Kamer Ali Yuksel (kamer@aixplain.com), **Thiago Castro Ferreira** (thiago@aixplain.com), **Ahmet Gunduz** (ahmet@aixplain.com),
Mohamed Al-Badrashiny (mohamed@aixplain.com), **Golara Javadi** (golara@aixplain.com)

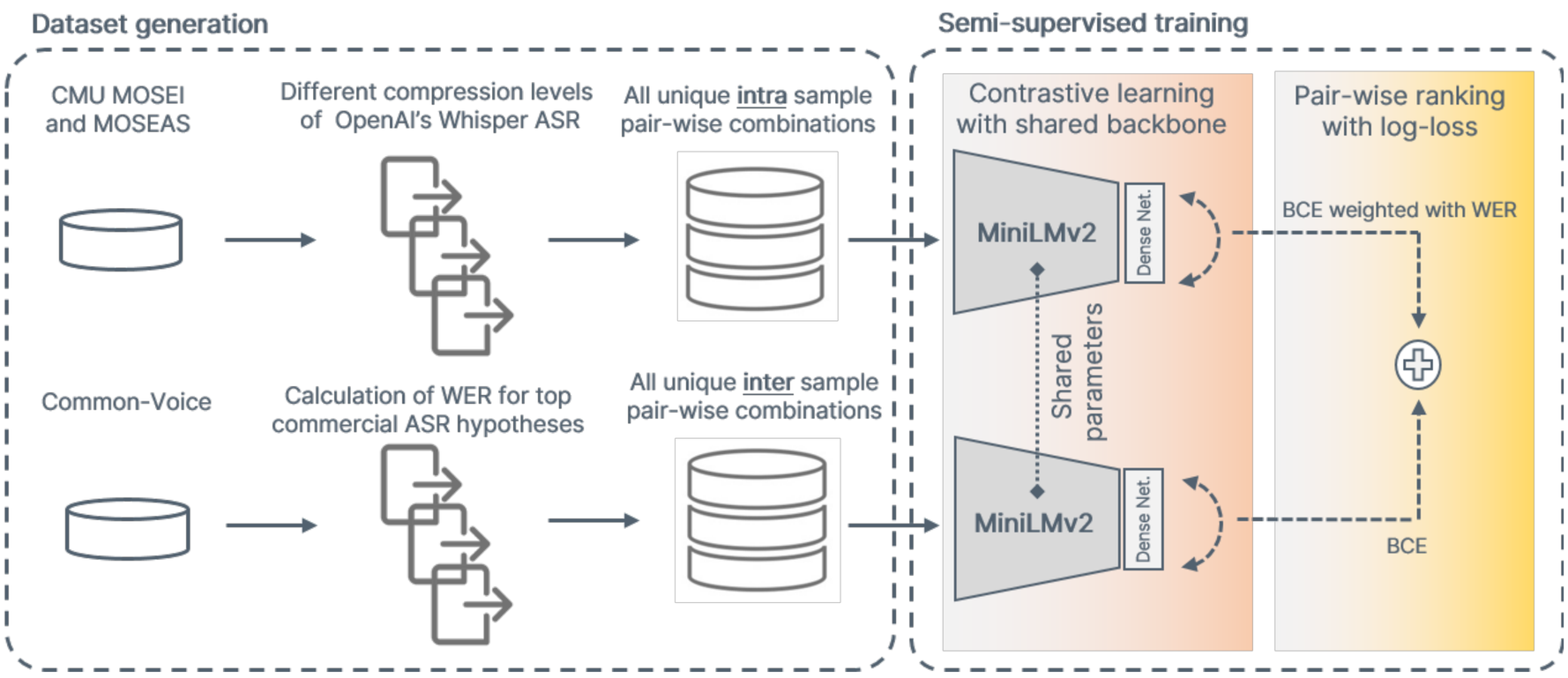
A Reference-less Quality Metric for Automatic Speech Recognition via Contrastive-Learning of a Multi-Language Model with Self-Supervision

INTRODUCTION

- Automatic Speech Recognition (ASR) systems are widely used, but evaluating their quality can be challenging due to the need for expensive ground-truth transcriptions.
- Traditional evaluation metrics such as Word Error Rate (WER) require ground-truth (reference) transcriptions, which are both time-consuming and expensive to obtain.
- We present a multi-language reference-less quality metric for ASR systems, which allows comparing ASR models without the need for reference transcriptions.

MOTIVATION

- Reference-based metrics like Word Error Rate (WER) are commonly used but have limitations such as requiring time-consuming and costly ground-truth transcriptions.
- Previous studies have used various approaches for ASR quality estimation. However, none of them were based solely on language features or trained without references.
- Recent efforts have focused on training regression or ordinal classification models for ASR quality estimation using both audio-input and text-output features that are available.
- Our motivation was to develop a reference-less metric that leverages only text features from the transcription output for faster and more cost-effective ASR quality estimation.
- Our work drew inspiration from referenceless quality estimation metrics in Machine Translation (WMT Metrics Shared Task), paving the way for a novel approach to ASR.

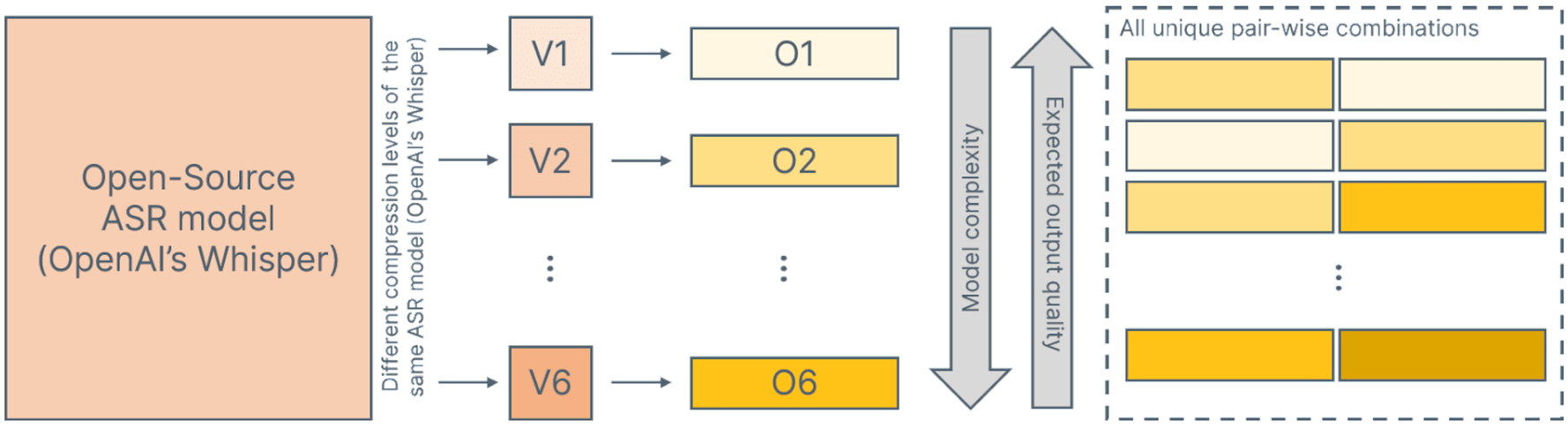


NoRefER fine-tunes a pre-trained language model (MiniLMv2) in a self-supervised learning manner with contrastive learning.

METHODOLOGY

- Our proposed method utilizes contrastive learning and self-supervision to fine-tune a pre-trained multi-language model using a Siamese network for pair-wise ranking.
- Contrastive learning with self-supervision helps LM learn discriminative representations and does not need reference transcriptions by training on ASR hypothesis pairs.
- Unique outputs from the Whisper model in multiple compression levels are used to form pairwise combinations as a training dataset consisting of ASR hypothesis pairs.

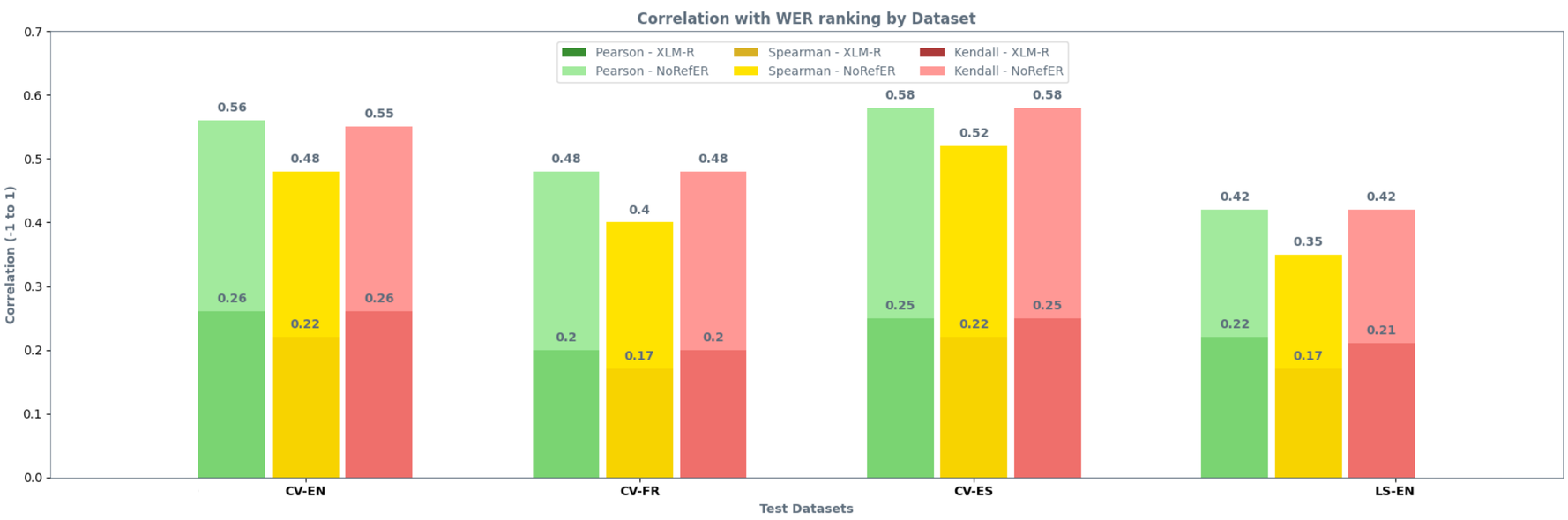
- The compression level of the Whisper model serves as a proxy for quality, with higher levels indicating lower-quality transcriptions, and used as labels for binary classification.
- Our reference-less metric leverages a pre-trained cross-lingual language model (MiniLMv2) in a Siamese network to enable binary classification of pair-wise quality.
- We employ a dense encoder to reduce language model embeddings of both pairs to scalar logits subtracted from each other for facilitating the pairwise quality ranking.



NoRefER fine-tunes a pre-trained language model (MiniLMv2) in a self-supervised learning manner with contrastive learning.

EXPERIMENTS

- We conducted extensive experiments to evaluate the performance of our reference-less quality metric. The metric was trained and validated on a large corpus of ASR outputs.
- The referenceless metric is trained and validated on a large corpus of ASR outputs. It achieves high validation accuracy in pair-wise ranking for reliable quality comparisons.
- The metric is blind-tested on multiple speech datasets and compared with a perplexity baseline, which it consistently outperformed on different datasets and languages.
- When the metric is used to pick the highest quality among hypotheses of ASR engines, the ensemble transcriptions reduced the best-performing engine's WER by +7%.



The correlation with WER scores and rankings against XLM-RoBERTa perplexity in Common Voice and Libri-Speech datasets.

CONCLUSION

- We have introduced a novel referenceless quality metric for ASR systems, which is based on contrastive learning of a multi-language model with self-supervision.
- The metric eliminates the need for references, reducing evaluation costs and time. Potential applications include ensembling ASR engines for improved quality.
- The metric's reliability and applicability are demonstrated through correlation coefficients with Word Error Rate (WER) ranks and scores on different datasets and languages.
- The proposed metric outperforms traditional metrics and shows potential for ensembling ASR engines. It provides reliable quality comparisons without requiring ground truths.