Task-Agnostic Low-Rank Dialectal Adapters for Speech-Text Models

Stanford CS224N Custom Project

Amy Guan
Department of Mathematics
Stanford University
amyguan@stanford.edu

Azure Zhou
Department of Mathematics
Stanford University
amysz@stanford.edu

Claire Shao
Department of Mathematics
Stanford University
clshao@stanford.edu

Abstract

Modern NLP systems have significantly lower performance on non-standard dialects, which can compound into higher bias and performance errors in downstream applications. We investigate a data-efficient, task-agnostic method of improving speech recognition for dialects by training LoRA adapters for cross-dialectal alignment of speech encodings. In particular, we minimize the earth mover's distance between the pre-trained encoder embeddings of a source dialect – North Indian English and Filipino English – and United States English. We find that this improves ASR performance for our primary dataset, but has poor performance on more general datasets, indicating a need for more varied training data. Attaching these modules to a pre-trained multimodal model, we find that the adapters result in minimal change in the performance on a downstream question-answering task. Our results suggest that alignment loss for speech encodings may not be immediately compatible with downstream systems.

1 Key Information

- Staff Mentor: Bessie Zhang
- External Mentor: Will Held

2 Introduction

Dialects refer to variations in language linked to regional or local communities, that can differ across ethnic, cultural, and socioeconomic groups (Haugen, 1966). As large language models become increasingly popular within our multicultural societies Chang et al. (2023), it is important that these models are built with consideration to the diversity in language of its users and are able perform invariantly to dialectal differences. Current NLP systems have significant performance disparities on non-standard dialects (Ziems et al., 2023; Blodgett et al., 2016; Okpala et al., 2022). In the case of English, these discrepancies arise in part because standard performance benchmarks evaluate primarily on Standard American English (SAE) and because the distribution of languages in training corpora is unclear (Joshi et al., 2024), which raises concerns for language variations under-represented within these corpora (Hovy and Spruit, 2016).

Speakers of non-standard dialects experience more failures and allocational harms when interacting with downstream applications (Jurgens et al., 2017; Weidinger et al., 2021; Zhou, 2021). In particular, there is substantial room for improvement in speech-text models (Hirayama et al., 2015; Martin and Tang, 2020; Rajpal et al., 2020). Faisal et al. (2021) show that models based on the Google speech-to-text API are much more equitable for US English speakers than other dialect speakers, such as South Indian English. There is additionally a significant downstream effect of noise in automatic speech recognition (ASR) within speech-to-text systems (Ravichander et al., 2021). As systems that

rely on speech-text models, such as Google Assistant and Apple Siri, have worldwide usage, the harms arising from models with poor accessibility may further perpetuate existing inequities.

This highlights the need for the development of dialect-robust technologies and efficient methods of improving dialect performance. To this end, we investigate a new method to reduce the performance gap of state-of-the-art ASR systems on non-standard English dialects. In our experiments, we use dialect interchangeably with regional/national variants of our language of interest, acknowledging that these regions are not monolithic and may contain diverse local, cultural, socio-economic, racial, or ethnic language variants. Following previous dialectal robustness work in the text domain, we adapt OpenAI's Whisper, which is a pre-trained robust ASR model (Radford et al., 2023), using low-rank adapters (LoRA) (Hu et al., 2021) trained on an optimal transfer loss to align the dialects at their encoder representations. The novelty of our approach lies in applying the cross-dialectal alignment methodology to the speech domain.

This method circumvents the need for task-specific dialectal data for downstream tasks, suggesting an inexpensive way to build robust, multilingual speech-text models. However, while we are able to improve ASR performance on a samples similar to our training set, ASR performance decreases on another set. This indicates potential of the dialectal alignment objective in improving ASR on similar datasets, but, given the richness of audio data, more work should be done by training on more varied samples of data. Furthermore, there was a decrease in performance on the downstream task, which suggests that the alignment loss was unable to shift encodings enough to significantly impact performance on the downstream QA task or that the model architecture was incompatible with our alignment objective.

3 Related Works

3.1 Speech-text Models

Hirayama et al. (2015) conduct ASR on a variety of Japanese dialects. They first simulate a dialect corpus via machine translation techniques, then assign the audio to a weighted mixture of dialects, and evaluate upon a corresponding mixed or single dialect language model. Both the mixed and single dialect language models outperformed the common language model but required individual dialectal language models.

Additional prior research by Thomas et al. (2022) demonstrates that applying adapters to self-supervised speech models for ASR, such as wav2vec 2.0, decreases the number of parameters required for downstream ASR tasks and increases scalability across various tasks and languages with little to no compromises in performance. While standard finetuning of wav2vec 2.0 for downstream tasks requires retraining 95.6% of total model parameters per task, adapters allow for training of less than 10% of model parameters per task .

3.2 Dialectal NLP

A growing body of research highlights the performance disparities in natural language processing (NLP) systems' treatment of various dialects, which have prompted an increasing trend towards dialect invariant NLP (Joshi et al., 2024). In natural language generation, recent work by Sun et al. (2022) develops formalized evaluation metrics for dialect robustness and dialect awareness, proposing NANO, an unsupervised pretraining step that distills dialect information into a model. Ziems et al. (2023) introduced Multi-VALUE, a rule-based translation system across dialects that can be used to generate benchmarks for evaluating English dialect invariance and as a data augmentation technique to improve existing systems. Transfer learning from a high-resource language to a dialect has shown strong results and become a dominant paradigm within Dialectal NLP, with multilingual models able to generalize to target languages even when labeled training data is available only for the source language (Held et al., 2023; Scherrer et al., 2023; Zampieri et al., 2020; Wang et al., 2020).

3.3 Cross-lingual Alignment

Bias can emerge in each stage of the machine learning pipeline, beginning with imbalanced training data (Zhang et al., 2018). In fairness research, Hardt et al. (2016) defines "equality of odds" as satisfied when a predictor $\hat{Y}$ and a protected attribute A are independent conditional on outcome Y.

As many downstream systems rely on pre-trained models, this notion of fairness motivates the need for developing large language models whose outputs are invariant to the input dialect.

Cross-lingual alignment methods are one approach for task-agnostic unsupervised transfer across languages and improving multilingual models, and are most effective when applied to the alignment of highly similar languages, making dialectal alignment particularly suitable (Cao et al., 2020; Conneau et al., 2018). Previous work attempting to perform cross-lingual alignment found that minimizing an approximated Wasserstein distance – including via Sinkhorn's divergence – was effective (Romanov et al., 2019; Zhang et al., 2017). Explicit, task-agnostic alignment of dialects with composable modules was first explored by Held et al. (2023), who adapted a model pre-trained on SAE to African American English and other global dialects using L2 alignment loss at the sequence level and an adversarial morphosyntactic alignment loss. Xiao et al. (2023) proposes an efficient adaption method using hypernetworks to generate dialect-specific LoRA adapters for token-level alignment, measured using the earth mover's distance.

4 Approach

In our project, we develop low-rank adapters which aim to improve task-agnostic dialect robustness for automatic speech recognition systems on English dialects (Figure 1). We use OpenAI's Whisper as our ASR model for its state-of-the-art performance. We first identified the top-performing English dialect based on Whisper's Word Error Rate (WER) of transcribing the Mozilla Common Voice 16.1 dataset (Ardila et al., 2019). We then trained our adapters using SD-QA (Faisal et al., 2021), a corpus of parallel spoken utterances across various dialects.

These utterances are already aligned at the word-level; we train our adapters to align the encoder representations by minimizing the earth mover's distance, discussed further below, between a source dialect and the fixed embeddings of the target top-performing dialect. We consider the model to be dialect-invariant for two dialects if the model outputs the same representation for sentence-equivalent input pairs. At test time, we attach our LoRA modules to the Whisper encoder head of the multimodal model SALMONN (Speech Audio Language Music Open Neural Network) (Tang et al., 2023), visualized in Figure 2. Building upon our mentor's existing adaptation of the SALMONN source code for a question-answering task, we evaluate performance both on transcription using Whisper and on question-answering using SALMONN.

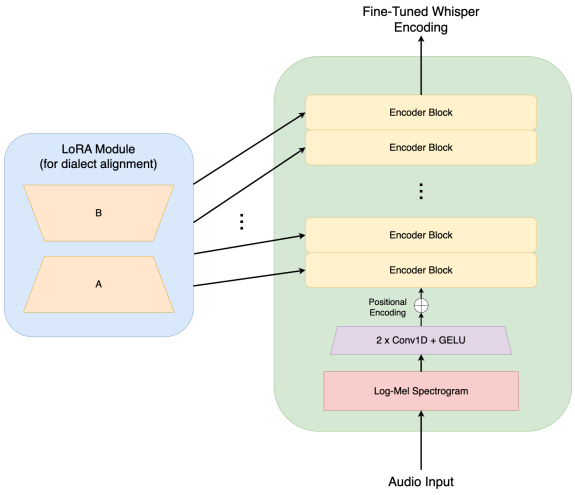

Figure 1: **Whisper encoder + LoRA module architecture.** A and B represent two low-rank weight matrices that are applied to all linear layers of the encoder.

The originality of our approach lies in applying the cross-dialectal alignment methodology to the audio domain. Similar to previous work in dialectal robustness, we use a limited amount of training data (1,000 samples) and a parameter-efficient method. We design our methodology this way to reduce computational barriers and increase the accessibility of our methods for non-English dialects and other data-limited languages, in attempt to address the "low-resource double bind"(Ahia et al.,

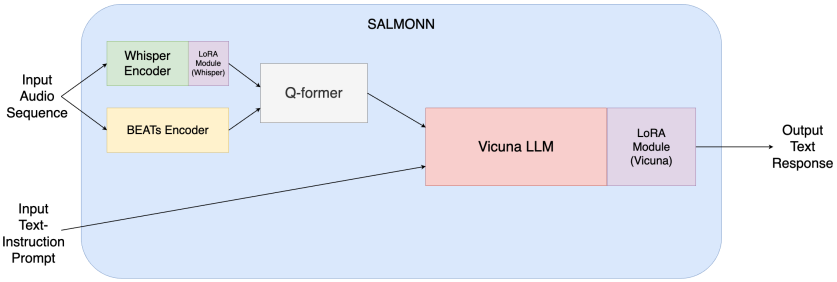

Figure 2: **The overall SALMONN architecture, augmented with our dialect-trained LoRA module.** Note that a more detailed representation of the Whisper Encoder + LoRA module architecture is depicted in Figure 1.

2021). Our approach aims to increase access to language technology for billions of low-resource English dialect speakers.

4.1 Unsupervised alignment loss

Treating word embedding spaces as analogous to distributions, we minimize the distance between the word embedding space of a source dialect and a target dialect using the earth mover's distance, a measure of distribution divergence (Zhang et al., 2017). The earth mover's distance can be approximated using Sinkhorn's divergence defined as:

$$S_\epsilon(\alpha, \beta) := W_\epsilon(\alpha, \beta) - \frac{1}{2}W_\epsilon(\alpha, \alpha) - \frac{1}{2}W_\epsilon(\beta, \beta), \tag{1}$$

where α and β are the distributions of interest and W_ϵ, given by is an efficient regularized Wasserstein distance (Feydy et al., 2019; Cuturi, 2013). Conceptually, this loss can be interpreted as the cost of moving mass from distribution α to β.

5 Experiments

5.1 Data

The datasets used in our experiments are the Spoken Dialectal Question Answering (SD-QA) and Common Voice 16.1 datasets (Faisal et al., 2021; Ardila et al., 2019). The SD-QA dataset is a multi-dialect, spoken data benchmark for five languages (Arabic, Bengali, English, Kiswahili, and Korean). For English, SD-QA consists of 2k audio prompts with 11 regional dialects samples for each. From the US English, North Indian English, and Filipino English dialect samples of SD-QA, we use 1k audio prompts for training and 1.03k audio prompts for testing.

The Common Voice dataset is an open-source, multi-language voice dataset consisting of 30,329 hours of recorded data and 19,916 hours of validated data in 120 languages, mainly selected for its large-scale availability of data. We specifically use a subset of English samples with accent tags "United States English," "India and South Asia (India, Pakistan, Sri Lanka)," and "Filipino," with 300 samples for validation and 1k for test.

5.2 Evaluation method

We originally determine the best-performing English dialect by transcribing a random subset of 5k labeled accented samples from the Common Voice dataset. Our ASR evaluation metric is the word error rate (WER), which is the number of (Substitutions + Insertions + Deletions) / (Number of Words Spoken) in the reference transcription. We compute the WER on our Common Voice validation set during hyperparameter tuning as well as on our Common Voice test set after final training. Excluding English dialects with less than 50 samples and those without a corresponding dialect in SD-QA, we found that USA English had the best performance with a WER of 8.600. The full set of metrics are included in the appendix.

Finally, we evaluate performance on QA by attaching our trained adapters to SALMONN. We assessed SALMONN 7B's CFMatch with no modifications on the SD-QA dataset to obtain our baselines. To ensure consistent and comparable results, we prompt SALMONN to give a simple one sentence answer to all audio prompts with the highest CFMatch score. Specifically, CFMatch is a question-answering (QA) evaluation metric that aims to determine answer equivalence (AE) by combining standard F_1 evaluation with a discriminative logistic regression classifier trained on an augmented AE dataset (Li et al., 2024). These modifications allow CFMatch to be less overly-strict and less sensitive to thresholding than other common evaluation metrics such as EM and F_1.

5.3 Experimental details

We trained our adapters using LoRA from HuggingFace PEFT (Mangrulkar et al., 2022) and the 'dev' split of SD-QA. Our custom trainer utilized Sinkhorn loss from the GeomLoss library (Feydy et al., 2019). We targeted all linear layers within the encoder of Whisper, used an alpha of 64, a dropout probability of 0.5, and no bias. For time constraints, we tuned hyperparameters by training on Whisper-small, using a sample from Common Voice as our evaluation data and WER as our evaluation metric. We performed a grid search over learning rates $1 \cdot 10^{-3}, 5 \cdot 10^{-2}, 1 \cdot 10^{-2}$, batch sizes $16, 32, 64$, and LoRA rank $32, 64, 128$. From this, we selected a learning rate of 1e-3, a batch size of 32, and a rank (the LoRA attention dimension) of 32. As LoRA adapters train quickly, we initially trained on 3 epochs. We found that our loss was still improving significantly and extended our training to 15 epochs, at which point the loss stabilized.

We trained our final adapters with fp16 precision on Whisper-large-v2 over 15 epochs. We trained our first adapter from the source North Indian English to the target USA English and repeated this process with the same hyperparameters from the source Filipino English to the target USA English. Training our final adapters took approximately 8 hours each.

To evaluate on the downstream task, we use the input text instruction prompt "You are a helpful assistant. Give a simple one sentence answer." to instruct SALMONN to answer open-ended questions about the SD-QA audio data. For LLM text response generation, we choose a top probability of 1.0.

5.4 Results

Table 1: WER for Baseline and Model on Source and Target for North Indian Model and Filipino Model

Model	**Dialect**	**WER**		**Baseline - Model WER**
		Baseline	Model	
North Indian Model	Northern Indian (SD-QA)	11.317	9.276	2.104
	US (SD-QA)	5.834	5.324	0.510
	South Asian (CV)	14.361	18.867	-4.506
	US (CV)	9.251	10.384	-1.133
Filipino Model	Filipino (SD-QA)	8.001	7.587	0.414
	US (SD-QA)	5.834	5.387	0.447
	Filipino (CV)	11.180	16.632	-5.452
	US (CV)	9.251	11.458	-2.207

We evaluated the word error rate (WER) of automatic speech recognition with the adapted Whisper model on the SD-QA 'test' set (Table 1). For each of the adapted models, there is an improvement in WER in both target and source dialects. The North Indian model shows a decrease in WER by 2.104 or 18.592% for Northern Indian English; the Filipino model shows a decrease in WER by 0.414 or 5.174% for Filipino English. For United States English, there is a decrease in WER by 0.510 (8.742%) and 0.447 (7.662%) for the North Indian and Filipino models, respectively.

We also evaluate the WER on a 1k sample subset of each dialect from Common Voice to observe the generalizability of the adapter across speakers and prompt types beyond the scope of QA. We find that there is an increase in WER for ASR with the adapter compared to the baseline across each tested dialect, with an average 2.82 and 3.82 increase in WER on the Common Voice samples for the North

Indian and Filipino models, respectively (Table 1). The change in WER for US English for each of these two models is of smaller magnitude compared to the change in WER for the source dialects.

To benchmark our model on downstream performance, we evaluated the performance of SALMONN 7B on the pretrained Whisper model embeddings and the SD-QA test dataset. Out of the baseline, North India model, and Filipino model, the best performance was obtained by the baseline on US English, with a CFMatch score of 0.3730 (Table 2). For each source dialect, the highest CFMatch score was similarly obtained by the baseline. Within each model, US English exhibited the greatest decreases in CFMatch score between baseline and model (0.0194 for the North India model, and 0.0269 for the Filipino model).

Table 2: SALMONN-7B CFMatch Scores on English Dialects

Model	**Dialect**	**CFMatch**		**Baseline - Model CFMatch**
		Baseline	Model	
North Indian Model	Northern Indian	0.3535	0.3532	0.0003
	US	0.3730	0.3536	0.0194
Filipino Model	Filipino	0.3551	0.3546	0.0005
	US	0.3730	0.3461	0.0269

6 Analysis

6.1 Automatic Speech Recognition

Given our training objective of minimizing the alignment loss between parallel encoder representations, we hypothesized that the LoRA adapter would minimize the difference in Whisper performance between the source and target dialect, bringing the WER of the source dialect closer to that of US English while potentially increasing word error rate for US English. On SD-QA test, we observed that the WER of the source dialect did decrease and shift toward the lower US English WER, which indicated that the dialectal alignment objective was effective at improving ASR performance. In particular, there was a greater improvement in the Northern Indian English WER with a decrease of 2.104 for Northern Indian English vs 0.414 for Filipino English (Table 1). This behavior is expected since the baseline WER for Northern Indian English was significantly higher than the baseline WER for Filipino English; intuitively, decreasing alignment loss provides more significant benefits for Northern Indian English. (Notably, there are more English speakers in India then USA, Australia, and England, yet the performance of NLP systems is worse for Indian English dialects (Joshi et al., 2024).)

Interestingly, both adapters marginally improved the US English WER as well. This was contrary to our expectation that additional training would result in negative transfer for US English, which has been observed in prior transfer learning (Wang et al., 2019). The US English WER improving on both adapters indicates against the improvement being only the result of randomness, and suggests that latent symmetries or abstractions shared across dialects could have been captured during training (Wu et al., 2019). The result that our unsupervised alignment loss is able to improve automatic speech recognition for both the source and target language is not obvious, and supports further study into audio embedding spaces across dialects.

On the Common Voice dataset, our adapters show poor performance on WER across each model and dialect. This increase in Common Voice WER was observed when training with a low number of epochs and rose when training over many epochs. This is likely a result of SD-QA being more standardized than Common Voice – which is a open-source, decentralized dataset featuring crowd-sourced samples with varying speech input systems and featuring diverse prompts outside the realm of question-answering. As our adapters show improved performance on the held-out SD-QA dataset, this result indicates that our adapter may have learned unwanted attributes specific to the SD-QA benchmark. This provides the insight that our alignment loss method requires training examples which vary in speakers, prompt type, and microphone characteristics in order to generalize to all speech input.

6.2 Downstream Question-Answering Task

By measuring the performance of a downstream system relying on the Whisper encoder, we attempt to gauge the ability of each embedding to be meaningful in terms of a real-world task. Despite the fact that the WER for each dialect increased with each adapter on SD-QA, our results indicate that the baseline model with no adapter had the overall best performance for the question-answering task. This is surprising given strong prior evidence from the text domain that cross-dialectal alignment of word representations can improve overall performance of multilingual models on dialectal data, see Cao et al. (2020); Xiao et al. (2023).

It is intuitive that the CFMatch for the US dialect performed worse since the US dialect was not part of the training data during adapter training. However, the CFMatch for North Indian and Filipino dialects reported very minor decreases in performance. It is possible that the alignment loss was unable to shift the encoder representations of the source dialects significantly enough to effect performance on the downstream SALMONN model. It is also notable that there is not a large gap between CFMatch performance on US and the two source dialects; it may be useful to explore dialects that began with greater performance disparity. Finally, a potential interpretation for the large change in US dialect may be found in the pre-training of the SALMONN model. SALMONN trains a window-level Q-former, which fuses the Whisper encoder output with a non-speech audio encoder output, and a LoRA module, for its base large language model, on datasets such as Librispeech and Gigaspeech, which to our knowledge do not have documented distributions of the dialects included in them. It is possible that the alignment loss objective shifted the US English feature representations in a way which is interpreted differently by the LoRA-adapted language model. In addition, the impact of the US embedding shift may have been amplified by the Q-former and the structure of SALMONN.

Our results indicate that there may be more complexities within speech embeddings that must be accounted for, and that our embedding-alignment loss adapters may not be immediately suitable for plug-and-play into additional models. There may be a disconnect between the training objective of minimizing alignment loss with retaining meaning for downstream tasks. Expanding upon considerations mentioned in the above section, we may include incorporating more variation in speakers, audio quality, and within the audio samples, as well as diversity of the downstream task.

7 Conclusion

In this paper, we develop LoRA adapters for speech-text models that minimizes the alignment loss between the speech embeddings of a source dialect of English – North Indian or Filipino – and United States English. We find that the adapter and dialectal alignment objective hold potential in improving ASR, but, given the richness of audio data, the adapter is likely learning content and audio qualities specific to SD-QA. Thus, more work should be done by training the adapter with more varied samples of data from different speakers with different speech characteristics, microphone qualities, and prompts. On a downstream QA task, the adapter is not able to adequately gain necessary information after alignment. As a result, we may need to investigate how the shifted embeddings may interact with other components of pre-trained composite models. This could then help reframe a more pertinent training objective that would more effectively generate task-agnostic adapters.

One limitation of our work is that our method was only evaluated on English dialects. Due to time and computational restraints, we also restricted our experiments to analyzing dialect-specific adapters, mapping a specific source dialect to our best performing dialect. However, a dialect-invariant generic adapter that maps all dialects to our target dialect may alleviate the large distributional shift we see in US English on SALMONN, and would be an interesting avenue for further exploration.

8 Acknowledgements

We would like to acknowledge William Held in the Stanford NLP group for his guidance and mentorship and Bessie Zhang for her support.

8.1 Team Member Contributions

Amy and Azure contributed equally to pre-processing data, tuning and training the LoRA models, evaluating the Whisper ASR, and analyzing results. Claire set up code for downstream task evaluation

(SALMONN), ran baselines for SALMONN + SD-QA, set up code for integrating LoRA adapter with downstream task, and built a system demo for the final poster presentation. Amy additionally did final ASR evaluations and troubleshot compute issues. Azure additionally conducted literature review and ran final SALMONN evaluations.

References

Orevaoghene Ahia, Julia Kreutzer, and Sara Hooker. 2021. The low-resource double bind: An empirical study of pruning for low-resource machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3316–3333, Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

Su Lin Blodgett, Lisa Green, and Brendan O'Connor. 2016. Demographic dialectal variation in social media: A case study of African-American English. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1119–1130, Austin, Texas. Association for Computational Linguistics.

Steven Cao, Nikita Kitaev, and Dan Klein. 2020. Multilingual alignment of contextual word representations. *arXiv preprint arXiv:2002.03518*.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2023. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Marco Cuturi. 2013. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26.

Fahim Faisal, Sharlina Keshava, Antonios Anastasopoulos, et al. 2021. Sd-qa: Spoken dialectal question answering for the real world. *arXiv preprint arXiv:2109.12072*.

Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-ichi Amari, Alain Trouvé, and Gabriel Peyré. 2019. Interpolating between optimal transport and mmd using sinkhorn divergences. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2681–2690. PMLR.

Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of opportunity in supervised learning. *CoRR*, abs/1610.02413.

Einar Haugen. 1966. Dialect, language, nation 1. *American anthropologist*, 68(4):922–935.

William Held, Caleb Ziems, and Diyi Yang. 2023. TADA : Task agnostic dialect adapters for English. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 813–824, Toronto, Canada. Association for Computational Linguistics.

Naoki Hirayama, Koichiro Yoshino, Katsutoshi Itoyama, Shinsuke Mori, and Hiroshi G Okuno. 2015. Automatic speech recognition for mixed dialect utterances by mixing dialect language models. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(2):373–382.

Dirk Hovy and Shannon L. Spruit. 2016. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598, Berlin, Germany. Association for Computational Linguistics.

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

Aditya Joshi, Raj Dabre, Diptesh Kanojia, Zhuang Li, Haolan Zhan, Gholamreza Haffari, and Doris Dippold. 2024. Natural language processing for dialects of a language: A survey. *arXiv preprint arXiv:2401.05632*.

David Jurgens, Yulia Tsvetkov, and Dan Jurafsky. 2017. Incorporating dialectal variability for socially equitable language identification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 51–57, Vancouver, Canada. Association for Computational Linguistics.

Zongxia Li, Ishani Mondal, Yijun Liang, Huy Nghiem, and Jordan Boyd-Graber. 2024. Cfmatch: Aligning automated answer equivalence evaluation with expert judgments for open-domain question answering. *arXiv preprint arXiv:2401.13170*.

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

Joshua L Martin and Kevin Tang. 2020. Understanding racial disparities in automatic speech recognition: The case of habitual" be". In *Interspeech*, pages 626–630.

Ebuka Okpala, Long Cheng, Nicodemus Mbwambo, and Feng Luo. 2022. Aaebert: Debiasing bert-based hate speech detection models via adversarial learning. In *2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 1606–1612.

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

Avni Rajpal, Achuth Rao, Chiranjeevi Yarra, Ritu Aggarwal, and Prasanta Kumar Ghosh. 2020. Pseudo likelihood correction technique for low resource accented asr. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7434–7438. IEEE.

Abhilasha Ravichander, Siddharth Dalmia, Maria Ryskina, Florian Metze, Eduard Hovy, and Alan W Black. 2021. NoiseQA: Challenge set evaluation for user-centric question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2976–2992, Online. Association for Computational Linguistics.

Alexey Romanov, Anna Rumshisky, Anna Rogers, and David Donahue. 2019. Adversarial decomposition of text representation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 815–825, Minneapolis, Minnesota. Association for Computational Linguistics.

Yves Scherrer, Tommi Jauhiainen, Nikola Ljubešić, Preslav Nakov, Jörg Tiedemann, and Marcos Zampieri. 2023. Tenth workshop on nlp for similar languages, varieties and dialects (vardial 2023). In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*.

Jiao Sun, Thibault Sellam, Elizabeth Clark, Tu Vu, Timothy Dozat, Dan Garrette, Aditya Siddhant, Jacob Eisenstein, and Sebastian Gehrmann. 2022. Dialect-robust evaluation of generated text. *arXiv preprint arXiv:2211.00922*.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. 2023. Salmonn: Towards generic hearing abilities for large language models. *arXiv preprint arXiv:2310.13289*.

Bethan Thomas, Samuel Kessler, and Salah Karout. 2022. Efficient adapter transfer of self-supervised speech models for automatic speech recognition. *CoRR*, abs/2202.03218.

Zirui Wang, Zihang Dai, Barnabás Póczos, and Jaime Carbonell. 2019. Characterizing and avoiding negative transfer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11293–11302.

Zirui Wang, Zachary C Lipton, and Yulia Tsvetkov. 2020. On negative interference in multilingual models: Findings and a meta-learning treatment. *arXiv preprint arXiv:2010.03017*.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. 2021. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*.

Shijie Wu, Alexis Conneau, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Emerging cross-lingual structure in pretrained language models. *arXiv preprint arXiv:1911.01464*.

Zedian Xiao, William Held, Yanchen Liu, and Diyi Yang. 2023. Task-agnostic low-rank adapters for unseen English dialects. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7857–7870, Singapore. Association for Computational Linguistics.

Marcos Zampieri, Preslav Nakov, and Yves Scherrer. 2020. Natural language processing for similar languages, varieties, and dialects: A survey. *Natural Language Engineering*, 26(6):595–612.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating unwanted biases with adversarial learning. *CoRR*, abs/1801.07593.

Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. 2017. Earth mover's distance minimization for unsupervised bilingual lexicon induction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1934–1945, Copenhagen, Denmark. Association for Computational Linguistics.

Xuhui Zhou. 2021. *Challenges in automated debiasing for toxic language detection*. University of Washington.

Caleb Ziems, William Held, Jingfeng Yang, Jwala Dhamala, Rahul Gupta, and Diyi Yang. 2023. Multi-VALUE: A framework for cross-dialectal English NLP. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 744–768, Toronto, Canada. Association for Computational Linguistics.

A Appendix

Table 3: Whisper Large ASR Error Rates for Selected English Dialects

Dialect	**WER**	**Norm. WER**
Australian English	9.92	5.62
Indian & South Asian English	12.42	7.37
Irish English	14.36	6.85
New Zealand English	10.44	6.30
Scottish English	38.06	34.15
Southern African English	14.51	10.74
United States English	8.60	4.57