ENHANCING SPEECH RECOGNITION WITH LLMS IN POST-CORRECTION SETTINGS

Anonymous authors
Paper under double-blind review

ABSTRACT

The rapid development of Automatic Speech Recognition (ASR) systems in audio transcription tasks to get text content. However, even State-of-the-Art systems do not always provide excellent results and can make mistakes, especially in the new speech domain. To address this problem, developers either fine-tune this system on specific data to adapt the ASR model to their domain or incorporate Language Models, which gained success in Natural Language understanding to the overall prediction re-scoring. In this work, we decided to improve the quality of transcriptions in a post-correction setting, fine-tuning the external Large Language Model (LLM) without tuning the ASR system. We demonstrated that this approach is prominent, and one fine-tuned LLM improves the results of different ASR models. We significantly enhanced the quality metrics compared to the baselines and competitors.

1 INTRODUCTION

The fast-paced development of deep learning affected different parts of human life. For example, advances in deep learning models for sound processing allowed the creation of ASR (automatic speech recognition) systems with superhuman performance on different benchmarks (transcribing audio better than a person). Such transcribed text can be used in downstream tasks like call center automation, text summarization, etc. Nevertheless, the place for research and transcription quality enhancement still exists.

Top-level ASR quality improvement approaches can be divided into a few main categories. The first is based on scaling the ASR model size and train data volume and the train model in a multitask mode (Radford et al., 2023). Despite the solid relative improvement on some benchmarks (See Common Voice (Ardila et al., 2019), CHiME6 (Watanabe et al., 2020), WSJ (Paul & Baker, 1992), etc), this approach clearly has some limitations related to the model size. It could be computationally expensive to fine-tune such a model on a particular domain. Also, it could be difficult to label paired audio-text data for the supervised fine-tuning (SFT) of a downstream task. Self-supervised learning for ASR models were introduced to deal with the latter problem.

The idea of SSL in ASR is quite straightforward - the model is trained to compress input data to a meaningful hidden state, and the training procedure is quite similar to the masked language modeling (MLM) task in BERT (Devlin et al., 2018). Such an approach allows us to get more or less acceptable transcription quality with a small dataset size for SFT (about 10 minutes) (Baevski et al., 2020). Nevertheless, such an approach is also computationally expensive because it requires a lot of unlabeled train data (about 1 billion hours, for example).

Another approach that is widely used in ASR production systems is the rescoring technique. There are two main types of such approach: the first one is called first-pass rescoring (Toshniwal et al., 2018). In this method, a Language model is used during hypothesis decoding. Given the hidden state from the encoder

of the ASR model, the possible tokens distribution is computed, and then the distribution is blended with the LM distribution. Then, the token is picked from the distribution obtained in the previous step. Another approach is second-pass rescoring, and it is based on rearranging the transcribed N hypothesis (See Chiu et al. (2018)). The idea is to take the top N hypothesis after beam-search decoding and then rearrange them according to acoustic and language model probability. It allows for the incorporation of information not only from the ASR model but also from the language model, which is usually trained on a huge number of texts and thus has better language understanding. Both approaches are used on inference time and pre-trained LMs, because text data is widely available for model pre-train, so enhancing ASR system quality by incorporating knowledge about language through LM, is much easier.

Also, one of the last trends in Large Language Models (LLM) is to add the audio domain to use a well-trained language domain and audio capturing with the encoder in the audio transcription task (See et al. (2024)). Linguistic and audio domains in pairs can lead to better error correction for transcription from audio. This approach also leads to more data for model training and can lead to better performance of LLM.

However, these approaches have potential disadvantages. Firstly, one could hurt performance on non-domain-specific tasks in ASR. Secondly, one can lead that LLM could not capture some specific details from the audio encoder, as it happens with visual parts (See Rahmanzadehgervi et al. (2024)). Our idea is to use the ASR model with great audio capturing and LLM with great language understanding and combine them. For that purpose, LLM will be tuned to understand the errors the ASR model is prone to. To give the LLM model better context about audio, we will provide it with not only one ASR transcription but 5 hypotheses. This could lead to a lower Word Error Rate (WER) cause sometimes hypotheses 2-5 can have better transcription. WER has a formula $\frac{S+D+I}{S+D+C}$, where S is a number of substituted, D – deleted, I – inserted words, C – correct words in prediction. Also, we applied not classic fine-tuning but rather Low-Rank Adaptation (LoRA, Hu et al. (2021)), more details in Section 4.1.3.

Our key contributions are the following:

- We significantly improved the quality of transcription compared to the ASR. Namely, WER reduced by up to 80 % on some data and 14 % on average;
- We demonstrated that our models can correct errors for various audio sources and different ASR models, even for those that were not in the training;
- We demonstrated that the NEFTune regularization technique (Jain et al., 2023) is useful in our setting with encoder-decoder architecture. Also, we show that even relatively small LLMs are strong correctors.

2 Related Work

In this section, we discuss related work and approaches relevant to understanding the topic of this work, such as main ASR and LLM models and previous post-correction approaches.

2.1 ASR models

Most ASR models utilize Mel-Spectrograms or MFCC as input features instead of raw waveform. Like Deep Speech (Hannun et al., 2014; Amodei et al., 2016), CTC-based approaches introduced Connectionist Temporal Classification (CTC) loss to align input speech sequences with output text without requiring per-letter or per-phoneme pre-segmented data, simplifying training and forcing the model to learn the optimal alignment between speech frames and text transcriptions. During inference, the model maintains several top hypotheses over paths in beam search. Despite the simplicity and performance of such approach it has the disadvantage: CTC decoding is context independent so it could possible fail on word-to-word matching in a

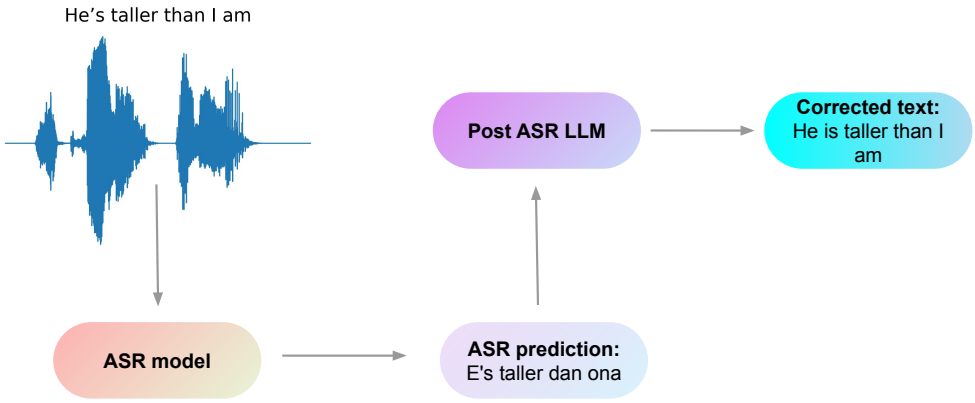

Figure 1: Overview of the proposed method during inference.

hypothesis. Listen, Attend and Spell (LAS) (Chan et al., 2016) uses an encoder-decoder architecture where the encoder processes the input speech signal, and the attention mechanism allows the decoder to focus on different parts of the input sequence dynamically. The Conformer model Gulati et al. (2020) integrates convolutional neural networks (CNNs) with Transformer architectures, enabling capturing local features through convolution and long-range dependencies through self-attention.

Whisper (Radford et al., 2023) leverages a transformer-based architecture optimized in a weak-supervised regime on a colossal amount of data. It focuses on robustness and scalability to different languages and speech domains, making it adaptable to diverse datasets and conditions. However, it does not demonstrate the best results on common benchmarks and hallucinates if the audio contains a lot of silence, as it misleads attention. Wav2Vec2 (Schneider et al., 2019) employs a self-supervised learning approach to pre-train the model on unlabeled speech data. This model uses contrastive learning to understand audio representation as in masked language modeling.

2.2 ROVER

Recognizer Output Voting Error Reduction (ROVER) (Fiscus, 1997) is a method that reduces WER by aligning hypotheses with a voting mechanism. It operates by first generating a Word Transition Network (WTN), a directed acyclic graph representing all possible word sequences from the input hypotheses. Each node in the WTN corresponds to a word, and edges represent transitions between words with associated probabilities. ROVER then employs dynamic programming to align these sequences. It uses a voting scheme to select the most likely word at each position based on frequency counts or confidence scores from the individual recognizers. This method can improve ASR quality, leveraging multiple outputs.

2.3 HyPoradise

HyPoradise Chen et al. (2023) is a dataset for post-ASR correction with LLM and also the open baseline for this task. It consists of several popular audio datasets listed in Table 1.

The audio was transcribed with Whisper-Large ASR system with BeamSearch and divided into training and test parts. In this dataset, in both parts, there are examples of different errors provided by ASR: **insertion** (ASR: "various sizes to you", Ground-Truth (GT): "various sizes"); **deletion** (ASR: "this was a great", GT:

Table 1: Description of various datasets.

Dataset	Description
ATIS Hemphill et al. (1990)	Airline travel information
CHiME4 Vincent et al. (2016)	Audio with some background noise
CORAAL Farrington & Kendall (2021)	Interviews with speakers born between 1888 and 2005
Common Voice	Crowdsourced dataset from different speakers collected via Internet
LRS2 Son Chung et al. (2017)	BBC television recordings
LibriSpeech Panayotov et al. (2015)	Audiobooks
SwithBoard Godfrey et al. (1992)	Telephone speech corpus
TED-LIUM 3 Hernandez et al. (2018)	TED talks
WSJ	Reading of Wall Street Journals

"this was great"); **consonant** (ASR: "additionally", GT: "traditionally"); **hallucinations** due to background noise (ASR: "sorry sorry sorry", GT: "despite the decline in stock prices").

The main idea of the dataset is that the choice of ASR system is not always the most accurate. In the CHiME4 test split, only half the time, the top 1 had the lowest WER among all 5 hypotheses. Also, sometimes, "parts" of an "ideal" sentence can be represented in several hypotheses. So, we move from the task of re-ranking to the task of generation. See Figure 2 for CHiME4 test results.

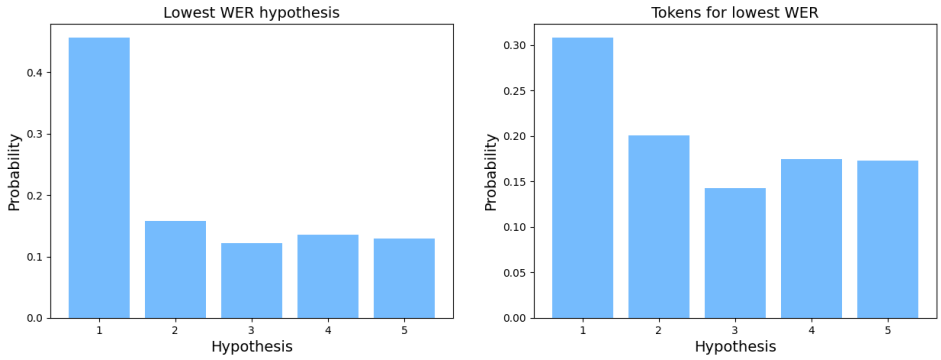

Figure 2: Distribution of hypothesis by lowest WER.

2.4 LLMs

Recent advancements in LLMs have revolutionized natural language processing (NLP). BERT (Devlin et al., 2018) employs a bidirectional transformer to pre-train on masked language modeling (MLM) and next-sentence prediction, capturing context from both directions. GPTs series uses a unidirectional transformer decoder. GPT-2, with 1.5 billion parameters, excels in generative tasks. GPT-3 scales up to 175 billion parameters for improved performance. GPT-4 enhances alignment with user intent using fine-tuning and reinforcement learning from human feedback (RLHF). Low-rank adaptation (LoRA) reduces trainable parameters in large models by decomposing learnable parameters into low-rank factors, cutting computational costs and memory usage while maintaining performance.

One way to correct hypotheses from ASR systems is to use proprietary LLM with a great linguistic domain such as ChatGPT (Ma et al., 2023). They tested OpenAI's model with several datasets, like LibriSpeech TED-LIUM 3 and Artie Bias Corpus (Meyer et al., 2020), and they investigated such ideas as hypothesis generation and selection, number of provided hypotheses and zero-shot and one-shot scenarios. They achieved great results and discussed several topics related to ChatGPT performance. However, this approach is limited to using the OpenAI API.

Another straightforward approach is to fine-tune the causal or sequence-to-sequence LLM. In Chen et al. (2023), the authors tuned T5-0.75B (Raffel et al., 2020) and Llama-13B (et al., 2023) LLMs for their task and measured WER on their dataset to show the ability of models to perform such tasks. However, they fine-tuned each model to each dataset separately, which, as was said, could lead to reduced generalization and decreased accuracy. Also, their Whisper as a Baseline results do not correlate with the WER values obtained on these datasets, so they will not be included in the comparing table.

3 METHODOLOGY

The pipeline consists of additional data collection, diverse audio and transcription augmentation, and further fine-tuning of language models with different hyperparameters. We considered the top-1 ASR model's hypothesis for the baselines and ROVER over 5 hypotheses. We used ROVER to align all hypotheses with the top 1 based on the idea that the top 1 hypothesis is the most accurate, so the hypothesis with the lowest possible WER will also be close to the top 1. We also compared our result with Llama3-8B untuned for this correction task.

The ASR system takes audio as an input and performs all the pre-processing steps that vary from model to model. Then, with BeamSearch, ASR generates 5 hypotheses from this audio. This text is post-processed to the required letter case and style (e.g., without commas). To collect additional data that helps to improve generalization, we also used the base version of Whisper instead of Large (See section 4.1.1 for more details). The Whisper model was fully frozen; we did not tune any parameters for our task.

We utilized the T5, which is an encoding-decoding transformer architecture. The idea behind this model is that the encoder gathers all information from instruction and hypotheses, transmits it through cross-attention to the decoder part, and the decoder generates his own hypothesis. We removed special tokens from LLM's hypothesis at the post-processing stage. The T5 model was fine-tuned on the Flan (Wei et al., 2021) dataset, leading to better generalization for different tasks, which could help the model better understand what kind of correction we want to get. In addition, we used few-shot prompting with 3 examples from our validation dataset, so results from it won't be included.

4 EXPERIMENTS

4.1 DATASETS

We used HyPoradise with training and test splits made by authors, except LibriSpeech-clean and WSJ from the test, which were used as validation tests. Also, we added to train Multi-Speaker Corpora of English native speakers (Demirsahin et al., 2020). It was transcribed using the Whisper-Base model to reduce transcription quality. Audio was loaded, padded, and trimmed, then converted to Log Mel Spectrogram. This spectrogram was fed into Whisper and decoded. We took the top 5 hypotheses from the decoded text and normalized them; all numbers were converted to strings, i.e., "42" was converted to "forty-two."

Part of the audio transcriptions were augmented. This was used to simulate common 2 types of errors that occur during audio transcription: random deletion of a word and random insertion. Insertion was made using context BERT embeddings with NLPAug (Ma, 2019) library. This augmentation gave the model more

examples that sometimes not all information is included in only one hypothesis and can be located in several hypotheses.

We considered the Mozilla Common Voice 18 (CV) English part for the second branch of experiments. We took 900000 audios for training and validation. Each audio was randomly augmented with several transformations: voice activity detection, Gaussian noise, room impulse response (RIR), gain, polarity inversion, pitch shift, low-pass filtering, color noise, and audio pre-emphasizing. Pre-emphasis is defined as follows $x_l = x_l - \gamma \cdot x_{l-1}$, where $l = 1, L$ is an amplitude index of a given audio of length L, and γ is the pre-emphasis factor (we used $\gamma = 0.97$). Every augmentation was applied independently. Augmented train audios were transcribed with Whisper tiny, base, and medium. To better evaluate the generalization power of our model, the non-augmented test audios (CV18 test set) were transcribed with Whisper small, large, and Wav2Vec2 base models. We provide LLM only top-1 ASR hypothesis in this branch of experiments.

4.2 Models

As stated in Section 3.2, we took Flan-T5 due to its instruction performance. The model was represented in the XXL (11B) variant with few-shot prompts of pairs hypotheses-GT from the validation dataset to give the model an example of how to correct these hypotheses. We compared its performance with Whisper-Large as the baseline (top-1 hypothesis), ROVER, and Llama3 8B with the same 3 shots we used with our model to show the performance of untuned SOTA LLM with a great linguistic domain. For experiments with CV, we fine-tuned Qwen2-1.5B with LoRA in a non-instruct setting.

4.3 Training and Inference

For fine-tuning, we used LoRA with several combinations of layers as a target for efficient fine-tuning our model for correction tasks with lower resource usage. All weights were initialized randomly due to the bad performance of PiSSA initialization. One more thing we used with LoRA was NEFTune regularization with alpha 0.1. It added some random noise to embeddings that can lead to better instruction execution, as it did with Llama2 on the Alpaca dataset in the original paper. NEFTune α value adjustment was performed on the same validation set, and a subset of the training set was used for model fine-tuning. α values were 0.01, 0.05 and 0.1. After testing the model with α equal to 0.1, we gained the lowest WER, so it was used with our model and "Flan-T5 (ours)" and "Full+NEFTune" in Table 3 and Table 4 respectively.

All code was executed with PyTorch 2.1.2 Hugging Face Transformers 4.42.4 (Wolf et al., 2020) and Parameter-Efficient Fine-Tuning 0.11.1 Mangrulkar et al. (2022) libraries. Optimizer AdamW (Loshchilov & Hutter, 2017) with lr 1e-4 and warm-up 0.1 ratio, fused implementation. Forward pass was mixed-precision (Micikevicius et al., 2017) with bfloat16 type (Kalamkar et al., 2019), float16 was unstable. Number of train epochs was set to 1. The batch size for both training and validation was set to 16. Inference BeamSearch has 6 beams, temperature 0.8, top_k 40, top_p 0.75. The maximum of new tokens parameter was limited to 224. The hardware we used was two NVIDIA H100 GPUs with CUDA 11.8.

5 Results and Discussion

This section presents our results: compare our model with Whisper's top 1 hypothesis, ROVER, Llama3, and demonstrate the result on the CV dataset. See Tables 2 and 3.

As you can see, our model outperforms Whisper on all test datasets except LibriSpeech-other. The greatest improvement we gained with the ATIS dataset could be the cause of the dataset's nature, such as Whisper translating mainly audio without splitting some place names into parts. For example, Whisper writes "washington dc" when the GT variant is "washington d c," and there are a lot of geographical places in this dataset. The second dataset by WER improvement is CHiME4. This one consists of audio with different background

Table 2: WER on Common Voice EN test set for different ASR models with and without post-correction.

LLM \ASR	Whisper-Small	Whisper-Large	Wav2Vec2-Base
No LLM	30.1	28.9	43.9
Qwen2-1.5B	19.8	18.5	30.5
Qwen2-0.5B	21.6	18.5	32.7

Table 3: WER (%) results of different models comparing with Whisper.

Test Set	Whisper	ROVER	Llama3	Flan-T5 (ours)
ATIS	8.4	$9.1_{+8.33\%}$	$8.32_{-0.1\%}$	$\mathbf{1.58}_{-81.19\%}$
CHiME4	12.05	$12.48_{+3.57\%}$	$44.73_{+257.84\%}$	$\mathbf{4.84}_{-59.83\%}$
CORAAL	24.38	$24.08_{-1.23\%}$	$94.99_{+287.57\%}$	$\mathbf{22.47}_{-7.83\%}$
CV	15.72	$16.38_{+4.2\%}$	$23.24_{+47.83\%}$	$\mathbf{10.85}_{-30.98\%}$
LRS2	12.89	$13.43_{+4.19\%}$	$261.86_{+1931.5\%}$	$\mathbf{8.72}_{-32.35\%}$
LS-other	**5.15**	$9.84_{+91.07\%}$	$6.64_{+28.93\%}$	$11.37_{+120.78\%}$
SWBD	17.03	$17.31_{+1.64\%}$	$196.6_{+1054.43\%}$	$\mathbf{14.94}_{-12.27\%}$
TD-3	4.77	$4.89_{+2.52\%}$	$334.19_{+6906.8\%}$	$\mathbf{4.54}_{-4.82\%}$

noises, which sometimes can lead to inaccuracy or even hallucinations, but it seems that 2 to 5 hypotheses still provide enough information for LLM to perform error correction. Common Voice dataset has a lot of speakers with different pronunciations, and improvement could be a possible cause of an additional dataset of various speakers from different parts of England that we used for training, which could show our model more variants of Whisper failure caused by pronunciation. Llama3, without fine-tuning, mostly failed with this task, also like a ROVER. Overall, our model performs well on all test datasets representing different sources, like books, phone calls, and others from HyPoradise. Although we did not include HyPoradise's results due to a bad correlation between Whisper-Whisper results, our result in 5 out of 8 datasets showed better WER decrease as a percentage of the baseline in comparison.

Here, "QV" and "QKVO" refer to attention modules we fine-tune, and "Full" refers to all linear modules, i.e., attention and post-attention "T5LayerFF" parts. So "Full" can be interpreted as "QKVO + wi_0 + wi_1 + wo". One can see that tuning different modules can sometimes lead to decreased and increased WER on different datasets, like QKVO, compared to QV's reduced WER on ATIS but increased it on LibriSpeech-other. Overall, "Full+NEFTune" performs best even though embeddings with LoRA are frozen during training. So "Full+NEFTune" we used it as "Flan-T5 (ours)" in Table 3.

We compare the results of experiments with different target modules for LoRA in Table 4 and present speech recognition and correction examples in Table 5. However, we should admit that our study has several limitations. The LLM model heavily depends on relatively good hypotheses produced by ASR models because these hypotheses are the only source for the language model to perform post-correction. If the ASR model has serious hallucinations due to loud noise or silence, the model cannot "repair" information. The language model could hallucinate, either. In addition, if train and test domains differ significantly (e.g. average train sentence length is much smaller than in test, such as CV and CORALL), the results might be

Table 4: WER (%) results of different target modules and NEFTune with Whisper.

Test Set	Whisper	QV	QKVO	Full	Full+NEFTune
ATIS	8.4	$2.72_{-67.62\%}$	$1.67_{-80.12\%}$	$1.75_{-79.17\%}$	$\mathbf{1.58_{-81.19\%}}$
CHiME4	12.05	$6.21_{-48.46\%}$	$5.32_{-55.85\%}$	$6_{-50.21\%}$	$\mathbf{4.84_{-59.83\%}}$
CORAAL	24.38	$23.65_{-2.99\%}$	$27.56_{+13.04\%}$	$23.6_{-3.2\%}$	$\mathbf{22.47_{-7.83\%}}$
CV	15.72	$11.4_{-27.48\%}$	$11.39_{-27.54\%}$	$11.01_{-29.96\%}$	$\mathbf{10.85_{-30.92\%}}$
LRS2	12.89	$10.18_{-21.02\%}$	$10.16_{-21.18\%}$	$9.93_{-22.96\%}$	$\mathbf{8.72_{-32.35\%}}$
LS-other	**5.15**	$5.29_{+2.72\%}$	$5.3_{+2.91\%}$	$5.36_{+4.08\%}$	$11.37_{+120.78\%}$
SWBD	17.03	$15.45_{-9.28\%}$	$\mathbf{14.75_{-13.39\%}}$	$15.24_{-10.51\%}$	$14.94_{-12.27\%}$
TD-3	4.77	$\mathbf{4.03_{-15.51\%}}$	$4.09_{-14.26\%}$	$4.2_{-11.95\%}$	$4.54_{-4.82\%}$

negative (Table 6). But despite all of that, overall WER reduction leads to more accurate translations in most cases regardless of the audio source.

Table 5: Examples of ASR post-correction on CV18 test set. LLM is QWEN2-1.5B Bai et al. (2023)

ASR Model	GT	ASR prediction	LLM correction	WER ASR	WER LLM
Whisper-Base	the chersky range is part of the south siberian system	the cheer sky range is part of the south cerebrian system	the chersky range is part of the south siberian system	30.0	20.0
Whisper-Base	interment in the woodlands cemetery	intermined in the woodland symmetry	interment in the woodland cemetery	60.0	20.0
Whisper-Base	basil of annonay france	basil of annoyed france	basil of annonay france	25.0	0.0
Whisper-Base	bundesliga where he played for two seasons	bunder siglor where he played for two seasons	bundesliga where he played for two seasons	28.6	0.0
Whisper-Large	he was born in pittsburgh	hän oli nainen pittsburghissa	he had been in pittsburgh pennsylvania	100.0	60.0
Wav2Vec2	its stigmas are bilobed	tstigmas are billo bed	its stigmas are bilobed	100.0	0.0
Wav2Vec2	flett played rugby union for edinburgh university	lit play rug beunon far edinburg universipi	flett played rugby union for edinburgh university	100.0	0.0
Wav2Vec2	i am sure of it	am shored offit	i am sure of it	80.0	20.0
Wav2Vec2	the area is roughly triangular	he areas rochly criangular	the area is roughly craggy	100.0	20.0

6 Conclusion and Future Work

In this article, we tackled the problem of ASR post-correction. We proposed using the Flan-T5 and Qwen2 models, fine-tuning them on various datasets with various waveform and text-level augmentations. Our approach successfully handles errors and enhances transcription obtained from diverse sources distinct from train speech recognition models, demonstrating prominent and robust results. Future work might be devoted to improving quality and language understanding results, expanding the procedure in a multilingual setting, and model distillation to be suitable for low-resource applications.

REFERENCES

Dario Amodei et al. Deep speech 2: End-to-end speech recognition in english and mandarin. In *International conference on machine learning*, pp. 173–182. PMLR, 2016.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber. Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*, 2019.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020.

Jinze Bai et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

William Chan, Navdeep Jaitly, Quoc Le, and Oriol Vinyals. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 4960–4964. IEEE, 2016.

Chen Chen et al. Hyporadise: An open baseline for generative speech recognition with large language models. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023.

Chung-Cheng Chiu, Tara N Sainath, Yonghui Wu, Rohit Prabhavalkar, Patrick Nguyen, Zhifeng Chen, Anjuli Kannan, Ron J Weiss, Kanishka Rao, Ekaterina Gonina, et al. State-of-the-art speech recognition with sequence-to-sequence models. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 4774–4778. IEEE, 2018.

Isin Demirsahin, Oddur Kjartansson, Alexander Gutkin, and Clara Rivera. Open-source Multi-speaker Corpora of the English Accents in the British Isles. In *Proceedings of The 12th Language Resources and Evaluation Conference (LREC)*, pp. 6532–6541, Marseille, France, May 2020. European Language Resources Association (ELRA). ISBN 979-10-95546-34-4. URL https://www.aclweb.org/anthology/2020.lrec-1.804.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Abhimanyu Dubey et al. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Hugo Touvron et al. Llama 2: Open foundation and fine-tuned chat models, 2023. URL https://arxiv.org/abs/2307.09288.

Charlie Farrington and Tyler Kendall. The corpus of regional african american language. 2021.

J.G. Fiscus. A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (rover). In *1997 IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings*, pp. 347–354, 1997. doi: 10.1109/ASRU.1997.659110.

John J Godfrey, Edward C Holliman, and Jane McDaniel. Switchboard: Telephone speech corpus for research and development. In *Acoustics, speech, and signal processing, ieee international conference on*, volume 1, pp. 517–520. IEEE Computer Society, 1992.

Anmol Gulati et al. Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100*, 2020.

Awni Hannun et al. Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*, 2014.

Charles T Hemphill, John J Godfrey, and George R Doddington. The atis spoken language systems pilot corpus. In *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27, 1990*, 1990.

François Hernandez et al. Ted-lium 3: Twice as much data and corpus repartition for experiments on speaker adaptation. In *Speech and Computer: 20th International Conference, SPECOM 2018, Leipzig, Germany, September 18–22, 2018, Proceedings 20*, pp. 198–208. Springer, 2018.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Neel Jain et al. Neftune: Noisy embeddings improve instruction finetuning. *arXiv preprint arXiv:2310.05914*, 2023.

Dhiraj Kalamkar et al. A study of bfloat16 for deep learning training. *arXiv preprint arXiv:1905.12322*, 2019.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Edward Ma. Nlp augmentation. https://github.com/makcedward/nlpaug, 2019.

Rao Ma, Mengjie Qian, Potsawee Manakul, Mark Gales, and Kate Knill. Can generative large language models perform asr error correction? *arXiv preprint arXiv:2307.04172*, 2023.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. Peft: State-of-the-art parameter-efficient fine-tuning methods. `https://github.com/huggingface/peft`, 2022.

Josh Meyer, Lindy Rauchenstein, Joshua D Eisenberg, and Nicholas Howell. Artie bias corpus: An open dataset for detecting demographic bias in speech applications. In *Proceedings of the twelfth language resources and evaluation conference*, pp. 6462–6468, 2020.

Paulius Micikevicius et al. Mixed precision training. *arXiv preprint arXiv:1710.03740*, 2017.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 5206–5210. IEEE, 2015.

Douglas B Paul and Janet Baker. The design for the wall street journal-based csr corpus. In *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*, 1992.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pp. 28492–28518. PMLR, 2023.

Colin Raffel et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL `http://jmlr.org/papers/v21/20-074.html`.

Pooyan Rahmanzadehgervi, Logan Bolton, Mohammad Reza Taesiri, and Anh Totti Nguyen. Vision language models are blind, 2024. URL https://arxiv.org/abs/2407.06581.

Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint arXiv:1904.05862*, 2019.

Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. Lip reading sentences in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6447–6456, 2017.

Shubham Toshniwal, Anjuli Kannan, Chung-Cheng Chiu, Yonghui Wu, Tara N Sainath, and Karen Livescu. A comparison of techniques for language model integration in encoder-decoder speech recognition. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pp. 369–375, 2018. doi: 10.1109/SLT.2018.8639038.

Emmanuel Vincent, Shinji Watanabe, Jon Barker, and Ricard Marxer. The 4th chime speech separation and recognition challenge. *URL: http://spandh. dcs. shef. ac. uk/chime_challenge/(last accessed on 1 August, 2018)*, 2016.

Shinji Watanabe, Michael Mandel, Jon Barker, Emmanuel Vincent, Ashish Arora, Xuankai Chang, Sanjeev Khudanpur, Vimal Manohar, Daniel Povey, Desh Raj, et al. Chime-6 challenge: Tackling multispeaker speech recognition for unsegmented recordings. *arXiv preprint arXiv:2004.09249*, 2020.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021.

Thomas Wolf et al. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, Online, October 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.emnlp-demos.6.

A ADDITIONAL EXPERIMENTS

Table 6: WER on different datasets. The LLMs were fine-tuned on the CV18 EN train set.

LLM \Dataset	**ATIS**	**CHiME4**	**CORAAL**	**LRS2**
No LLM	18.2	18.0	27.4	25.9
Qwen2-0.5B	13.5	17.0	54.6	15.9
Qwen2-1.5B	12.6	17.4	64.6	17.7