# To what extent do traditional ASR metrics like CER fail to predict downstream intent classification performanc

Assignee Research

June 10, 2026

**Abstract**

Sequence-to-sequence models, such as attention-based models in automatic speech recognition (ASR), are typically trained to optimize the cross-entropy criterion which corresponds to improving the log-likelihood of the data. However, system performance is usually measured in terms of word error rate (WER), not log-likelihood. Traditional ASR systems benefit from discriminative sequence training which optimizes criteria such as the state-level minimum Bayes risk (sMBR) which are more closely related to WER. In the present work, we explore techniques to train attention-based models to directly mi

## 1 Introduction

This paper examines: Minimum Word Error Rate Training for Attention-based Sequence-to-Sequence Models. Research question: To what extent do traditional ASR metrics like CER fail to predict downstream intent classification performance in noisy environments for large-scale speech foundation models?.

## 2 Methodology

Systematic literature search across multiple databases yielded 4 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.8/10.

## 3 Results

4 papers retrieved. 8 claims extracted; 0 independently verified. Quality review score: 3.8/10.

## 4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

## 5 Extracted Claims

| Claim | Verified | Confidence |
|---|---|---|
| Optimizing the sample-based approximation, LSample, reduces the expected number of word errors by $\sim$50% after training. | × | 0.12 |
| The WER for the top-hypothesis computed using beam search does not improve, but instead degrades as a result of training | × | 0.04 |
| Optimizing LN-best significantly improves WER by about 10.4% on the held-out portion of the training set. | × | 0.04 |
| Performance seems to be similar even when just the top four hypotheses are considered during the optimization of LN-best | × | 0.04 |
| It is important to also interpolate with CE loss function during optimization (i.e., setting $\lambda$ > 0). | × | 0.01 |
| After expected minimum WER training (MWER) of the uni-directional LAS model, the WER is 7.5% and after rescoring, it is | × | 0.10 |
| After expected minimum WER training (MWER) of the bi-directional LAS model, the WER is 7.2% and after rescoring, it is 6 | × | 0.07 |
| The CD-phone (CE + sMBR) system has a WER of 7.5% and after rescoring, it is 6.7%. | × | 0.02 |

## References


- http://arxiv.org/abs/1911.00566v2
- http://arxiv.org/abs/2304.00649v1
- http://arxiv.org/abs/1712.01818v1