speech. ASR models with higher WER may still be useful for speakers with accented speech or non-ideal acoustic settings for example.

5. References

[1] Synnaeve, "WER are we?" github.com/syhw/wer_are_we, 2022, accessed: 2022-04-09. [Online]. Available: https://github.com/syhw/wer_are_we

[2] J. R. Green, R. L. MacDonald, P.-P. Jiang, J. Cattiau, R. Heywood, R. Cave, K. Seaver, M. A. Ladewig, J. Tobin, M. P. Brenner, P. C. Nelson, and K. Tomanek, "Automatic Speech Recognition of Disordered Speech: Personalized Models Outperforming Human Listeners on Short Phrases," in *Proc. Interspeech 2021*, 2021, pp. 4778–4782.

[3] J. Tobin and K. Tomanek, "Personalized automatic speech recognition trained on small disordered speech datasets," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 6637–6641.

[4] Y.-Y. Wang, A. Acero, and C. Chelba, "Is word error rate a good indicator for spoken language understanding accuracy," in *2003 IEEE Workshop on Automatic Speech Recognition and Understanding (IEEE Cat. No.03EX721)*, 2003, pp. 577–582.

[5] A. C. Morris, V. Maier, and P. Green, "From WER and RIL to MER and WIL: improved evaluation measures for connected speech recognition," in *Proc. Interspeech 2004*, 2004, pp. 2765–2768.

[6] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: A method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ser. ACL '02. USA: Association for Computational Linguistics, 2002, p. 311–318. [Online]. Available: https://doi.org/10.3115/1073083.1073135

[7] S. Banerjee and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," in *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*. Ann Arbor, Michigan: Association for Computational Linguistics, Jun. 2005, pp. 65–72. [Online]. Available: https://aclanthology.org/W05-0909

[8] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, Jul. 2004, pp. 74–81. [Online]. Available: https://aclanthology.org/W04-1013

[9] P. Anderson, B. Fernando, M. Johnson, and S. Gould, "Spice: Semantic propositional image caption evaluation," in *Computer Vision – ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham: Springer International Publishing, 2016, pp. 382–398.

[10] T. Sellam, D. Das, and A. Parikh, "BLEURT: Learning robust metrics for text generation," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, Jul. 2020, pp. 7881–7892. [Online]. Available: https://aclanthology.org/2020.acl-main.704

[11] T. Zhang*, V. Kishore*, F. Wu*, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," in *International Conference on Learning Representations*, 2020.

[12] S. Kim, A. Arora, D. Le, C.-F. Yeh, C. Fuegen, O. Kalinli, and M. L. Seltzer, "Semantic Distance: A New Metric for ASR Performance Analysis Towards Spoken Language Understanding," in *Proc. Interspeech 2021*, 2021, pp. 1977–1981.

[13] R. Sharma, S. Palaskar, A. W. Black, and F. Metze, "End-to-end speech summarization using restricted self-attention," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 8072–8076.

[14] R. L. MacDonald, P.-P. Jiang, J. Cattiau, R. Heywood, R. Cave, K. Seaver, M. A. Ladewig, J. Tobin, M. P. Brenner, P. C. Nelson, J. R. Green, and K. Tomanek, "Disordered Speech Data Collection: Lessons Learned at 1 Million Utterances from Project Euphonia," in *Proc. Interspeech 2021*, 2021, pp. 4833–4837.

[15] R. Prabhavalkar, T. N. Sainath, Y. Wu, P. Nguyen, Z. Chen, C.-C. Chiu, and A. Kannan, "Minimum word error rate training for attention-based sequence-to-sequence models," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE Press, 2018, p. 4839–4843. [Online]. Available: https://doi.org/10.1109/ICASSP.2018.8461809

[16] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, 1960.

[17] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, vol. 33, no. 1, pp. 159–174, 1977. [Online]. Available: http://www.jstor.org/stable/2529310

[18] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[19] H. Akaike, "Information theory and an extension of the maximum likelihood principle," in *Selected papers of hirotugu akaike*. Springer, 1998, pp. 199–213.