

Ground Truth-Free WER Prediction for ASR via Audio Quality and Model Confidence Features

Anton Polevoi¹[0009–0000–2216–0468][†], Alexander Kragin²[0009–0009–5152–887X][†],
and Natalia Loukachevitch³[0000–0002–1883–4121]

¹ Lomonosov Moscow State University

² AI Talent Hub, ITMO University

³ Research Computing Center, Lomonosov Moscow State University

Abstract. We propose a data-driven approach for predicting Word Error Rate (WER) without requiring ground truth transcriptions. Our method involves creating diverse audio datasets by applying various noise types, acoustic degradations, and room impulse responses to clean speech samples across many fine-grained quality and intelligibility levels. Unlike previous work, we extract and analyze a comprehensive set of speech quality features including signal-to-noise ratio (SNR) estimates, modern neural audio quality metrics (such as NISQA), and ASR (Automatic Speech Recognition) model confidence scores to train WER prediction models. We conduct experiments across multiple languages with state-of-the-art ASR architectures (Whisper and FastConformer) to demonstrate our method’s effectiveness in predicting WER in diverse acoustic conditions. We also show that our approach generalizes in a multilingual model-unified setting. We provide feature importance analysis to identify key metrics needed to predict WER. This work enables practical applications such as quality-based filtering of audio inputs, allowing ASR systems to assess expected performance and estimate transcription reliability without ground truth transcripts.

Keywords: ASR · Word error rate prediction · Audio augmentation dataset.

1 Introduction

Deep neural networks have demonstrated remarkable performance in controlled settings, yet their robustness in real-world applications—particularly in automatic speech recognition (ASR)—remains a persistent challenge. A critical gap exists in diagnosing ASR failures before deployment, especially when acoustic degradations or linguistic complexity induce errors ranging from minor inaccuracies to catastrophic hallucinations. Traditional evaluation relies on post-hoc transcription analysis ([40]), which is impractical for real-time systems. Instead, we argue that predicting Word Error Rate (WER) from simulated degradations offers a proactive solution: by synthetically replicating real-world distortions

[†] Equal contribution.

(e.g., noise, reverberation, or packet loss), we can directly correlate degradation severity with ASR failure rates, where higher predicted WER signals imminent model unreliability.

The primary causes of faulty transcriptions fall into two categories:

- **Acoustic Degradations:** Environmental noise, microphone artifacts, reverberation, and signal distortion—all synthetically simulatable to stress-test ASRs.
- **Speech Complexity:** Accents, rare words, and domain-specific terms, which require curated datasets to evaluate robustness.

Crucially, while linguistic challenges are harder to simulate, acoustic degradations can be precisely controlled and reintroduced into clean audio. This enables systematic analysis of how specific distortion types (e.g., additive noise vs. clipping) and their intensity (e.g., SNR levels) degrade ASR performance. By modeling the relationship between these degradations and WER, we can preemptively flag audio inputs likely to cause failures, even without ground-truth transcriptions.

Our work is based on the core theses:

- **Simulated degradations expose ASR vulnerabilities.** By artificially degrading clean audio with real-world distortions (e.g., dynamic noise profiles or impulsive interruptions), we create a controlled environment to quantify how specific degradation types and levels correlate with WER elevation—a direct proxy for model failure.
- **WER prediction is feasible without transcripts.** Using two feature groups—(1) ASR Model Confidence Metrics, (2) Audio Quality Features (WADA SNR, SI-SNR, NISQA), we demonstrate that WER can be accurately predicted, enabling real-time diagnostics.

Our contributions include:

- A framework for simulating diverse real-world degradations (e.g., variable SNR noise, reverberation tails, codec artifacts) and evaluating their impact on state-of-the-art ASR models (Whisper, FastConformer).
- A regression model to predict WER solely from Speech Quality Features and ASR Model Confidence Metrics.

2 Background

2.1 Acoustic Degradations for ASR Model

Existing research on ASR robustness had primarily focused on defending models against acoustic degradations. According to Shah et al. [2], these approaches can be categorized as either model-based or feature-based.

Model-based approaches include adapting pre-trained models [4, 5], pre-processing audio with denoising techniques [6, 7], and training directly on noisy data [8].

These strategies typically require access to representative noisy data and are most effective when the deployment environment and noise characteristics are known in advance.

Feature-based approaches focus on developing noise-invariant representations of speech. These include biologically-inspired features [9] and signal processing techniques [1] designed to extract speech-relevant components while filtering out irrelevant signal elements [10].

In recent years it has been demonstrated that large-scale training on diverse acoustic scenarios can produce inherently robust ASR systems without specialized techniques. Notably, Radford et al. [3] showed that models like Whisper achieve substantial robustness through exposure to varied audio conditions during training.

Typically, the performance of a new ASR system is evaluated on multiple standard speech datasets, which inherently contain varying levels of noise and signal quality. This approach by itself provides some insight into the system’s robustness. However, a more direct method to assess robustness across different conditions involves simulating various signal degradations and measuring the Word Error Rate (WER) at different Signal-to-Noise Ratio (SNR) levels, as demonstrated in numerous studies on ASR robustness [3] [32] [33] [34].

2.2 Word Error Rate (WER)

Word Error Rate (WER) is the standard evaluation metric for automatic speech recognition systems. It measures the distance between a reference transcript and the ASR model’s hypothesis by calculating the minimum number of word-level operations (substitutions, insertions, and deletions) required to transform the hypothesis into the reference.

WER is defined as:

$$\text{WER} = \frac{S + D + I}{N} \quad (1)$$

where S is the number of substituted words, D is the number of deleted words, I is the number of inserted words, N is the total number of words in the reference transcript. Lower WER values indicate better ASR performance, with 0.0 representing perfect transcription. WER can exceed 1.0 in cases where the number of insertion errors is particularly high.

2.3 WER Prediction

Because WER serves as an effective indicator of ASR performance, the main motivation for developing WER prediction methods is to assess the quality of ASR predictions without ground-truth transcriptions, which are often unavailable in real-world applications.

Several methods have been proposed to predict WER without ground truth transcriptions. For instance, eWER [11] detects disagreement between 2 ASR systems (world-level and character-level) and uses that to predict WER. eWER-2

[12] utilizes acoustic, lexical, and phonotactic features for the same task. eWER-3 [13] demonstrated that their multilingual model outperforms previous monolingual WER prediction methods (eWER2) by achieving a 9% absolute increase in Pearson correlation coefficient (PCC), showing better correlation between predicted and reference WER. A common downside of these methods is the need for either 2 separate ASR models or additional models to process text and extract phonemes, which increases computational requirements and overall complexity.

Other approaches include Litman et al. [14], who calculated prosodic statistics (frequencies, energy values, tempo, pauses) and used an ML model to predict WER for telephone dialogue recordings. Fish et al. [15] estimated SNR and decomposed WER into base WER (from inherent model limitations) and delta-WER (from audio quality degradations). Additionally, Gallardo et al. [16] investigated speech quality measurement techniques like POLQA [22] to predict WER. Their results led to polynomial models for predicting speech recognition accuracy from instrumental measures across various channel distortions in different bandwidths.

Some of the most recent works by Park et al. [24, 23] achieve impressive prediction accuracy, but rely on using natural language processing techniques to incorporate information from ASR hypothesis transcriptions.

To summarise, our approach differs in several key ways:

- We predict WER using only audio quality features and confidence scores, avoiding extra complexity associated with using additional natural language processing and/or phone recognition models.
- Our training and evaluation pipelines both use speech samples in different languages under diverse acoustic conditions, including different types of noise and various signal degradations.
- We conduct experiments with the most common modern ASR architectures, namely Whisper [3] and FastConformer [30].
- Our approach relies on a lightweight WER regressor model combined with a modular feature extraction architecture. This design ensures flexibility, allowing researchers to incorporate new speech quality estimation methods as they emerge.
- We provide feature importance comparisons to facilitate further research into ASR WER prediction and speech quality metrics.

3 Proposed Solution

3.1 Degradation Framework

To train a robust WER prediction model that performs reliably across diverse acoustic conditions, we need a comprehensive dataset. We create this dataset by creating speech recordings that span a wide range of scenarios. Starting with clean speech recordings, we systematically add different types of noise at varying intensity levels and apply additional augmentations (aliasing, real impulse responses, MP3 compression, room simulation), as detailed in Algorithm 1. This

synthetic data generation strategy allows us to produce a large, diverse dataset with known ground truth transcriptions and controlled degradations, enabling our WER prediction model to learn patterns of ASR performance across a broad spectrum of real-world acoustic environments.

We design the data generation pipeline as a highly modular system to support adding arbitrary ASR models and audio feature extractors, both local and available through external APIs.

Algorithm 1 Dataset Creation

Require: Clean speech dataset with transcriptions, noise files, gain levels, ASR transcriber, feature extractors

Ensure: Dataset of features paired with ground truth WER values

```

1: Filter dataset to desired duration range
2: Sample subset of audio files from filtered dataset
3: Initialize results collection
4: for each audio file with transcript do
5:   Load and normalize audio
6:   for each noise gain level do
7:     Initialize record with metadata
8:     Select random noise and extract segment matching audio length
9:     Scale noise by gain factor and mix with clean audio
10:    Apply normalization if needed
11:    Apply augmentations
12:    Generate ASR transcription
13:    Compute ASR confidence scores from model logits
14:    Calculate SNR and WER
15:    Extract all quality features using feature extractors
16:   end for
17: end for
18: Save raw results
19: Compute aggregated statistics by gain level
20: return Complete dataset

```

3.2 Speech Quality Feature Extractors

To obtain audio quality and speech intelligibility estimates we leverage 3 distinct speech quality assesment methods as feature extractors for the WER prediction model:

1. WADA-SNR (Waveform Amplitude Distribution Analysis) [26]: a classical statistical approach that estimates signal-to-noise ratio by modeling clean speech as a Gamma distribution and noise as a Gaussian distribution.
2. SpeechBrain SI-SNR Estimator [25]: a modern neural approach that blindly estimates Scale-Invariant Signal-to-Noise Ratio.

3. NISQA (Non-Intrusive Speech Quality Assessment) [27]: a modern neural approach to provide multidimensional audio quality assessment. NISQA evaluates not only overall quality but also specific dimensions (noisiness, coloration, discontinuity, and loudness).

These feature extractors are selected for speed, ease of use and the ability to capture different aspects of audio quality.

3.3 ASR Model Confidence Metrics

To quantify ASR model uncertainty, we utilise a variety of confidence metrics derived from the model’s token-level or char-level log probabilities (log-probs). These confidence metrics are used as features to provide signals about the model’s certainty in its predictions. We implement and evaluate multiple confidence formulations to determine which best correlate with transcription accuracy:

- **Least Confidence:** Normalizes the uncertainty represented by the difference between the maximum probability and 1.0.

$$LC = \frac{1 - \max_i(p_i)}{1 - \frac{1}{|\mathcal{V}|}} \times \frac{|\mathcal{V}|}{|\mathcal{V}| - 1} \quad (2)$$

- **Margin Confidence:** Measures the difference between the probabilities of the top two predictions.

$$MC = 1 - (p_{(1)} - p_{(2)}) \quad (3)$$

- **Ratio Confidence:** Calculates the ratio between the second highest and highest probabilities.

$$RC = \frac{p_{(2)}}{p_{(1)} + \epsilon} \quad (4)$$

- **Entropy-Based Confidence:** Uses normalized Shannon entropy of the probability distribution.

$$EC = -\frac{1}{\log_2(|\mathcal{V}|)} \sum_{i=1}^{|\mathcal{V}|} p_i \log_2(p_i + \epsilon) \quad (5)$$

- **Perplexity Confidence:** Applies normalized perplexity measure.

$$PC = \frac{\exp(-\sum_{i=1}^{|\mathcal{V}|} p_i \log p_i) - 1}{|\mathcal{V}| - 1} \quad (6)$$

- **Gini Coefficient Confidence:** Adapts the Gini coefficient to measure prediction inequality.

$$GC = \frac{2 \sum_{i=1}^{|\mathcal{V}|} i \cdot p_{(i)}}{|\mathcal{V}| \sum_{i=1}^{|\mathcal{V}|} p_{(i)}} - \frac{|\mathcal{V}| + 1}{|\mathcal{V}|} \quad (7)$$

- **Mean Chosen Confidence:** Takes the mean of maximum logprobs across tokens.

$$\text{MCC} = \frac{1}{n} \sum_{j=1}^n \max_i (\log p_{i,j}) \quad (8)$$

Where p_i represents the probability of token i , $p_{(i)}$ denotes the i -th largest probability in the distribution, $|\mathcal{V}|$ is the vocabulary size, ϵ is a small constant to prevent division by zero, and n is the number of tokens in the sequence. We calculate these metrics for each token in the ASR output and average them to obtain utterance-level confidence scores.

These confidence metrics capture different aspects of prediction uncertainty: entropy-based measures assess overall distribution spread, margin and ratio confidence focus on ambiguity between top predictions, while Gini coefficient measures the inequality of the probability mass.

3.4 WER Prediction Model

We propose a robust WER prediction framework that utilises ASR model confidence scores as well as the speech quality feature extractors listed above to account for various audio degradations. These features serve as inputs to a regression model. We evaluate several approaches: classical (Linear Regression, Random Forest, CatBoost [38]), AutoML Ensembles (LightAutoML [39]) and modern DL tabular methods (TabPFN [37], [36]). TabPFN with vanilla parameters was used for the final experiments.

4 Experiments

Data. We run our experiments on 2 distinct datasets to account for differences in data and explore a multilingual setting.

- 1,000 randomly selected recordings from Fleurs [18] (Russian)
- 1,000 randomly selected recordings from LibriSpeech [19] (English)

First, we randomly split the original recordings into train and test sets, then apply degradations separately within each set to avoid mixing augmented versions across splits.

Noise Application. We implement a linear noising schedule with 64 discrete steps, ranging from SNR +60 dB (clean audio) to SNR -10 dB (extremely degraded audio that is challenging or impossible for humans to transcribe). For each recording, we randomly apply one of the following noise types:

1. Bank-easy: Background sounds from a bank environment with minimal background speech
2. Office-easy: Background sounds from an office environment with minimal background speech

3. Office-hard: Background sounds from an office environment containing audible background speech
4. White noise: Standard white noise distortion

Audio Augmentation. In addition to controlled noise addition, we apply various audio augmentations using the Audiomentations library [35]. Our augmentation pipeline includes aliasing effects (50% probability), impulse response convolution (30% probability) using responses from the MIT IR Survey [31], MP3 compression artifacts (50% probability), and room simulation effects (70% probability). These augmentations introduce realistic audio degradations commonly encountered in real-world scenarios.

ASR Models. Our experiments utilize a diverse set of ASR models to represent modern speech recognition approaches across different parameter scales:

- Whisper [3] variants: Turbo (800M) and Small (240M)
- FastConformer [30] with CTC [28] Decoder (120M)

For inference, we use the original OpenAI Whisper implementation for the Whisper variants and Nvidia NEMO Toolkit [29] for FastConformer.

All models are evaluated in both English and Russian languages to assess cross-lingual robustness. This selection is meant to encompass modern SOTA ASR approaches across different model sizes.

4.1 Adequacy of Proposed Features

SNR Estimation in the Presence of Audio Augmentations. Our experiments provide insights into how different SNR estimation methods perform under varying audio conditions, as illustrated in Figure 1. We find that WADA SNR, which relies on simple statistical assumptions about the distribution of speech and noise, performs admirably under standard noising conditions (correlation of 0.8983). However, its performance deteriorates substantially (dropping to 0.6246) when confronted with complex audio augmentations such as MP3 compression artifacts, impulse response convolution, and room simulations.

In contrast, more sophisticated modern neural-based techniques like Speech-Brain SI-SNR models appear to be more resilient. This suggests that modern neural approaches to SNR estimation have implicitly learned more robust representations that can better generalize across various audio degradation types beyond simple additive noise.

Correlation Between Speech Quality Features, ASR Model Confidence Metrics and WER. We observe that ASR Model Confidence Metrics exhibit the strongest correlation with WER, generally achieving correlation coefficients of approximately 0.8, as shown in Table 1. Despite our intuitive expectations that Speech Quality Features would serve as the most reliable predictors of ASR performance, the model’s own confidence score proved to be better indicators of WER.

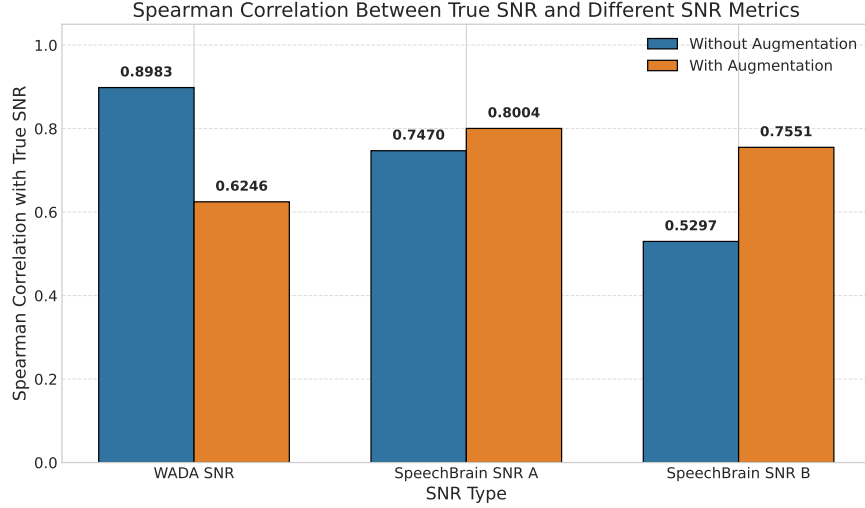


Fig. 1. Spearman correlation between true SNR and different SNR estimation techniques, comparing performance with and without audio augmentations. WADA SNR shows significant performance degradation with augmentations, while SpeechBrain model demonstrates greater robustness.

4.2 Results

WER Prediction Performance Across Models and Languages. Our WER prediction models demonstrate strong performance across different ASR architectures and languages, as shown in Figure 2. There are several notable patterns:

First, we observe that WER prediction accuracy varies by language, with consistently better performance on English data across all ASR models. This is likely attributable to the prevalence of English in training data for both ASR systems and audio quality assessment models, resulting in more reliable feature extraction and confidence scoring. For instance, Whisper Turbo achieves an RMSE of 0.0633 on LibriSpeech (English) compared to 0.0768 on Fleurs (Russian).

Second, our multilingual approach combining Russian and English data into a single prediction model maintains strong performance despite the challenges of cross-language generalization. As illustrated in Figure 2, the combined multilingual model achieves competitive RMSE values (0.0747 for Whisper Turbo, 0.0943 for Whisper Small, and 0.1060 for FastConformer), demonstrating the feasibility of language-agnostic WER prediction. This suggests that the fundamental relationship between audio quality metrics, confidence scores, and transcription accuracy generalizes across languages, at least for the high-resource languages tested.

Table 1. Correlation coefficients between various features and WER. ASR confidence metrics consistently show higher correlation (approximately 0.8) compared to speech quality metrics.

Correlation Group	Feature	Correlation Coefficient (r)
High ($ r \geq 0.7$)	Entropy Confidence	0.82
	Variance Confidence	-0.81
	Least Confidence	0.81
	Temperature Scaled Confidence	0.80
	Margin Confidence	0.79
	Jensen Shannon Confidence	-0.78
	Mean Chosen Confidence	-0.78
	Top K Entropy Confidence	0.76
Mid ($0.2 \leq r < 0.7$)	Gini Confidence	-0.69
	Ratio Confidence	0.69
	NISQA MOS	-0.56
	SpeechBrain SNR A	-0.51
	Perplexity Confidence	0.50
	NISQA Discontinuity	0.28
	SpeechBrain SNR B	-0.26
	NISQA Noisiness	-0.25
	NISQA Coloration	-0.25
Low ($ r < 0.2$)	NISQA Average	-0.19
	WADA SNR	-0.09
	NISQA Loudness	0.05

Third, among the ASR architectures evaluated, Whisper Turbo consistently outperforms smaller models in WER prediction accuracy across all language settings. With an RMSE of 0.0747 on the combined multilingual dataset, Whisper Turbo demonstrates a 20.8% relative improvement over Whisper Small (0.0943) and a 29.5% improvement over FastConformer (0.1060). This performance gap suggests that larger, more capable ASR models not only transcribe more accurately but also provide more reliable confidence scores that correlate better with actual transcription quality.

A detailed breakdown of the metrics including correlation coefficients is available in Table 2.

Table 2. RMSE and correlation results when predicting WER for a particular ASR model.

Model	Size	LibriSpeech (en) Only			Fleurs (ru) Only			Combined Multilingual		
		RMSE	Pearson	Spearman	RMSE	Pearson	Spearman	RMSE	Pearson	Spearman
3-11										
whisper-large-v3-turbo	809M	0.0633	0.9891	0.9424	0.0768	0.9827	0.9145	0.0747	0.9847	0.9179
whisper-small	244M	0.0752	0.9817	0.9545	0.1077	0.9527	0.9399	0.0943	0.9690	0.9541
fast-conformer	120M	0.0758	0.9828	0.9340	0.0866	0.9682	0.8942	0.1060	0.9604	0.9168

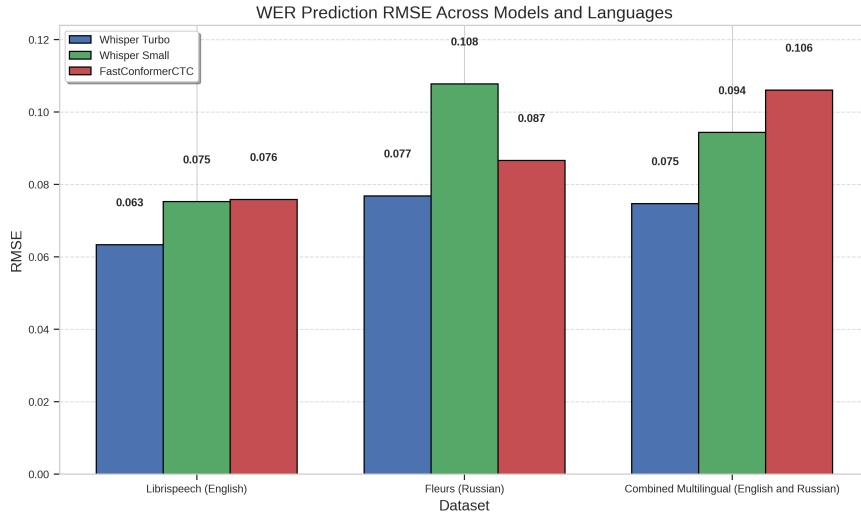


Fig. 2. RMSE for WER prediction across different models and datasets. Lower values indicate better prediction accuracy. The multilingual dataset combines data from Russian and English datasets.

Model-unified Multilingual WER Prediction. While model-specific WER predictors demonstrate superior performance, maintaining separate predictors for each ASR model and language combination quickly becomes impractical in real-world deployments. To address this challenge, we investigate a model-unified approach that enables WER prediction across different ASR architectures and languages using a single prediction model. This approach offers significant practical advantages: it eliminates the need for continuous retraining as new ASR models are developed, provides a unified quality assessment framework across multilingual applications, and simplifies system architecture. Furthermore, as speech recognition technology evolves, a generalizable predictor that works reasonably well with new ASR models without immediate retraining provides a more sustainable and scalable solution for real-world production deployment.

Our model-unified approach involves training a regression model on the combined data from all ASR systems (Whisper Turbo, Whisper Small, and FastConformer) across both English and Russian languages. This unified predictor relies on the same feature set described earlier, including audio quality metrics and ASR confidence scores, but learns to generalize across different ASR models and architectures.

As shown in Figure 3, the model-unified approach achieves an RMSE of 0.0994 on the combined multilingual dataset. While this performance is slightly worse than the best model-specific predictor (Whisper Turbo at 0.0747, representing a 24.9% relative difference), it outperforms the FastConformer-specific predictor (0.1060) by 6.2%. This indicates that a single prediction model can esti-

mate WER reasonably well across different ASR architectures without requiring separate predictors for each system.

The effectiveness of this approach suggests that despite architectural differences between ASR models, there exist common patterns in how speech quality and model confidence scores correlate with transcription accuracy. As new ASR models are developed, existing model-unified WER predictors may still provide reasonable estimation accuracy without requiring immediate retraining, although periodic fine-tuning with data from newer models would likely improve performance over time.

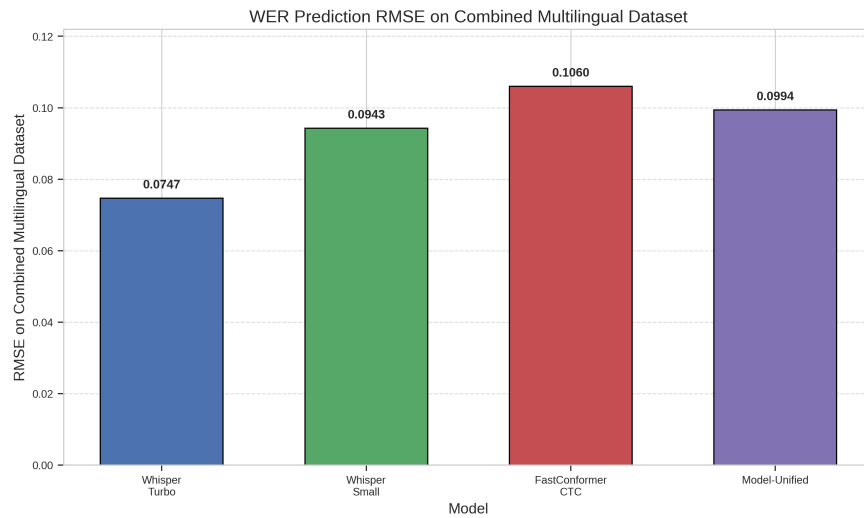


Fig. 3. RMSE for WER prediction for the combined multilingual dataset. Lower values indicate better prediction accuracy. The multilingual dataset combines data from Russian and English datasets. Model-unified approach involves training a single predictor for all 3 ASR models.

5 Ablation Studies

To understand the relative importance of different feature groups in our WER prediction framework, we conducted ablation experiments by systematically removing feature categories and evaluating the impact on prediction accuracy. Figure 4 shows the results of these tests for Whisper Turbo on the LibriSpeech (English) dataset.

Our ablation results reveal that confidence metrics are the most important features for accurate WER prediction. Removing them causes the RMSE to increase dramatically from 0.0633 to 0.1609, representing a 154.2% relative degra-

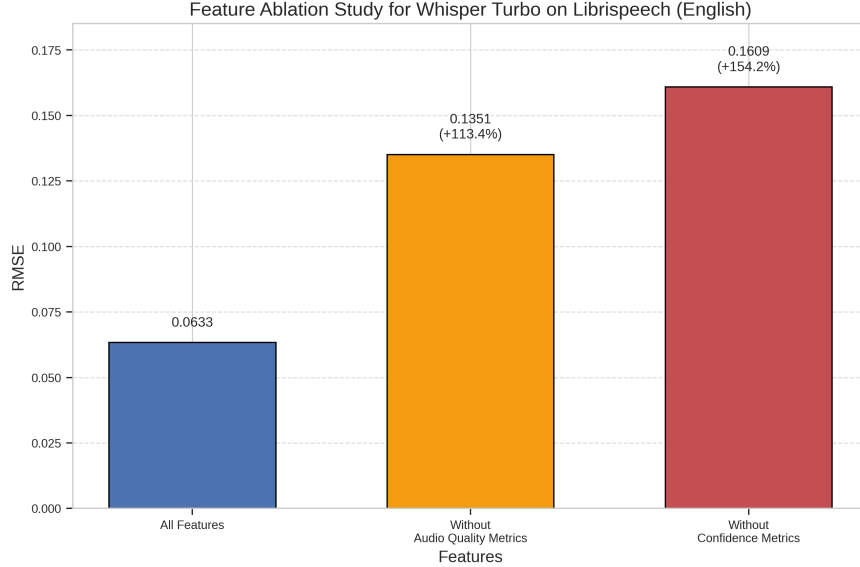


Fig. 4. Feature ablation experiments showing the impact of removing different feature categories on WER prediction performance. Percentages indicate relative RMSE increase compared to using all features.

dation. This result aligns with our correlation analysis in Figure 1, which showed confidence scores having the strongest relationship with WER.

Audio quality metrics also contribute substantially to prediction accuracy, though to a lesser extent than confidence features. Without audio quality metrics, the RMSE increases to 0.1351, a 113.4% relative increase. This shows that while confidence scores provide the strongest predictive signal, audio quality metrics also capture important information that significantly improves overall prediction accuracy.

Interestingly, even without confidence metrics (which require running ASR inference), the model still achieves reasonable prediction performance using only audio quality features. In scenarios where computational resources are limited or where a quick quality assessment is needed before deciding whether to run a full ASR pipeline, audio quality metrics alone can provide a useful rough approximation of expected transcription accuracy. For applications such as filtering out low-quality audio that would likely result in unreliable transcriptions, this approach offers a computationally efficient pre-screening mechanism without requiring full ASR inference.

6 Conclusion

This paper presents a robust approach for WER prediction without requiring ground truth transcriptions. Our key contributions include:

First, we developed a synthetic data generation pipeline that creates diverse acoustic scenarios ranging from clean to heavily degraded speech. This approach enables comprehensive evaluation of ASR robustness across controlled degradation conditions.

Second, we demonstrated that a combination of speech quality metrics and ASR confidence scores can effectively predict WER across different ASR architectures and languages. Our analysis reveals that while modern neural audio quality metrics show promise, ASR confidence scores remain the strongest predictors of transcription accuracy.

Third, we showed that multilingual WER prediction is feasible, with our combined English-Russian model achieving strong performance across languages. Furthermore, our model-unified approach enables WER prediction across different ASR architectures using a single prediction system, simplifying deployment in real-world applications.

Our findings have significant implications for practical ASR deployments, enabling systems to estimate transcription reliability without ground truth, facilitating quality-based filtering, and improving overall robustness in diverse acoustic environments. In future work we intend to explore and integrate additional features to further improve prediction accuracy.

References

1. Li, J., Deng, L., Gong, Y., Haeb-Umbach, R.: An overview of noise-robust automatic speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* **22**(4), 745–777 (2014)
2. Shah, M.A., Noguero, D.S., Heikkilä, M.A., Raj, B., Kourtellis, N.: Speech robust bench: a robustness benchmark for speech recognition. *arXiv preprint arXiv:2403.07937* (2024)
3. Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I.: Robust speech recognition via large-scale weak supervision. In: *International conference on machine learning*, pp. 28492–28518. PMLR (2023)
4. Baevski, A., Zhou, Y., Mohamed, A., Auli, M.: wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems* **33**, 12449–12460 (2020)
5. Hsu, W.N., Bolte, B., Tsai, Y.H., Lakhotia, K., Salakhutdinov, R., Mohamed, A.: Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing* **29**, 3451–3460 (2021)
6. Tsilfidis, A., Mporas, I., Mourjopoulos, J., Fakotakis, N.: Automatic speech recognition performance in different room acoustic environments with and without dereverberation preprocessing. *Computer Speech & Language* **27**(1), 380–395 (2013)
7. Loweimi, E., Ahadi, S.M., Drugman, T., Loveymi, S.: On the importance of pre-emphasis and window shape in phase-based speech recognition. In: *Advances in Nonlinear Speech Processing: 6th International Conference, NOLISP 2013, LNCS*, vol. 7911, pp. 160–167. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-38847-7_21

8. Yin, S., Liu, C., Zhang, Z., Lin, Y., Wang, D., Tejedor, J., Zheng, T.F., Li, Y.: Noisy training for deep neural networks in speech recognition. *EURASIP Journal on Audio, Speech, and Music Processing* **2015**(1), 1–14 (2015)
9. Kim, C., Stern, R.M.: Power-normalized cepstral coefficients (PNCC) for robust speech recognition. *IEEE/ACM Transactions on audio, speech, and language processing* **24**(7), 1315–1329 (2016)
10. Stern, R.M., Morgan, N.: Hearing is believing: Biologically inspired methods for robust automatic speech recognition. *IEEE signal processing magazine* **29**(6), 34–43 (2012)
11. Ali, A., Renals, S.: Word error rate estimation for speech recognition: e-WER. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 20–24. *ACL* (2018)
12. Ali, A., Renals, S.: Word error rate estimation without asr output: E-wer2. *arXiv preprint arXiv:2008.03403* (2020)
13. Chowdhury, S.A., Ali, A.: Multilingual word error rate estimation: e-WER3. In: *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1–5. *IEEE* (2023)
14. Litman, D., Hirschberg, J., Swerts, M.: Predicting automatic speech recognition performance using prosodic cues. In: *1st Meeting of the North American Chapter of the Association for Computational Linguistics*, pp. 1–8. *ACL* (2000)
15. Fish, R., Hu, Q., Boykin, S.: Using audio quality to predict word error rate in an automatic speech recognition system. Unpublished technical report (2006)
16. Gallardo, L.F., Möller, S., Beerends, J.: Predicting automatic speech recognition performance over communication channels from instrumental speech quality and intelligibility scores. In: *INTERSPEECH 2017*, pp. 2939–2943. *ISCA* (2017)
17. Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., Morais, R., Saunders, L., Tyers, F.M., Weber, G.: Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670* (2019)
18. Conneau, A., Ma, M., Khanuja, S., Zhang, Y., Axelrod, V., Dalmia, S., Riesa, J., Rivera, C., Bapna, A.: FLEURS: Few-shot Learning Evaluation of Universal Representations of Speech. *arXiv preprint arXiv:2205.12446* (2022)
19. Panayotov, V., Chen, G., Povey, D., Khudanpur, S.: Librispeech: an ASR corpus based on public domain audio books. In: *ICASSP 2015*, pp. 5206–5210. *IEEE* (2015)
20. Rekish, D., Koluguri, N.R., Krizan, S., Majumdar, S., Noroozi, V., Huang, H., Hrinchuk, O., Puvvada, K., Kumar, A., Balam, J., et al.: Fast conformer with linearly scalable attention for efficient speech recognition. In: *ASRU 2023*, pp. 1–8. *IEEE* (2023)
21. Pallett, D.S., Fiscus, J.G., Fisher, W.M., Garofolo, J.S., Lund, B.A., Przybocki, M.A.: 1993 Benchmark Tests for the ARPA Spoken Language Program. In: *Proceedings of the Workshop on Human Language Technology*, pp. 1–12. *ACL* (1994). <https://doi.org/10.3115/1075812.1075824>
22. Beerends, J., Schmidmer, C., Berger, J., Obermann, M., Ullmann, R., Pomy, J., Keyhl, M.: Perceptual Objective Listening Quality Assessment (POLQA), The Third Generation ITU-T Standard for End-to-End Speech Quality Measurement Part I-Temporal Alignment. *AES: Journal of the Audio Engineering Society* **61**(6), 366–384 (2013)
23. Park, C., Lu, C., Chen, M., Hain, T.: Fast Word Error Rate Estimation Using Self-Supervised Representations for Speech and Text. *arXiv preprint arXiv:2310.08225* (2025)

24. Park, C., Chen, M., Hain, T.: Automatic Speech Recognition System-Independent Word Error Rate Estimation. arXiv preprint arXiv:2404.16743 (2024)
25. Subakan, C., Ravanelli, M., Cornell, S., Grondin, F.: REAL-M: Towards Speech Separation on Real Mixtures. arXiv preprint arXiv:2110.10812 (2021)
26. Kim, C., Stern, R.: Robust signal-to-noise ratio estimation based on waveform amplitude distribution analysis. In: INTERSPEECH 2008, pp. 2598–2601. ISCA (2008). <https://doi.org/10.21437/Interspeech.2008-644>
27. Mittag, G., Naderi, B., Chehadi, A., Möller, S.: NISQA: A Deep CNN-Self-Attention Model for Multidimensional Speech Quality Prediction with Crowdsourced Datasets. In: INTERSPEECH 2021, pp. 1–5. ISCA (2021). <https://doi.org/10.21437/Interspeech.2021-299>
28. Graves, A., Fernández, S., Gomez, F., Schmidhuber, J.: Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In: ICML 2006, pp. 369–376. ACM (2006). <https://doi.org/10.1145/1143844.1143891>
29. Harper, E., Majumdar, S., Kuchaiev, O., et al.: NeMo: a toolkit for Conversational AI and Large Language Models. NVIDIA (2019). <https://nvidia.github.io/NeMo/>
30. Rekesch, D., Koluguri, N.R., Krizan, S., et al.: Fast Conformer with Linearly Scalable Attention for Efficient Speech Recognition. arXiv preprint arXiv:2305.05084 (2023)
31. Traer, J., McDermott, J.H.: Statistics of natural reverberation enable perceptual separation of sound and space. *Proceedings of the National Academy of Sciences* **113**(48), E7856–E7865 (2016). <https://doi.org/10.1073/pnas.1612524113>
32. Bouchakour, L., Debyeche, M.: Noise-robust speech recognition in mobile network based on convolution neural networks. *International Journal of Speech Technology* **25**(1), 1–9 (2022). <https://doi.org/10.1007/s10772-021-09950-9>
33. Duarte, J., Colcher, S.: Noise-Robust Automatic Speech Recognition: A Case Study for Communication Interference. *Journal on Interactive Systems* **15**(1), 670–681 (2024). <https://doi.org/10.5753/jis.2024.4267>
34. Chen, G., O’Shaughnessy, D., Tolba, H.: A performance investigation of noisy voice recognition over IP telephony networks. In: INTERSPEECH 2005, pp. 2681–2684. ISCA (2005). <https://doi.org/10.21437/Interspeech.2005-259>
35. Jordal, I., Tamazian, A., Dhyani, T., et al.: iver56/audiomentations: v0.39.0. Zenodo (2025). <https://doi.org/10.5281/zenodo.14856562>
36. Hollmann, N., Müller, S., Purucker, L., et al.: Accurate predictions on small data with a tabular foundation model. *Nature* **625**(7993), 1–9 (2025). <https://doi.org/10.1038/s41586-024-08328-6>
37. Hollmann, N., Müller, S., Eggensperger, K., Hutter, F.: TabPFN: A transformer that solves small tabular classification problems in a second. In: ICLR 2023 (2023)
38. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A.V., Gulin, A.: CatBoost: unbiased boosting with categorical features. arXiv preprint arXiv:1706.09516 (2019)
39. Vakhrushev, A., Ryzhkov, A., Savchenko, M., Simakov, D., Damdinov, R., Tuzhilin, A.: LightAutoML: AutoML Solution for a Large Financial Services Ecosystem. arXiv preprint arXiv:2109.01528 (2022)
40. Koenecke, A., et al.: Racial disparities in automated speech recognition. *PNAS* **117**(14), 7684–7689 (2020). <https://doi.org/10.1073/pnas.1915768117>