Primal-OWSM: Speech Foundation Model with Parameter-efficient Primal Attention for Low-resource Dutch Speech Recognition

Pu Wang[0000−0001−8725−6225] and Hugo Van hamme[0000−0003−1331−5186]

KU Leuven, Department of Electrical Engineering-EAST,
Kasteelpark Arenberg 10, 3001, Belgium
{pu.wang, hugo.vanhamme}@esat.kuleuven.be

Abstract. Large-scale transformer-based language models, such as ChatGPT from OpenAI, have revolutionized the natural language processing (NLP) landscape. These models are pre-trained on massive, industry-scale datasets, allowing them to perform remarkably well across diverse downstream tasks, often with minimal fine-tuning. This success has also extended to the field of speech processing. For example, OpenAI's Whisper, released in 2022, is an automatic speech recognition (ASR) system that has achieved impressive results. Whisper is trained on a large web-sourced dataset, totaling approximately 680,000 hours of data across more than 96 languages. This extensive multilingual supervised training enables Whisper to generalize across languages.

Though according to OpenAI, Whisper's English ASR performance reaches human-level accuracy and robustness, researchers have observed frequent spelling mistakes in transcriptions of non-English audio. For instance, our results show that the zero-shot ASR performance of the Whisper-small model on the Spoken Dutch Corpus (CGN) is less satisfactory, with a word error rate (WER) above 30%. While its performance can be enhanced through fine-tuning on specific low-resource language corpora, studies have shown that Whisper tends to overfit quickly on small datasets. Consequently, fine-tuning with limited data can result in worse performance compared to the zero-shot capabilities of the pre-trained model. For example, our results show that the Whisper-large model achieves a 19% WER on CGN, but after fine-tuning with the Dutch part of Multilingual Librispeech (MLS), its WER increases to 21%. This issue is particularly pronounced in low-resource languages, where publicly available corpora are either small or lack diversity, leading to poor generalization.

Since the comprehensive pipeline for Whisper development (from data preparation to training) remains inaccessible to the public, it is challenging for researchers to understand the underlying mechanisms, address the robustness issues, and find effective methods to improve the model's performance. To promote transparency, WAVLab at Carnegie Mellon University has reproduced a Whisper-style training process using an open-source toolkit and publicly available data, releasing the Open

Whisper-style Speech Model (OWSM). Our study builds on the OWSM foundation model, focusing on efficient fine-tuning for low-resource Dutch ASR. Specifically, we modify OWSM by replacing the $Softmax$ self-attention mechanism with primal attention to reduce the computations, implemented using a primal representation of singular value decomposition (SVD) on the $Softmax$ attention. The modified model, named Primal-OWSM, is configured with three model scales similar to Whisper: small (272M parameters, comparable to Whisper-small), medium (712M, comparable to Whisper-medium), and large (889M, half the size of Whisper-large). The implementation of Primal-OWSM is openly accessible at `https://github.com/wangpuup/Primal-OWSM.git` and can be easily combined with other speech foundation models or parameter-efficient fine-tune algorithms for future studies.

Experiments with fine-tuning on CGN Dutch and Flemish Dutch data show that Primal-OWSM offers reduced memory usage and lower training complexity compared to the original OWSM. On the same device (a single NVIDIA A100 SXM4 80GB GPU), Primal-OWSM reduces training memory usage by approximately 40% and training time per epoch by 10%, while outperforming OWSM with the same number of training epochs. Whether fine-tuning on a sufficient training scenario (236 hours of training data) or a limited training scenario (10 hours of training data), Primal-OWSM consistently converges significantly faster and achieves a lower WER with fewer training epochs (or training hours). For example, after 4 to 5 epochs of training on 236 hours of data, Primal-OWSM (large) reduces the WER to around 15%, whereas OWSM requires 3 additional training epochs (approximately 16 additional hours, or 60% more training hours) to reach comparable performance. It should be noted that the zero-shot WER of OWSM (large or v3) on CGN is 52.44%. The Primal-OWSM speech foundation model for Dutch is available on Huggingface `https://huggingface.co/wangpuupup/primal_owsm_large`. It is zero-shot evaluated on Common Voice-Dutch and MLS-Dutch to assess its generalization capabilities. Additionally, Primal-OWSM is compared with the parameter-efficient fine-tuning (PEFT) method LoRA applied to the Whisper model.

Keywords: Speech foundation model · Primal attention · Whisper · OWSM · Parameter-efficient training · Dutch speech recognition.

Acknowledgments. The research was supported by the Flemish Government under "Onderzoeksprogramma AI Vlaanderen" and FWO-SBO grant S004923N: NELF.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.