A Paradigm for Interpreting Metrics and Measuring Error Severity in Automatic Speech Recognition

Thibault Bañeras-Roux, Mickael Rouvier, Jane Wottawa, Richard Dufour

A Paradigm for Interpreting Metrics and Measuring Error Severity in Automatic Speech Recognition

Thibault Bañeras Roux[1], Mickael Rouvier[2,3], Jane Wottawa[3], and Richard Dufour[4]

[1] Nantes University, LS2N, France
[2] Le Mans University, LIUM, France
[3] Avignon University, LIA, France
{thibault.roux, richard.dufour}@univ-nantes.fr
mickael.rouvier@univ-avignon.fr
jane.wottawa@univ-lemans.fr

Abstract. The evaluation of automatic speech transcriptions relies heavily on metrics such as Word Error Rate (WER) and Character Error Rate (CER). However, these metrics have faced criticism for their limited correlation with human perception and their inability to capture linguistic and semantic nuances accurately. Despite the introduction of metric-based embeddings to approximate human perception, their interpretability remains challenging compared to traditional metrics. In this article, we introduce a novel paradigm aimed at addressing these limitations. Our approach integrates a chosen metric to derive Minimum Edit Distance (minED), which serves as an indicator of the rate of serious errors in automatic speech transcriptions. Unlike conventional metrics, minED offers a more nuanced understanding of errors, accounting for both linguistic complexities and human perception. Furthermore, our paradigm facilitates the measurement of error severity from both intrinsic and extrinsic perspectives.

Keywords: automatic speech recognition · evaluation metrics · semantice evaluation · human perception.

1 Introduction

Despite significant advancements in speech processing and the widespread use of data in training, Automatic Speech Recognition (ASR) systems continue to exhibit transcription errors across various usage conditions.

Traditionally, the evaluation of ASR systems involves comparing manual (reference) and automatic (hypothesis) transcriptions using metrics such as Word Error Rate (WER) and Character Error Rate (CER). However, these metrics have been criticized for their inability to capture semantic nuances effectively [4,16,7,5], as they assign equal weight to all errors.

In response to these limitations, embedding-based metrics [18,9,1] have been proposed to incorporate semantic aspects into the evaluation process. Likewise, from a perceptive point-of-view, the speech community [7,10,5,2] used annotated data sets to rigorously evaluate the alignment of speech recognition metrics with human perception, revealing the superior correlation of semantic metrics with human judgment. While semantic metrics offer a different evaluation perspective, their scores, computed through cosine similarity, can be challenging to interpret compared to traditional metrics like WER.

In this article, we propose a novel paradigm called Minimum Edit Distance (minED) to address the interpretability issue of evaluation metrics in ASR systems. Unlike traditional metrics, minED calculates a serious error rate according to a chosen metric, making it more interpretable and reflective of error severity. Furthermore, we introduce minED as a tool for measuring error severity, offering insights into the performance of ASR systems from both intrinsic and extrinsic perspectives. Our code is openly available on a public GitHub repository[4].

The paper is organized as follows. Section 2 introduces ASR metrics and a data set with human perception annotations. Section 3 outlines the proposed minED paradigm for metric interpretability while Section 4 examines the paradigm's ability to measure error severity. We finally conclude the work and give perspectives in Section 5.

2 Methodology Overview

In Section 2.1, we provide details on the ASR metrics utilized in this study. Then in Section 2.2, we present the HATS data set, employed for evaluating both the metrics and our paradigm.

2.1 Metrics

In response to criticisms of metrics like WER and CER, the research community has introduced a variety of alternative evaluation measures. Leveraging techniques from BERT [3], semantic representations known as embeddings can be extracted from sentences. One notable metric, SemDist [9], quantifies the cosine similarity distance between reference and hypothesis embeddings at the sentence level. Another metric, BERTScore [18], widely applied across Natural Language Processing (NLP) tasks [17,6], computes a similarity score for each token in the candidate sentence against each token in the reference sentence using contextual embeddings. These two embedding-based metrics are examined. SemDist incorporates the Sentence-BERT [15] version of CamemBERT[5] [12], a French adaptation of BERT. Additionally, BERTScore utilizes a multilingual BERT model [3]. To ensure consistent comparison and interpretation across metrics, all values were normalized to a [0, 1] scale, adhering to a lower-is-better principle.

[4] https://github.com/thibault-roux/mined
[5] https://huggingface.co/dangvantuan/sentence-camembert-large

2.2 HATS Dataset

The openly accessible HATS dataset[6] [2] serves as the resource for evaluating the correlation between ASR evaluation metrics and human perception. The construction of the HATS dataset involved a side-by-side experiment [5,7,10]. In this experiment, a textual reference alongside two erroneous hypotheses generated by ASR systems (comprising 8 end-to-end systems [14] and 2 DNN-HMM-based systems[7] [13]) was presented to a minimum of 7 subjects who then selected the most appropriate hypothesis. The dataset encompasses 1,000 triplets, each containing one reference, along with two hypotheses and their corresponding number of votes.

By tallying the frequency with which a metric aligns with human annotations (*i.e.* indicating the best score for the hypothesis chosen by humans), we can compute a ratio indicative of the correlation with human perception.

The SemDist metric demonstrates the most robust correlation with human perception according to the findings from the HATS dataset. Our study aims to explore whether the integration of the minED paradigm reduces the correlation with human perception compared to the utilization of the metric in isolation.

	Transcription	**Translation**	**SemDist**	**BERTScore**
Reference	à nos résultats	to our results		
Hypothesis	un non résultat	a no result	57.8	28.1
Corrected Hypotheses	à non résultat	to no result	50.1 (+7.7)	23.6 (+4.5)
	un nos résultat	a our result	20.2 (+37.6)	21.4 (+6.7)
	un non résultats	a no results	52.7 (+5.1)	28.0 (+0.1)

Table 1: SemDist and BERTScore improvements due to correcting the hypothesis "*à nos résultats*". Scores are projected in a lower-is-better rule and a [0, 100] scale for better readability.

3 Integrating Metrics for Interpretability

The minED paradigm is designed to enhance the interpretability of metrics yielding scores that are challenging to comprehend. To do this, we integrate a non-interpretable metric such as SemDist into minED. This involves computing the minimum number of modifications required to make the hypothesis sufficiently close to the reference regarding human perception. We extend this method to both words (minWED) and characters (minCED). The paradigm is described in Section 3.1, while Section 3.2 addresses the parameter setting of the method. We then discuss two types of metrics (consistent, inconsistent) influencing computation cost (Section 3.3), and explore the correlation between minED and human perception (Section 3.4).

[6] https://github.com/thibault-roux/metric-evaluator
[7] https://github.com/kaldi-asr/kaldi/blob/master/egs/librispeech/s5/

3.1 Minimum Edit Distance (minED)

Word or character correction entails editing the hypothesis to eliminate substitutions, insertions, or deletions. The minED paradigm calculates the minimum number of corrections (words or characters) necessary to render a hypothesis "acceptable" based on a non-interpretable metric. To do so this, we construct a graph representing all possible modifications to the hypothesis that align it with the reference (refer to Appendix, Figure 4). For each corrected token, we compute a score between the reference and the adjusted hypothesis using the integrated metric. If the score falls below a predefined threshold, the hypothesis is considered "acceptable", obviating the need to traverse the rest of the graph. Human acceptability serves as a prerequisite, and the score denotes the minimum level of edits, allowing for some errors in the hypothesis. Establishing the threshold is pivotal, as detailed in Section 3.2.

3.2 Setting the threshold of acceptability

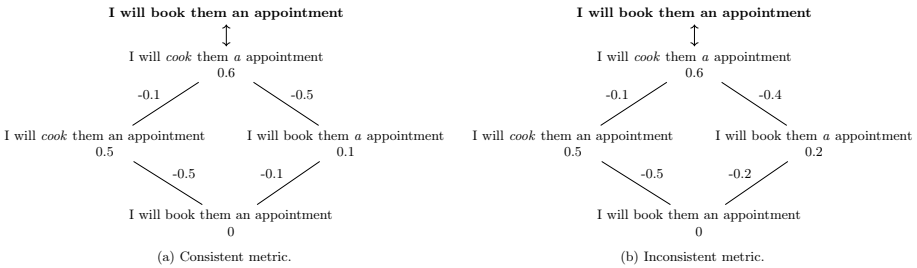
As discussed in Section 3.1, minED represents the required edits for an acceptable hypothesis. This concept hinges on identifying a metric value deemed acceptable to humans. For instance, if a semantic metric indicates a score below the threshold (lower-is-better) when a human reads an erroneous hypothesis, the original sentence's meaning should be comprehensible.

Setting the threshold too low tends to align minWED and minCED metrics with WER or CER values. Conversely, excessively high threshold values lead these metrics to converge towards zero scores, suggesting no corrections are necessary. One approach could involve selecting a threshold maximizing correlation with human perception, though alternative methods should not be discounted.

3.3 Consistency of metrics

Correcting a hypothesis to align it with the reference can yield improvements in the score according to the integrated metric. Such corrections can have two effects: they either enhance the score regardless of prior modifications (see Figure 1a) or improve the score based on preceding modifications (see Figure 1b). For instance, in Figure 1a, rectifying the substitution "cook/book" improves the metric performance by 0.5, irrespective of whether "an/a" was corrected. Conversely, in Figure 1b, rectifying "cook/book" enhances the metric performance by 0.5 or 0.4, contingent on whether "an/"" was corrected.

The consistency property allows faster computation of the minimum number of edits as it is no longer necessary to compute the entire graph. Instead, a pragmatic approach involves computing the second level, where a single error in the hypothesis is corrected. Subsequently, subtracting the original hypothesis score from the minimum improvements required for the resulting score to fall below the threshold. WER and CER exemplify consistent metrics, while BERTScore and SemDist exemplify inconsistent metrics.



(a) Consistent metric. (b) Inconsistent metric.

Fig. 1: Comparison of the impact of correction on consistent and inconsistent metrics. Metrics are based on a lower-is-better rule.

3.4 Correlation with Human Perception

Figure 2 illustrates the correlation between human perception and minED across various threshold (θ) values. Lower thresholds result in correlations closer to the embedded metric, while higher values lead to diminished performance. Conversely, excessively high values cause a decline in performance.

While minWED experiences a 21.56% reduction in correlation compared to SemDist, it achieves a 5.12% improvement over WER, rendering it more interpretable. Similarly, minCED correlates more strongly with human perception than CER but demonstrates a notable loss compared to SemDist. The limited granularity of metrics based on word/character edit distance constrains their efficacy in evaluating ASR transcripts from a human perspective.

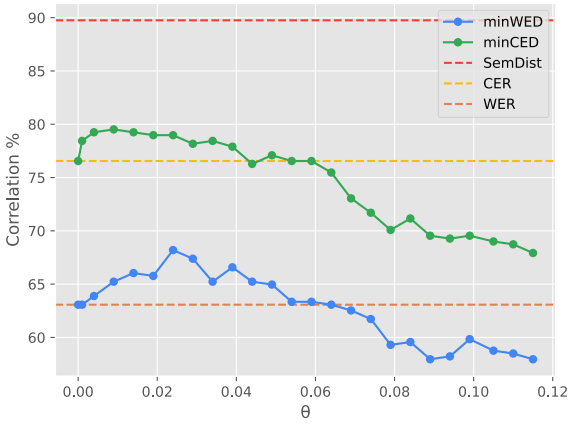


Fig. 2: Minimum Edit Distance's correlations with HATS data set according to various threshold values (θ).

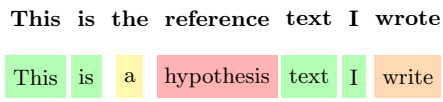


Fig. 3: Visualization of error severity according to our paradigm incorporating semantic metric.

4 Measuring Error Severity

In this section, we investigate the ability of our paradigm to identify errors and gauge their severity, as depicted in Figure 3. Section 4.1 outlines our evaluating method for error severity, while Section 4.2 delves into the results and analysis.

4.1 Evaluation protocol

To assess the paradigm's ability to measure error severity, we assume that rectifying a severe error should exert a more significant impact on a downstream task than rectifying a minor one.

In our study, we selected a French-to-English translation task from speech data. This task commences with automatic transcription, serving as the ASR intrinsic evaluation using SemDist and CER metrics. The transcription subsequently undergoes translation to produce the final hypothesis, facilitating the ASR extrinsic evaluation using BLEU and BERTScore metrics.

As outlined in Table 1, we generate corrections to an erroneous hypothesis proportional to the number of transcription errors. This approach enables us to ascertain, for each correction, the improvement score for both intrinsic and extrinsic metrics. The presence of a correlation between these values indicates the paradigm's efficacy in measuring error severity.

Our experimental setup leverages the HATS dataset to procure references and associated erroneous hypotheses, with translations generated using Google Translator.

Additionally, we conducted a secondary experiment utilizing the Word Importance corpus [8], comprising 25,000 English tokens. Word importance is defined as the impact of omitting a word from a transcription on overall comprehension. Consequently, we can compute the correlation between SemDist improvement after the correction of a deletion and each importance score.

4.2 Results and analysis

Table 2 showcases Pearson's correlation between intrinsic and extrinsic automatic transcription improvement for the translation task. A notable correlation between SemDist and BERTScore is observed, especially in comparison with the correlation obtained with CER as an intrinsic metric. Different correlations emerge for intrinsic and extrinsic metrics, suggesting potential variations in results for tasks other than translation. For the second experiment, the best English

Sentence-BERT [8] yielded a Pearson correlation of 0.69 on the Word Importance corpus.

The results of these experiments, carried out for the first time to our knowledge, demonstrate the ability of this paradigm to identify error severity, and could be used as a baseline.

Intrinsic/Extrinsic	**BERTScore**	**BLEU**
SemDist	0.41	0.27
CER	0.24	0.23

Table 2: Average Pearson's correlation between intrinsic and extrinsic improvements across various metrics for the translation task.

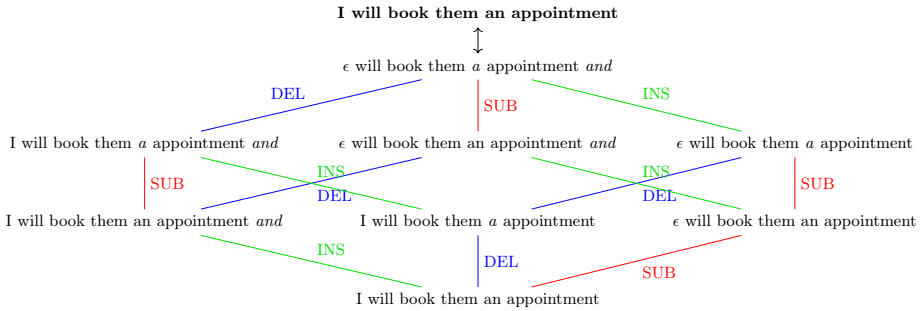


Fig. 4: Computed graph of each possible modification to an error-free hypothesis with the minWED paradigm. Each edge corresponds to a corrected error. The token ϵ corresponds to deletions.

5 Conclusions and perspectives

We have introduced a paradigm that not only enhances the interpretability of ASR metrics but also facilitates the measurement of error severity. The minED approach offers a transparent framework for evaluating ASR systems. While our investigation revealed a noticeable decrease in correlation with human perception when integrating a metric in minWED (on words), our findings demonstrate that minCED (on characters) maintains relatively strong performance in capturing error perception compared to a broad range of previously evaluated metrics [2].

Our study highlights a significant loss of correlation with interpretability, indicating that a mere count of errors - even with semantic consideration - does

[8] https://huggingface.co/sentence-transformers/paraphrase-MiniLM-L3-v2

not reflect human judgment accurately. It appears that humans prioritize error severity over the mere frequency of serious errors.

Moreover, an alternative approach for developing interpretable metrics that closely align with human perception could involve the development of qualitative rather than quantitative metrics. Datasets such as HypRatings [10] incorporate qualitative annotations like '`exact match`', '`useful hyp`', '`wrong hyp`', and '`nonsense hyp`'. Exploring the development of metrics predicting these qualitative features could represent a promising avenue for future research.

In conclusion, our work underscores the importance of not only understanding the numerical output of ASR metrics but also considering the perceptual aspects of error evaluation. By embracing both interpretability and error severity, we can advance the effectiveness of ASR evaluation methods, ultimately enhancing the quality and usability of transcription systems.

6 Appendices

6.1 Properties of edit graph

The graph is constructed with a node representing the hypothesis produced by the ASR system. If the hypothesis contains no errors, there is no edit edge, and one node represents both the hypothesis and the reference. The hypothesis corresponds to the first level, and the reference to the last level, with the number of levels equal to the number of errors + 1. When there are N errors in the hypothesis, there are N possible edits, resulting in N nodes in the second level. As depicted in Figure 4, different edit paths can lead to the same node. If we consider correction as a set (*i.e.* empty when no corrections are made and full when all corrections have been made), we can analogize this graph to a Hasse diagram of a graded partially ordered set of a Power set. Consequently, the graph inherits its properties:

- The number of nodes at level k with n errors = $\binom{n}{k}$
- The total number of nodes given n errors = 2^n

Due to the exponential complexity of the calculation, the process can be computationally expensive. For instance, for a hypothesis with 5 errors, we must calculate the metric for a maximum of 32 nodes. To address this challenge, we propose optimization solutions in Section 3.3.

6.2 Linguistic analysis

Each error in the hypothesis corresponds to either a word in the reference (substituted or deleted) or to an insertion (i.e., a word only present in the hypothesis). Leveraging a state-of-the-art part-of-speech (POS) tagger for French [11], we associate a POS with the words in the reference and analyze which POS holds the most significance in the context of error perception.

Figure 5 presents the SemDist gains per POS tag. Across POS, gains vary: the highest gains are observed for nouns or proper nouns, followed by verbs.

These three word categories carry essential lexical information crucial for sentence meaning. Conversely, POS such as conjunctions or pronouns carry minimal lexical information.

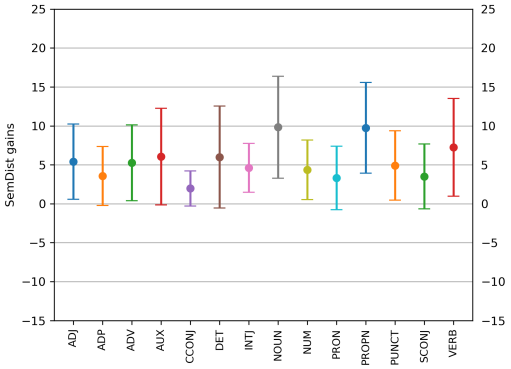


Fig. 5: SemDist gains for each POS tag corrected.

7 Limitations

While this paradigm enhances interpretability, integrating metrics may lead to a reduction in correlation with human perception. The magnitude of this reduction, which depends on the selected threshold, could potentially diminish the relevance of the employed metric. Moreover, the computational overhead of minED can be significant, especially in situations where contemporary metrics exhibit inconsistency, particularly in the presence of a high error rate.

References

1. Bañeras-Roux, T., Rouvier, M., Wottawa, J., Dufour, R.: Qualitative Evaluation of Language Model Rescoring in Automatic Speech Recognition. In: Interspeech (2022)
2. Bañeras-Roux, T., Wottawa, J., Rouvier, M., Merlin, T., Dufour, R.: Hats: An open data set integrating human perception applied to the evaluation of automatic speech recognition metrics. In: Text, Speech, and Dialogue (TSD) (2023)
3. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: North American Chapter of the Association for Computational Linguistics (NAACL). pp. 4171–4186 (2019)
4. Favre, B., Cheung, K., Kazemian, S., Lee, A., Liu, Y., Munteanu, C., Nenkova, A., Ochei, D., Penn, G., Tratz, S., et al.: Automatic human utility evaluation of ASR systems: Does WER really predict performance? In: Interspeech. pp. 3463–3467 (2013)

5. Gordeeva, L., Ershov, V., Gulyaev, O., Kuralenok, I.: Meaning Error Rate: ASR domain-specific metric framework. In: ACM SIGKDD Conference on Knowledge Discovery & Data Mining. pp. 458–466 (2021)
6. Hanna, M., Bojar, O.: A fine-grained analysis of bertscore. In: Machine Translation. pp. 507–517 (2021)
7. Kafle, S., Huenerfauth, M.: Evaluating the usability of automatically generated captions for people who are deaf or hard of hearing. In: International ACM SIGACCESS Conference on Computers and Accessibility. pp. 165–174 (2017)
8. Kafle, S., Huenerfauth, M.: A corpus for modeling word importance in spoken dialogue transcripts. In: Language Resources and Evaluation (LREC) (2018)
9. Kim, S., Arora, A., Le, D., Yeh, C.F., Fuegen, C., Kalinli, O., Seltzer, M.L.: Semantic Distance: A New Metric for ASR Performance Analysis Towards Spoken Language Understanding. In: Interspeech. pp. 1977–1981 (2021). https://doi.org/10.21437/Interspeech.2021-1929
10. Kim, S., Le, D., Zheng, W., Singh, T., Arora, A., Zhai, X., Fuegen, C., Kalinli, O., Seltzer, M.: Evaluating User Perception of Speech Recognition System Quality with Semantic Distance Metric. In: Interspeech. pp. 3978–3982 (2022). https://doi.org/10.21437/Interspeech.2022-11144
11. Labrak, Y., Dufour, R.: Antilles: An open french linguistically enriched part-of-speech corpus. In: Text, Speech, and Dialogue (TSD). pp. 28–38. Springer (2022)
12. Martin, L., Muller, B., Suárez, P.J.O., Dupont, Y., Romary, L., De La Clergerie, É.V., Seddah, D., Sagot, B.: CamemBERT: a Tasty French Language Model. In: Association for Computational Linguistics (ACL). pp. 7203–7219 (2020)
13. Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., et al.: The Kaldi speech recognition toolkit. In: Automatic Speech Recognition and Understanding Workshop (ASRU). Institute of Electrical and Electronics Engineers (IEEE) (2011)
14. Ravanelli, M., Parcollet, T., Plantinga, P., Rouhe, A., Cornell, S., Lugosch, L., Subakan, C., Dawalatabad, N., Heba, A., Zhong, J., Chou, J.C., Yeh, S.L., Fu, S.W., Liao, C.F., Rastorgueva, E., Grondin, F., Aris, W., Na, H., Gao, Y., Mori, R.D., Bengio, Y.: SpeechBrain: A general-purpose speech toolkit (2021), arXiv:2106.04624
15. Reimers, N., Gurevych, I.: Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). pp. 3982–3992 (2019)
16. Ruiz, N., Federico, M.: Phonetically-oriented word error alignment for speech recognition error analysis in speech translation. In: Automatic Speech Recognition and Understanding Workshop (ASRU). pp. 296–302. IEEE (2015)
17. Yilmaz, Z.A., Wang, S., Yang, W., Zhang, H., Lin, J.: Applying bert to document retrieval with birch. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations. pp. 19–24 (2019)
18. Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: Bertscore: Evaluating text generation with bert. In: International Conference on Learning Representations (ICLR) (2020), https://openreview.net/forum?id=SkeHuCVFDr