

ASR Performance Evaluation Post Speech Enhancement: WER Analysis and Tuning

Objective

Investigate why Word Error Rate (WER) increased after applying speech enhancement (NVIDIA and Own Model) compared to the original audios, and evaluate if tuning enhancement intensity can recover accuracy.

1. Data Sources

- Original audios: Baseline ASR output.
- NVIDIA enhanced: Same audio processed with NVIDIA speech enhancement (intensity=1.0).
- Own model enhanced: Same audios processed with a custom enhancement model.

Each transcript was aligned against ground-truth text, and word-level error breakdown (substitutions, insertions, deletions) was computed.

2. Key Findings

Condition	WER	Ins Rate	Del Rate	Sub Rate	Primary Issue
Original Audio	0.1786	0.0727	0.0204	0.0855	
NVIDIA Enhanced (1.0)	0.1913	0.0574	0.0255	0.1084	Substitutions + Deletions
Own Model Enhanced	0.1977	0.0714	0.0217	0.1046	Substitutions (Main Driver)

Main Contributors to WER Increase:

- Substitution errors dominate (the largest contributor in both enhancers).
- Diacritic confusion: č/š/ž → c/s/z.
- Morphology errors: Last-letter flips (e.g., dolivke → dolivka).
- Boundary effects: Increased first/last word mismatches.
- NVIDIA-specific: More deletions, suggesting dropped words.

3. Example Error Patterns

- Diacritic loss: “čas izpihovanja dolivke” → “cas izpihovanja dolivke”
- Morphology flip: “dolivke” → “dolivka”
- Boundary mismatch: Missing short words like “in”, “na” more often after enhancement.

4. Conclusion

Enhancement (at default intensity) is counterproductive for ASR accuracy. Own model causes more substitutions, NVIDIA adds deletions. Original audio consistently yields the best WER and CER before intensity tuning.

5. Recommendations

- Reduce enhancement aggressiveness to preserve phoneme detail.
- Use speech-preservation losses during training (SI-SNR, phoneme-consistency).
- Consider fine-tuning ASR on the enhanced data if the enhancement is necessary.
- Perform intensity sweeps to find the optimal trade-off.

6. Intensity Sweep & Optimisation Results (NVIDIA BNR)

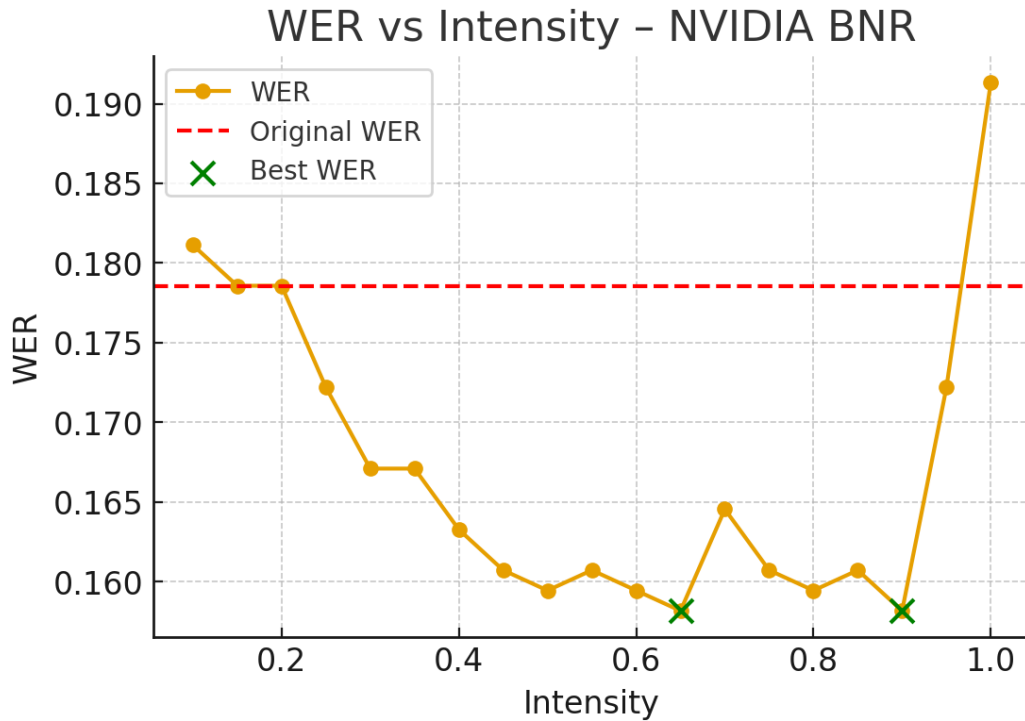
A sweep from 0.10–1.00 was performed using NVIDIA BNR. WER decreased steadily until ~0.65, then worsened beyond 0.70, with a sharp increase at 0.95–1.00.

Intensity	WER	Ins	Del	Sub	Δ WER vs Orig (%)
0.10	0.181122	0.072704	0.021684	0.086735	+1.43
0.15	0.178571	0.072704	0.020408	0.085459	+0.00
0.20	0.178571	0.072704	0.020408	0.085459	+0.00
0.25	0.172194	0.070153	0.020408	0.081633	-3.57
0.30	0.167092	0.070153	0.017857	0.079082	-6.43
0.35	0.167092	0.068878	0.017857	0.080357	-6.43
0.40	0.163265	0.066327	0.017857	0.079082	-8.57
0.45	0.160714	0.065051	0.019133	0.076531	-10.00
0.50	0.159439	0.063776	0.019133	0.076531	-10.71
0.55	0.160714	0.065051	0.019133	0.076531	-10.00
0.60	0.159439	0.065051	0.017857	0.076531	-10.71
0.65	0.158163	0.063776	0.017857	0.076531	-11.43
0.70	0.164541	0.063776	0.020408	0.080357	-7.86
0.75	0.160714	0.063776	0.020408	0.076531	-10.00
0.80	0.159439	0.062500	0.021684	0.075255	-10.71
0.85	0.160714	0.059949	0.022959	0.077806	-10.00
0.90	0.158163	0.057398	0.021684	0.079082	-11.43
0.95	0.172194	0.058673	0.020408	0.093112	-3.57
1.00	0.191327	0.057398	0.025510	0.108418	+7.14

Best WER = 0.158163 at intensities 0.65, 0.9. Choose 0.65 as the default: it offers the lowest Sub/Del, while slightly higher Ins is acceptable.

7. Residual Issues at 0.65

- Substitutions remain dominant (diacritic drops, morphology flips).
- Occasional insertion bursts in long sentences.
- Short phrases can still fail with high WER if misrecognized completely.



8. Focused Comparison: 0.65 vs 0.90

Intensity	WER	Ins	Del	Sub	Δ WER vs Orig (%)
0.65	0.158163	0.063776	0.017857	0.076531	-11.43
0.90	0.158163	0.057398	0.021684	0.079082	-11.43

Comparison shows 0.65 has slightly lower Sub/Del rates, while 0.90 mainly reduces Ins. Overall, WER is equal, so 0.65 is a safer default as it preserves word correctness better.

9. Detailed Error Analysis at Best Intensities

This section summarises the most frequent word-level substitution patterns observed at the best-performing intensities (0.65 and 0.90). These patterns explain the residual WER despite optimal enhancement.

Top Substitutions (0.65)

Ref Word	Hyp Word	Count
slajderju	sliderju	1
kateri	katerem	1
stroju	stroj	1
doziranje	zaziranje	1
kosov	kotcev	1
izkop	izklop	1

prebrisavati	prebrizgavati	1
spet	pip	1
vaga	vlaga	1
zalaufat	zalaufa	1
dolivke	dolivka	1
brizgalice	brizgance	1
y	xy	1
postrani	strani	1
dolivka	dolivke	1

Top Substitutions (0.90)

Ref Word	Hyp Word	Count
okvara	gora	2
slajderju	sliderju	1
kateri	katerem	1
stroju	stroj	1
doziranje	zaziranje	1
kosov	hodcev	1
izkop	izklop	1
prebrisavati	prebrizgavati	1
kosov	kotov	1
vaga	vlaga	1
zalaufat	zalaufa	1
zalaufat	laufati	1
dolivke	dolivka	1
brizgalice	brizgance	1
y	xy	1

Suffix Flips (End-Letter Changes)

Intensity	Pair (Ref→Hyp)	Count
0.65	dolivke → dolivka	1
0.65	dolivka → dolivke	1
0.65	okvara → okvare	1
0.90	dolivke → dolivka	1

10. Final Recommendation

Set NVIDIA BNR intensity to 0.65 as the default. Validate range 0.60–0.70 on held-out data. Mitigate residual errors by dictionary/LM biasing for critical words and diacritic-aware post-processing.