What is lost in Normalization? Exploring Pitfalls in Multilingual ASR Model Evaluations

Kavya Manohar[1,2] **Leena G Pillai**[1,3]

[1]Digital University Kerala
[2]Swathanthra Malayalam Computing [3]University of Kerala
Correspondence: kavya.manohar@duk.ac.in, leena.g@duk.ac.in

Abstract

This paper explores the pitfalls in evaluating multilingual automatic speech recognition (ASR) models, with a particular focus on Indic language scripts. We investigate the text normalization routine employed by leading ASR models, including OpenAI Whisper, Meta's MMS, Seamless, and Assembly AI's Conformer, and their unintended consequences on performance metrics. Our research reveals that current text normalization practices, while aiming to standardize ASR outputs for fair comparison, by removing inconsistencies such as variations in spelling, punctuation, and special characters, are fundamentally flawed when applied to Indic scripts. Through empirical analysis using text similarity scores and in-depth linguistic examination, we demonstrate that these flaws lead to artificially improved performance metrics for Indic languages. We conclude by proposing a shift towards developing text normalization routines that leverage native linguistic expertise, ensuring more robust and accurate evaluations of multilingual ASR models.

1 Introduction

Automatic speech recognition (ASR) systems have become increasingly relevant in various applications, ranging from voice assistants and transcription services to accessibility tools for the disabled population. The performance and usability of ASR models are evaluated in terms of their error rates. Recent advancements in open ASR models pretrained in self-supervised (Schneider et al., 2019; Babu et al., 2022; Chung et al., 2021) manner or weakly supervised (Radford et al., 2023) manner are capable of handling various languages and scripts. These models can be fine-tuned for improved performance in domains or languages of interest. This capability has revolutionized speech recognition in ultra low resource languages and scenarios (Rouditchenko et al., 2023). Many of these models have brought down state of the art (SOTA) word error rates (WERs) on popular benchmarks.

Evaluation of the performance of ASR models are often affected by the prediction differing from the ground truth in letter casing, punctuation, spelling variants etc. leading to inflated WERs. To mitigate this, a text normalization routine is employed (Deviyani and Black, 2022; Zhang et al., 2021). A proper text normalization routine is required to minimize penalization of non-semantic differences by aligning the predicted output more closely with the ground truth.

The study presented in this paper examines the pitfalls in the current normalizations routines employed in the latest ASR models on the banchmarking of non-English languages, specifically on many Asian languages that use Indic scripts[1]. Our empirical analysis reveals that the current normalization practices can result in significant errors, particularly in many low-resource languages, by boosting the model performance on many benchmarks and misleading the research community. We propose for the development of linguistically informed normalization routines that account for the unique characteristics of each language, ensuring a fair and reasonable evaluation and benchmarking process for multilingual ASRs.

2 Background and Related Works

Prior to the introduction of OpenAI's Whisper model (Radford et al., 2023), most ASR systems were trained on normalized text transcripts and produced output without punctu-

[1]https://en.wikipedia.org/wiki/Brahmic_scripts

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 10864–10869
November 12-16, 2024

Language	Normalization Type	Transcription		Similiarity (METEOR)
		Native Script	Rough IPA	
English	Unnormalized: Normalized:	This is an example. this is an example	ðɪs ɪz ən ɪgˈzɑːmpl ðɪs ɪz ən ɪgˈzɑːmpl	0.97
Finnish	Unnormalized: Normalized:	Tämä on esimerkki. tämä on esimerkki	tiːäɛmä ɒn esiˌmerkːi tiːäɛmä ɒn esiˌmerkːi	0.95
Hindi	Unnormalized: Normalized:	यह एक उदाहरण है । यह एक उद हरण ह	jəɦə ekə udaːɦrəɳə ɦæː jəɦə ekə udə ɦərɳə ɦə	0.38
Tamil	Unnormalized: Normalized:	இது ஒரு உதாரணம். இத ஒர உத ரணம	iṭu oɾu uṭaːɾaɳam iṭa oɾa uṭa ɾaɳama	0.00
Malayalam	Unnormalized: Normalized:	ഇതൊരു ഉദാഹരണമാണ്. ഇത ര ഉദ ഹരണമ ണ	iṭoɾu uḍaːɦaɾaɳamaːɳ iṭa ɾa uḍa ɦaɾaɳama ɳa	0.00
Thai	Unnormalized: Normalized:	นี่คือตัวอย่าง น ค อต วอย าง	niːkʰɯːɔːtuːa̯ɔːjaːŋ na kʰa ɔːta wɔːja aːŋa	0.00

Table 1: A demonstration of the effect of Whisper normalization. While diacritics are retained in non-English languages (eg: Finnish) that uses latin script, the relevant vowel signs and virama sign are lost in Indic scripts. Rough Romanized transcript in IPA is also provided. Text similarity between original and Whisper-normalized text are indicated using METEOR score (Banerjee and Lavie, 2005).

ation or casing. Whisper, however, outputs UTF-8 text, requiring a comprehensive normalization process to accurately evaluate its performance. This ensures that the evaluation metric, WER penalizes only actual word mistranscriptions, not formatting or punctuation differences.

Whisper's normalization routine for English extends beyond basic casing and punctuation, incorporating transformations such as converting contracted abbreviations to expanded forms and expanding currency symbols. However, this approach would require a language-specific set of transformations for non-English text. Due to the lack of linguistic knowledge to develop such normalizers for all languages, the Whisper's normalization relies on a basic data-driven approach, which includes replacement of characters in the `mark` class with spaces and removes successive whitespace characters to a single instance (Radford et al., 2023).

The non-English normalization routine employed by Whisper, inadvertently removes vowel signs (*matras*), that belong to the the `mark` class of Unicode characters. These vowel signs, essential for correct word formation and pronunciation, are removed along with other punctuation marks, leading to significant distortions in the text in languages such as Hindi, Bengali, Tamil, and others (O'Connell, 2023; Manohar, 2024). This results in words being broken down into consonants without their associated vowels, causing a loss of meaning and intelligibility. This also leads to incorrect WER calculations for languages written in Indic scripts. Additionally, Thai, which does not use spaces between words but relies on spaces to delimit sentences, is also affected. The normalization process inserts spaces instead of vowel signs, effectively distorting the nature of the language. See Table 1 for examples with detailed analysis provided in section 3.1.

This normalization routine has been adopted by various later models, including Meta's MMS, Seamless series (Pratap et al., 2024; Barrault et al., 2023a,b) and AssemblyAI's Conformer-1 (AssembyAI, 2023) for evaluation and benchmarking and is integrated into Huggingface transformers[2], thus amplifying its impact.

[2]Normalization in Huggingface Whisper Transformer

3 Methodology

In this study, we present two complementary empirical evaluations to assess the impact of Whisper's normalization routine on different languages. First, we conduct an intrinsic evaluation by comparing the similarity of example sentences from various languages before and after normalization, using the METEOR score as a text similarity metric. Second, we perform an extrinsic evaluation by measuring the WER on a multilingual benchmark dataset for the same set of languages, with and without the application of Whisper's normalization. For our case study, we used both the baseline and fine-tuned Whisper ASR because their outputs include punctuation, unlike other ASR models in the literature. This allowed us to demonstrate the impact of normalization on ASR outputs with punctuation. All the datasets and the models used in this experiments are available under permissive licenses in Huggingface repositories and listed in Appendix A. All the evaluations were run on a single NVIDIA A100 GPU.

3.1 Analysis of Text Similarity after Whisper Normalization

To empirically assess the impact of Whisper's normalization routine on different languages, we conducted a comparative analysis of example sentences from languages that employ various script systems. Specifically, we selected languages that use Latin script (English and Finnish), Indic scripts (Hindi, Tamil, and Malayalam), and South East Asian scripts (Thai). For each language, we prepared a set of example sentences that were identical in meaning but differed in their script and formatting as presented in Table 1. The METEOR score (Banerjee and Lavie, 2005) was employed to quantify the similarity between the original and normalized sentences. It is a text similarity metric that considers the precision, recall, and F-score of the machine-translated text, providing a comprehensive measure of its similarity to the reference text, while also placing importance on the order of words in the text.

The similarity scores we obtained demonstrate the varying impact of Whisper's normalization routine on different languages. The high similarity scores for English (0.97) and Finnish (0.95) indicate that the normalization process preserves the linguistic structure and meaning of these languages very well. The diacritic marks in Finnish are retained without any distortion as indicated in the Table 1. This is because the normalization routine ensures the diacritic marks gets converted to `letter` class of characters using NFKC compatibility composition rules of Unicode[3], before `mark` class of characters are replaced by space.

In contrast, as illustrated in Table 1, the normalization process severely distorts the text in languages other than English and Finnish. The replacement of Unicode characters in the `mark` class, including vowel signs and *virama* symbols, by spaces after Whisper normalization significantly alters the linguistic structure of these languages. While Hindi, with a METEOR score of 0.38, is less affected due to its analytic typology, Malayalam and Tamil are severely impacted (Kumar et al., 2007; Manohar et al., 2020) by the splitting of morphologically complex words at every occurrence of vowel signs and *virama* symbols, leading to similarity scores of 0. Thai, which typically does not use spaces between words, is also affected by the removal of important vowel signs, resulting in a text that is unusable due to excessive spacing and a similarity score of 0.

3.2 Impact of Whisper Normalization on WER

To empirically analyze the impact of normalization on the WER, we present the results of evaluating the original Whisper-small model, referred to as the baseline model, with and without the application of Whisper's normalization on the test split of Google FLEURS (Conneau et al., 2022) multilingual speech dataset.

The left side bar graph in Figure 1 shows that the WER of the baseline model is significantly high for languages other than English and Finnish, with values of 86.9% for Hindi, 93.3% for Tamil, and 287.4% for Malayalam. The baseline ASR model exhibits a WER exceeding 100% for Malayalam due to

[3] http://unicode.org/reports/tr15/

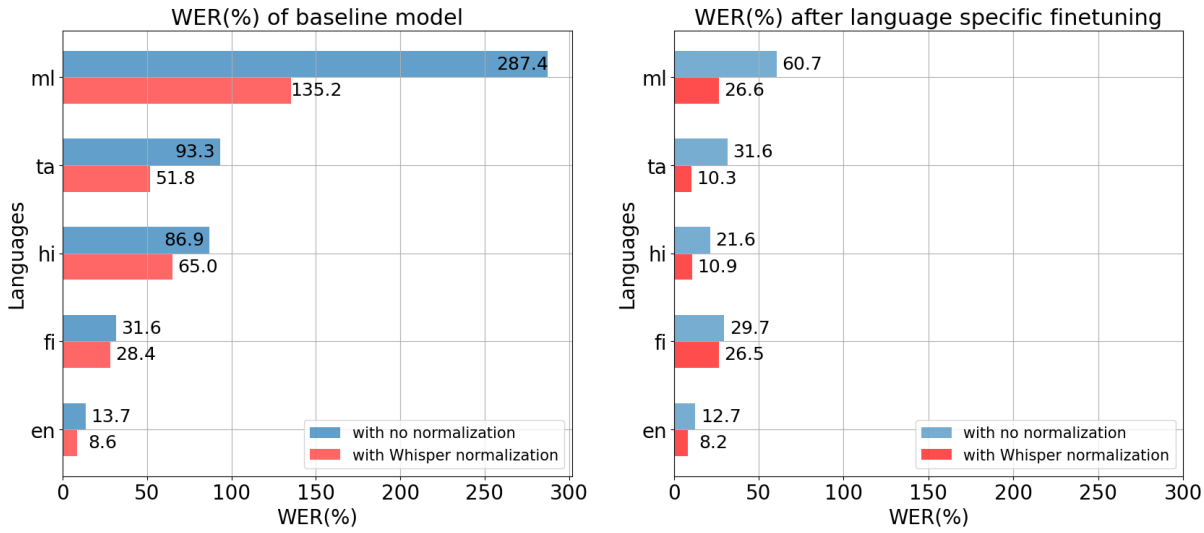

Figure 1: Performance comparison of the OpenAI Whisper-Small model across different languages. The graph on left shows WER on the original model and the one on right shows the result after language specific finetuning. Regular WER are computed on raw transcripts and normalized WER are computed on Whisper normalized transcripts.

a high number of insertion errors, leading to the combined total of substitutions, deletions, and insertions surpassing the total word count in the reference transcript. While the application of Whisper's normalization results in modest WER improvements for English and Finnish, with an absolute reduction of 5.1% and 3.2% respectively, Indic languages experience suspicious absolute WER reductions: 21.9% for Hindi, 41.5% for Tamil, and a substantial 152.2% for Malayalam.

Due to the poor performance of the baseline model on many Indian languages, we conducted a further comparison of WER with and without Whisper's normalization on publicly available models that have been derived from the baseline model after language-specific fine-tuning. The fine-tuned models used in these evaluations are listed in Appendix A. Fine-tuning has significantly improved the performance of the Hindi, Tamil, and Malayalam models.

Fine-tuned models of English and Finnish exhibit a reasonable absolute reduction of 4.5% and 3.2% on WER respectively. In contrast, Indic languages exhibit a substantial absolute reduction in WER, with decreases of 10.7% for Hindi, 21.3% for Tamil, and 34.1% for Malayalam. Notably, the languages that showed the worst similarity scores exhibit the maximum improvement in WER after normalization. This suggests that the normalization process, which breaks most words into a series of consonants and adds spaces, artificially increases the number of words in the reference, thereby reducing the WER.

4 Recommendations

Findings from our empirical evaluation underscore the importance of language-specific normalization routines to ensure accurate text representation and reliable performance evaluation in many underrepresented Indic languages. Building up on our findings, we propose a collaborative approach, leveraging the collective efforts of native speakers and linguistic experts to develop effective normalization routines for diverse linguistic contexts.

5 Conclusions

The empirical evaluation conducted in this study highlights that the current practice of normalization severely affects the text representation across languages, resulting in artificially boosted WER and SOTA performance. By adopting a more tailored approach to evaluations, we can enhance the reliability of multilingual ASR models, making them truly inclusive and effective across diverse linguistic landscapes.

6 Limitations

1. Being a position paper, this study highlights only the limitations of existing normalization techniques, but does not propose new normalization algorithms.

2. The results are based on specific datasets and publicly available models used for evaluating WER. Variability in datasets (e.g., different accents, dialects, or recording conditions) might influence the reported values.

3. The primary metric discussed is WER. Other evaluation metrics (e.g., phoneme error rate, semantic error rate, match error rate) might provide additional insights into the impacts of text normalization.

4. We used the `raw transcription` field of the FLEURS corpus, which could be a reason for the difference from the WER values reported in Radford et al. (2023).

5. While the paper focuses on text normalization on Indian languages there could be other languages which gets affected by the normalization differently.

6. We omitted Thai from WER comparison charts because for languages where space is not a word delimiter, character error rate is the metric reported in Radford et al. (2023).

References

AssemblyAI. 2023. Conformer-1. Accessed on June 8, 2024.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2022. XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale. In *Proc. Interspeech 2022*, pages 2278–2282.

Satanjeev Banerjee and Alon Lavie. 2005. 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*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Loïc Barrault, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, et al. 2023a. SeamlessM4T-Massively Multilingual & Multimodal Machine Translation. *arXiv preprint arXiv:2308.11596*.

Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, Mark Duppenthaler, Paul-Ambroise Duquenne, Brian Ellis, Hady Elsahar, Justin Haaheim, et al. 2023b. Seamless: Multilingual Expressive and Streaming Speech Translation. *arXiv preprint arXiv:2312.05187*.

Yu-An Chung, Yu Zhang, Wei Han, Chung-Cheng Chiu, James Qin, Ruoming Pang, and Yonghui Wu. 2021. W2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 244–250. IEEE.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. 2022. Fleurs: Few-shot learning evaluation of universal representations of speech. *arXiv preprint arXiv:2205.12446*.

Athiya Deviyani and Alan W Black. 2022. Text Normalization for Speech Systems for All Languages. In *Proc. 1st Workshop on Speech for Social Good (S4SG)*, pages 20–25.

G Bharadwaja Kumar, Kavi Narayana Murthy, and BB Chaudhuri. 2007. Statistical analyses of telugu text corpora. *IJDL. International journal of Dravidian linguistics*, 36(2):71–99.

Kavya Manohar. 2024. Indian Languages and Text Normalization: Part 1. Accessed on June 8, 2024.

Kavya Manohar, AR Jayan, and Rajeev Rajan. 2020. Quantitative analysis of the morphological complexity of Malayalam language. In *Text, Speech, and Dialogue: 23rd International Conference, TSD 2020, Brno, Czech Republic, September 8–11, 2020, Proceedings 23*, pages 71–78. Springer.

Ross O'Connell. 2023. Breaking Brahmic: How OpenAI's Text Cleaning Hides Whisper's True Word Error Rate for Many South Asian Languages. Accessed on June 8, 2024.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, et al. 2024. Scaling speech technology to 1,000+ languages. *Journal of Machine Learning Research*, 25(97):1–52.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya

Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR.

Andrew Rouditchenko, Sameer Khurana, Samuel Thomas, Rogerio Feris, Leonid Karlinsky, Hilde Kuehne, David Harwath, Brian Kingsbury, and James Glass. 2023. Comparison of Multilingual Self-Supervised and Weakly-Supervised Speech Pre-Training for Adaptation to Unseen Languages. In *Proc. INTERSPEECH 2023*, pages 2268–2272.

Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. 2019. wav2vec: Unsupervised Pre-Training for Speech Recognition. In *Proc. Interspeech 2019*, pages 3465–3469.

Yang Zhang, Evelina Bakhturina, and Boris Ginsburg. 2021. NeMo (Inverse) Text Normalization: From Development to Production. In *Proc. Interspeech 2021*, pages 4857–4859.

A Resources

We have used the following publicly available models and datasets for our experiments.

ASR Models

1. The baseline model: `https://huggingface.co/openai/whisper-small`
2. The Fine-tuned English: `https://huggingface.co/openai/whisper-small.en`
3. The Fine-tuned Finnish: `RASMUS/whisper-small-fi-15k_sample`
4. The Fine-tuned Hindi: `https://huggingface.co/vasista22/whisper-hindi-small`
5. The Fine-tuned Tamil: `https://huggingface.co/vasista22/whisper-tamil-small`
6. The Fine-tuned Malayalam: `https://huggingface.co/vrclc/Whisper_small_malayalam`

Speech Dataset

1. Google FLEURS: `https://huggingface.co/datasets/google/fleurs`