ASR Data Selection from Multiple Sources: A Practical Approach on Performance Scaling

Hoang Anh Just*
Virginia Tech
just@vt.edu

I-Fan Chen
Amazon Alexa AI
ifanchen@amazon.com

Feiyang Kang
Virginia Tech
fyk@vt.edu

Yuanzhi Zhang
Virginia Tech
yuanzhi@vt.edu

Anit Kumar Sahu
Amazon Alexa AI
anit.sahu@gmail.com

Ruoxi Jia
Virginia Tech
ruoxijia@vt.edu

Abstract

This paper proposes a framework leveraging small samples from different Automatic Speech Recognition (ASR) data sources to predict model performance and facilitate ASR data selection decisions. By utilizing data distribution distance and a mapping technique inspired by neural scaling laws, our framework estimates the model performance for various data mixtures within the disclosed range and extrapolates it onto much larger target data sizes. This is the first study on extending this novel approach to ASR problems. Experiments conducted on the LibriSpeech and the TED-LIUM3 datasets confirm the effectiveness of the proposed data selection framework. Compared to a heuristic-based selection baseline, our framework consistently demonstrates $13 \sim 17\%$ relative word error rate reductions under 40/ 50/ 100-hour fine-tuning data hour budgets.

1 Introduction

In Automatic Speech Recognition (ASR) domain, researchers have started training single ASR models capable of processing speech data from diverse acoustic channels, background environments, speaker groups, accents, and even languages [1]. Within the realm of general-purpose large language models (LLMs) and unified ASR models, scaling laws have been established that explore the relationship and interplay between model performance, model size, and the amount of training data [2, 3]. These studies emphasize the importance of preparing high-quality training data to optimize model performance. However, in real-world scenarios, multiple data sources are often available for model training, raising the question of how to effectively combine data from these different sources to extract good performance out of a model. Two prevalent approaches have been utilized for data combination in ASR training. The first approach involves using all available data from each source without re-weighting or re-balancing the datasets [1]. While this approach is simple, it may not be the most efficient allocation of the training computation budget, as some data could be duplicated and less relevant than others. The second approach entails applying a combination of weights to mix training data from different sources based on heuristic assumptions or manual tuning. Although this approach offers improvements over the non-weighted approach in terms of the training cost, it does not guarantee that the combination weights are geared for the target tasks due to a shift in the sampling distribution due to heterogeneous weights [4–6]. (For extended related work, please see Appendix A.) Additionally, manual weight tuning can also be expensive and time-consuming. Furthermore, existing scaling laws primarily focus on the overall size of the training data, neglecting the fact that different types of data might have distinct scaling relationships with model performance. This limitation underscores the need to explore alternative approaches to data combination. Another challenge in data selection arises when the full dataset from each source is not completely revealed, where only a subset of examples is disclosed or available, leaving the question of whether to utilize the

Appears in 3^{rd} *Efficient Natural Language and Speech Processing (ENLSP-III) Workshop* at 37th Conference on Neural Information Processing Systems (NeurIPS 2023), New Orleans, LA, USA.

dataset unanswered. To address these issues, a recent work proposed `projektor`, a data selection approach that combines performance scaling laws with optimal transport-based distribution distance computation [7]. This approach allows for the determination of appropriate mixing weights for each data source, given a total training data budget and target model performance evaluation setup, without actually performing full-scale model training. The proposed approach has demonstrated effectiveness in image recognition and Natural Language Processing (NLP) tasks. Therefore, in this paper, we aim to extend `projektor` to ASR tasks. The contributions of this paper include establishing the necessary adaptations required to implement the idea of `projektor` to ASR tasks and conducting experiments to analyze its effectiveness in model fine-tuning scenarios. By investigating the suitability of this approach in ASR, we aim to advance the understanding of data selection methods for ASR tasks and potentially contribute to more efficient and optimized ASR model training.

2 Methods

First proposed in [7], `projektor` integrates Optimal Transport (OT) and scaling laws to provide accurate predictions on the performance of machine learning models to guide the selection of training data. In this section, we first introduce the technical framework and formulate the data selection problem in ASR inspired by [7]. Then, we discuss the detailed operational pipeline for implementing the method before applying it to ASR tasks. We start with the preliminaries.

Consider m datasets $D_1, \ldots, D_m$ representing m different *data sources* and a practitioner with a validation set D^{val}, who would like to combine samples from these datasets to train a model $\mathcal{A}$ with performance metric $\mathcal{L}$. Given a selection budget of N samples and a mixing ratio of data sources $\mathbf{p} = \{p_1, \ldots, p_m\}, \forall_i\, 0 \leq p_i \leq 1$, and $\sum p_i = 1$, denote the selected dataset by $\mathcal{D}(N, \mathbf{p}) = S_1 \cup \cdots \cup S_m$, where $S_i \subseteq D_i$ is a random collection of subset of samples of D_i and $|S_i| = p_i N$. The practitioner seeks to maximize the resulting model performance by strategically choosing the mixing ratio $\mathbf{p}$ of m data sources for a given *target* dataset size N, i.e., $\max_{\mathbf{p}} \mathcal{L}(\mathcal{A}(\mathcal{D}(N, \mathbf{p})), D^{\text{val}})$.

2.1 `projektor`-based Data Selection

Optimal Transport is a distance metric between probability distributions [8] enjoying advantageous analytical properties [9, 10]. Given training and validation probability measures μ_t, μ_v over space $\mathcal{Z}$, OT distance is defined as $\text{OT}(\mu_t, \mu_v) := \min_{\pi \in \Pi(\mu_t, \mu_v)} \int_{\mathcal{Z}^2} \mathcal{C}(z, z') d\pi(z, z')$, where $\Pi(\mu_t, \mu_v)$ denotes the set of couplings over μ_t, μ_v and $C : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}^+$ is some positive cost function. Inspired by theoretical results that the upper bound on the difference between training loss and validation loss can be tightly bounded by an affine transformation of the OT distance [11, 12], [7] propose to directly estimate this transformation by empirically fitting OT to model performance and then the estimated transformation can be used for predicting the model performance for different data mixtures as $\hat{\mathcal{L}}\left(\mathcal{A}(\mathcal{D}(N, \mathbf{p})), D^{\text{val}}\right) = a_1 \cdot \text{OT}\left(\mathcal{D}(N, \mathbf{p}), D^{\text{val}}\right) + a_0$, where a_0, a_1 can be estimated via least-square fitting. An alternative nonlinear version is provided as [7] $\hat{\mathcal{L}}(\mathcal{A}(\mathcal{D}(N, \mathbf{p})), D^{\text{val}}) = \sum_{i=1}^{m} (c_2^i p_i^2 + c_1^i p_i + c_0) + \sum_{i=1}^{m} (b_2^i p_i^2 + b_1^i p_i + b_0) \cdot \text{OT}(\mathcal{D}(N, \mathbf{p}), D^{\text{val}})$, where b^i, c^i are additional parameters for fitting the performance predictor and p_i is the data ratio associated with the data source i. However, if the target size N is large, then fitting the predictor directly might be inefficient. Therefore, for better efficiency, we apply the neural scaling laws that enable predicting empirical performance changing with the size of the training dataset as $\mathbb{E}_V[\mathcal{L}(\mathcal{A}(\mathcal{D}(N, \mathbf{p})); D^{\text{val}})] \approx -\alpha \log(N) + C$, where α and C are some constants [13]. By first fitting the predictors at smaller scales N_0, N_1 (in practice, we use $N_0 < N_1 < 1\% \cdot N$), we then use smaller-scale predictors to directly fit neural scaling laws for this particular distribution and project it onto larger scale N, which is

$$\hat{\mathcal{L}}(\mathcal{A}(\mathcal{D}(N, \mathbf{p}); D^{\text{val}}) = \left(\log \frac{N_1}{N_0}\right)^{-1} \left[\log \frac{N}{N_0} \hat{\mathcal{L}}(\mathcal{A}(\mathcal{D}(N_1, \mathbf{p})); D^{\text{val}}) \; - \log \frac{N}{N_1} \hat{\mathcal{L}}(\mathcal{A}(\mathcal{D}(N_0, \mathbf{p}); D^{\text{val}})\right]. \tag{1}$$

The predictions are used to support determining optimal data source selection $\mathbf{p}^* = \arg\max_{\mathbf{p}} \hat{\mathcal{L}}(\mathcal{A}(\mathcal{D}(N_s, \mathbf{p})), D^{\text{val}})$. The problem is then solved effectively via gradient methods, where [12] allows almost free gradient calculation for the *calibrated gradient* of OT. Alg.1 in Appendix B provides a comprehensive explanation of how this pipeline is implemented in practice.

3 Experiments

3.1 Experimental Setup and Implementation

To simulate data selection for training a general ASR model for different types of tasks, we selected two distinct ASR tasks: LibriSpeech [14] and TED-LIUM3 [15]. Our objective was to train a single

ASR model that performs well on both datasets, considering a limited training data budget. We conduct comparisons of data selection strategies in the fine-tuning scenario, where a LibriSpeech 960 pretrained end-to-end ASR model [16] is fine-tuned using a combination of LibriSpeech and TED-LIUM3 data. This scenario aims to assess the effectiveness of data selection for adapting a pretrained model to a different task while maintaining a strong performance on the original task. A visualization of the task and further implementation details can be found in Appendix B and C.

Datasets and Models. We use two different training data sources with distinct types of data: LibriSpeech Clean 100 train and TED-LIUM3. The LibriSpeech Clean 100 training dataset consists of 100 hours of the read speech recordings extracted from audiobooks. The TED-LIUM3 training data consist of 546 hours of the speech data from TED Talks, where the speaker's speaking styles are quite different from that of LibriSpeech. To simulate the model training scenario with different training data budgets for LibriSpeech and TED-LIUM3 training data, we prepare mixed subsets of 10, 20, 40, 50, and 100 hours. The amount of pilot data is limited to $N_0 < N_1 = 1$ hour uniformly sampled from each data source. We use Torch Audio Emformer models [17, 18] for experiments. For the fine-tuning task, we use the LibriSpeech pretrained Emformer model [16]. To establish the scaling laws for data selection, we use even smaller-sized Emformer models (i.e., with two and four transformer layers). We evaluate the trained model performance by the averaged Word Error Rate (WER) on the LibriSpeech Clean Test and TED-LIUM3 Test datasets.

Data Selection Methods. We consider the following two data selection approaches:

- **Heuristic rule-based selection (50:50 Baseline)**: selects equal amounts of training data from the LibriSpeech and TED-LIUM3 datasets based on a given data budget. The selection rule was based on the averaged WER for both tasks, ensuring a balanced representation of data from both sources.
- **Proposed approach:** utilizes the performance of a small-scale model on the validation set to guide data selection with optimal transport distance. By training or fine-tuning models using pilot data and evaluating their performance on the validation set, the selection of data sources and amounts was optimized with the gradient computed from the dual solution of OT.

By comparing these approaches, we assessed their effectiveness in preparing training data for ASR models for different training data budgets. The heuristic rule-based selection offered a straightforward approach, while the proposed approach incorporated OT distance and validation performance to guide data selection, aiming to enhance overall performance on both LibriSpeech and TED-LIUM3 tasks.

3.2 Empirical Results

Scaling Laws for Different ASR Data Types. Our proposed approach assumes that different types of ASR data could have different scaling laws of loss vs. training data size. To validate our assumption, we follow the setup in [3] to derive the scaling laws of loss vs. training data size for both the LibriSpeech and the TED-LIUM3 datasets. The losses were computed on the held-out validation data. To derive the scaling law, we train a smaller Emformer model that keeps the same model architecture of the pretrained Emformer RNN-T model [16], except that we reduced the Emformer layer number from 20 to 10. The total model parameter for the smaller Emformer model is 45.2 M. We reuse the feature extractor and sentence-piece model from the pretrained Emformer Bundle [16]. For LibriSpeech and TED-LIUM3 data, we train a smaller model with 2, 5, 10, 20, 50, 100 hours of the training data on a single GPU and collect the minimal validation loss for each training. The scaling laws thus obtained are $\mathit{Loss} = -74.99\ln(N) + 1053.6$ for LibriSpeech and $\mathit{Loss} = -115.6\ln(N) + 1300.8$ for TED-LIUM3 depicting higher slope for TED-LIUM3, confirming our assumption that different types of ASR data may have different scaling laws, which motivates the proposed approach for data selection. The Figure 3 for log-linear fits is relegated to Appendix A.

WER Performance Prediction with OT

We demonstrate that the proposed method effectively captures the intricate relationship between the model's performance metric (e.g., loss, test WER) and the selected training data used for fine-tuning. Figure 4 demonstrating the effectiveness of the fitted function in terms of predicting the model's performance is relegated to Appendix C. To quantitatively validate the performance of our method,

	10Hr	20Hr	40Hr	50Hr	100Hr
Linear	0.36	0.20	0.31	0.24	0.22
Quadratic	0.22	0.22	0.29	0.21	0.19
Ours	**0.18**	**0.17**	**0.20**	**0.16**	**0.15**

Table 1: Comparison of MAE values for predicting WER performance on unseen data source compositions at different training data budgets: 10Hr, 20Hr, 40Hr, 50Hr, and 100Hr.

we compute the Mean Average Error (MAE) between the predicted WER and the actual WER of different data source compositions for each training data budget. We benchmark our results with baseline methods, Linear and Quadratic, which assume a linear and quadratic relationship with the training data compositions, respectively. As we observe in Table 1, by obtaining the lowest MAE values, our method outperforms the baseline methods in predicting WER in each of the data budget cases. By incorporating optimal transport distance into the fitting function, we can better predict average WER to further guide us in data selection.

Optimal Transport Guides Data Source Selection

We show the performance of our data selection method and compare it with the results of the heuristic-based selection. Our method selects a data source composition by following the OT-based gradient descent. We report the actual average WER performance for each data source composition selected by our method. Table 2 presents the WER results of fine-tuning with the baseline and the proposed selection for data budgets of 40, 50, and 100 hours. Our selection puts a higher selection weight onto TED-LIUM3 data than on the LibriSpeech as it is a more challenging task for the LibriSpeech pretrained Emformer. We observe that the proposed data selection method outperforms the baseline heuristic rule-based selection approach by consistently reducing average WER by an absolute of $1 \sim 1.5\%$ (or a $13 \sim 17\%$ relative WER reduction) in each budget cases. While our method maintains a similar WER performance of the original task (i.e., LibriSpeech), it also significantly improves the WER performance on the new task (i.e., TED-LIUM3).

Budget (Hr)	Method	Train Data L:T3 (Hr)	Test WER (%) L Clean	T3	Avg
40	Base	20:20	5.8	18.0	11.9
	Ours	6:34	6.0	14.6	**10.3**
50	Base	25:25	4.9	18.1	11.5
	Ours	8:42	5.0	13.9	**9.5**
100	Base	50:50	6.8	17.2	12.0
	Ours	18:82	6.5	14.2	**10.4**

Table 2: Comparison between the heuristic selection and the Optimal Transport (OT) Guided selection under different fine-tuning data budgets. L and T3 represent LibriSpeech and TED-LIUM3 data.

Performance Projection onto Larger Data Budgets

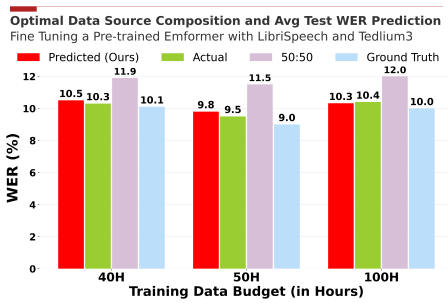

Figure 1: Projected WER performance using proposed data source mixtures. Comparison with baseline selection and actual WER performance.

After selecting the data source composition for each data budget, our method proceeds to project WER performance onto larger data scales without actually fine-tuning the model by incorporating the OT-based scaling law as derived in Eq. 1. We first fit two functions on smaller data budget scales, i.e., 10Hr and 20Hr, respectively, then we project the performance onto larger data budgets of 40Hr, 50Hr, and 100Hr. For each data budget, we apply the data selection ratio given in Table 2 and predict the average WER performance as proposed in Eq. 1. We compare our selection method's predicted WER with the actual average WER of that given selection. We additionally compare with WER performance using the heuristic-based selection baseline (50:50 ratio). Lastly, we attempt to find the "ground-truth" optimal WER performance by grid-searching over data compositions. We note that only our method projects the model's performance, while the heuristic-based selection cannot make such a projection prediction. In Figure 1, we observe that our predicted average WER does not depart from the actual WER by more than 0.3% and outperforms the baseline method in each data budget case by more than 1% of WER reduction. Moreover, we notice that our selection's WER performance is close to the "ground-truth" WER, which indicates the potential of our method in both selecting data source compositions and predicting average WER performance for large data scales.

4 Conclusion

In this paper, we explore the application of the performance scaling via optimal transport approach to ASR data selection from partially revealed sources. We demonstrate that different types of ASR data exhibit distinct scaling laws, which motivates the proposed data selection approach. We evaluate the proposed data selection approach in the setting of model fine-tuning. Preliminary results on LibriSpeech and TED-LIUM3 show the proposed approach outperforms the heuristic rule-based baseline. Future work involves conducting experiments with larger models/datasets and exploring other factors, such as data quality and data influence. Our research contributes to advancing automatic speech recognition with a better budget-efficient utilization of diverse ASR datasets, which could be key for building high-performance general ASR models given a data budget constraint.

Acknowledgments and Disclosure of Funding

RJ and the ReDS lab acknowledge support through grants from the Amazon-Virginia Tech Initiative for Efficient and Robust Machine Learning, the National Science Foundation under Grant No. IIS-2312794, NSF IIS-2313130, NSF OAC-2239622, and the Commonwealth Cyber Initiative.

References

[1] William Chan, Daniel Park, Chris Lee, Yu Zhang, Quoc Le, and Mohammad Norouzi, "SpeechStew: Simply Mix All Available Speech Recognition Data to Train One Large Neural Network," http://arxiv.org/abs/2104.02133, Apr. 2021. 1

[2] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei, "Scaling Laws for Neural Language Models," https://arxiv.org/abs/2001.08361, Jan. 2020. 1

[3] Jasha Droppo and Oguz Elibol, "Scaling laws for acoustic models," in *Interspeech 2021*, 2021. 1, 3

[4] Yi Wu, Rong Zhang, and Alexander Rudnicky, "Data selection for speech recognition," in *2007 IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU)*, Dec. 2007, pp. 562–565. 1, 7

[5] Jeff Bilmes, "Submodularity In Machine Learning and Artificial Intelligence," https://arxiv.org/abs/2202.00132, Oct. 2022. 7

[6] Chanho Park, Rehan Ahmad, and Thomas Hain, "Unsupervised Data Selection for Speech Recognition with Contrastive Loss Ratios," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2022, pp. 8587–8591. 1, 7

[7] Feiyang Kang, Hoang Anh Just, Anit Kumar Sahu, and Ruoxi Jia, "Performance scaling via optimal transport: Enabling data selection from partially revealed sources," *Advances in Neural Information Processing Systems*, vol. 36, 2023. 2, 7

[8] Cédric Villani, *Optimal transport: old and new*, vol. 338, Springer, 2009. 2

[9] Aude Genevay, Gabriel Peyré, and Marco Cuturi, "Learning generative models with sinkhorn divergences," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2018, pp. 1608–1617. 2

[10] Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-ichi Amari, Alain Trouvé, and Gabriel Peyré, "Interpolating between optimal transport and mmd using sinkhorn divergences," in *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR, 2019, pp. 2681–2690. 2

[11] David A Edwards, "On the kantorovich–rubinstein theorem," *Expositiones Mathematicae*, vol. 29, no. 4, pp. 387–398, 2011. 2

[12] Hoang Anh Just, Feiyang Kang, Tianhao Wang, Yi Zeng, Myeongseob Ko, Ming Jin, and Ruoxi Jia, "Lava: Data valuation without pre-specified learning algorithms," in *11th International Conference on Learning Representations, ICLR*, 2023, p. to appear. 2

[13] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei, "Scaling laws for neural language models," *arXiv preprint arXiv:2001.08361*, 2020. 2

[14] 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)*, Apr. 2015, pp. 5206–5210. 2

[15] Anthony Rousseau, Paul Deléglise, and Yannick Estève, "Enhancing the TED-LIUM Corpus with Selected Data for Language Modeling and More TED Talks," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, May 2014, pp. 3935–3939, European Language Resources Association (ELRA). 2

[16] "EMFORMER_RNNT_BASE_LIBRISPEECH — Torchaudio 2.1.0.dev20230717 documentation," https://pytorch.org/audio/main/generated/torchaudio.pipelines.EMFORMER_RNNT_BASE_LIBRISPEECH.html#torchaudio.pipelines.EMFORMER_RNNT_BASE_LIBRISPEECH. 3

[17] Yangyang Shi, Yongqiang Wang, Chunyang Wu, Ching-Feng Yeh, Julian Chan, Frank Zhang, Duc Le, and Mike Seltzer, "Emformer: Efficient Memory Transformer Based Acoustic Model for Low Latency Streaming Speech Recognition," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, June 2021, pp. 6783–6787. 3

[18] "Emformer — Torchaudio 2.1.0.dev20230627 documentation," https://pytorch.org/audio/main/generated/torchaudio.models.Emformer.html. 3

[19] Yao-Yuan Yang, Moto Hira, Zhaoheng Ni, Anjali Chourdia, Artyom Astafurov, Caroline Chen, Ching-Feng Yeh, Christian Puhrsch, David Pollack, Dmitriy Genzel, Donny Greenberg, Edward Z. Yang, Jason Lian, Jay Mahadeokar, Jeff Hwang, Ji Chen, Peter Goldsborough, Prabhat Roy, Sean Narenthiran, Shinji Watanabe, Soumith Chintala, Vincent Quenneville-Bélair, and Yangyang Shi, "Torchaudio: Building blocks for audio and speech processing," *arXiv preprint arXiv:2110.15018*, 2021. 7

[20] 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*, vol. 33, pp. 12449–12460, 2020. 8

[21] Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z. Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, Léo Gautheron, Nathalie T.H. Gayraud, Hicham Janati, Alain Rakotomamonjy, Ievgen Redko, Antoine Rolet, Antony Schutz, Vivien Seguy, Danica J. Sutherland, Romain Tavenard, Alexander Tong, and Titouan Vayer, "Pot: Python optimal transport," *Journal of Machine Learning Research*, vol. 22, no. 78, pp. 1–8, 2021. 8

Appendices

Appendix A Related Work

Existing research on selecting training data for ASR primarily focuses on training within homogenous data from a single source, rather than optimally mixing heterogeneous data from different sources to train general ASR models.

Current methods often employ heuristic selection criteria to identify informative training data. For instance, Wu et al.[4] utilized the maximum entropy principle to select training data that contains the most informative examples, aiming to encourage the phoneme distribution approximating a uniform distribution. Submodular optimization [5] has also been commonly employed in data selection/distillation for ASR tasks under a variety of formulations. For example, in a recent work [6], Park et al. proposed a submodular function incorporating a loss ratio to determine the importance of each data sample. While these approaches contribute to data selection within a single data type, our study focuses on the selection and mixing of different types of ASR data for optimal general ASR model training.

Appendix B Algorithm

Algorithm 1: `projektor` performance predictor (adapted from [7])

In : Pilot Datasets $D_1^{pi}, D_2^{pi}, \ldots, D_m^{pi}$; Query Data Budget N; Query Mixing Ratio $\mathbf{p}$; 0-Data Scale Size N_0; 1-Data Scale Size N_1; Learning Algorithm $\mathcal{A}$; Performance Metric Function $\mathcal{L}(\cdot, D^{val})$; OT Distance Function $OT(\cdot, D^{val})$.

Out : Projected Model Performance $\rightarrow [0, 1]$.

1 $\texttt{P} \leftarrow$ Generate mixing ratios

2 $\texttt{DT}_0, \texttt{DT}_1 \leftarrow$ Initialize empty lists to store OT distances

3 $\texttt{L}_0, \texttt{L}_1 \leftarrow$ Initialize empty lists to store performance values

4 **for** *Mixing Ratio* $\mathbf{p_i}$ *in* $\texttt{P}$ **do**

5 $\quad S_0, S_1 = \mathcal{D}(N_0, \mathbf{p_i}), \mathcal{D}(N_1, \mathbf{p_i})$ newly composed datasets of size N_0, N_1

6 $\quad \texttt{DT}_0 \leftarrow$ append $OT(S_0, D^{val})$ Optimal Transport distance between S_0 and D^{val}

7 $\quad \texttt{DT}_1 \leftarrow$ append $OT(S_1, D^{val})$ Optimal Transport distance between S_1 and D^{val}

8 $\quad \texttt{L}_0 \leftarrow$ append $\mathcal{L}(\mathcal{A}(S_0), D^{val})$ Performance of a model trained on S_0

9 $\quad \texttt{L}_1 \leftarrow$ append $\mathcal{L}(\mathcal{A}(S_1), D^{val})$ Performance of a model trained on S_1

10 $\hat{\mathcal{L}}(\mathcal{A}(\mathcal{D}(N_0, \cdot)), D^{\text{val}}) \leftarrow$ Fit the function from Eq. in Section 2.1 with $((\texttt{P}, \texttt{DT}_0), \texttt{L}_0)$

11 $\hat{\mathcal{L}}(\mathcal{A}(\mathcal{D}(N_1, \cdot)), D^{\text{val}}) \leftarrow$ Fit the function from Eq. in Section 2.1 with $((\texttt{P}, \texttt{DT}_1), \texttt{L}_1)$

12 $\hat{\mathcal{L}}(\mathcal{A}(\mathcal{D}(N, \mathbf{p})); D^{\text{val}}) \leftarrow$ Project performance by substituting $\hat{\mathcal{L}}(\mathcal{A}(\mathcal{D}(N_0, \mathbf{p})), D^{\text{val}})$ and $\hat{\mathcal{L}}(\mathcal{A}(\mathcal{D}(N_1, \mathbf{p})), D^{\text{val}})$ into Eq. 1

13 return $\hat{\mathcal{L}}(\mathcal{A}(\mathcal{D}(N, \mathbf{p}); D^{\text{val}})$

To fit the function that predicts the model performance for different data mixtures with Equation from Section 2.1 (Lines 10-11, Algorithm 1), we first need to generate training data for two data scales N_0 and N_1 (Lines 4-7 for generating OT distances and Lines 8-9 for generating model performance values). Then, we apply the neural scaling law in Equation 1 to project the performance onto a larger target data size, N (Line 12). We will make our code public after the review process is over.

Appendix C Experimental Details

We followed the official TorchAudio implementation for fine-tuning the LibriSpeech (LS) pretrained Emformer RNN-T (with LS SOTA performance) [19]. All our experiments are conducted on a single 48GB A6000 GPU. While model fine-tuning requires 2-3 GPU hours, our optimal transport distance can be efficiently computed within 2-3 seconds.

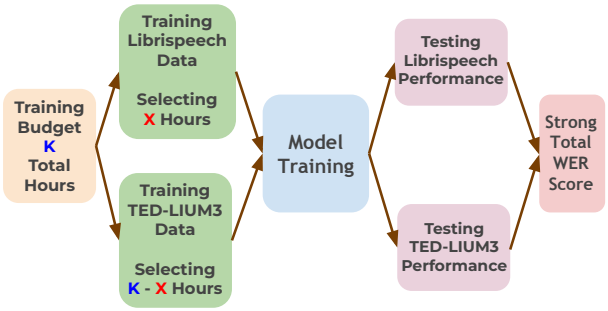

Figure 2: Problem visualization of training data selection. Given K hours of training data budget, the goal is to split the budget into training data sources to maximize model performance on all tasks combined (e.g., average WER score).

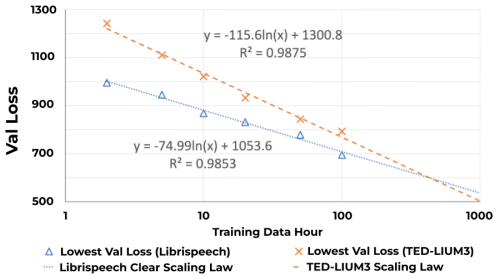

Figure 3: Comparing the scaling law between validation loss and training data size for the LibriSpeech and TED-LIUM3 training datasets.

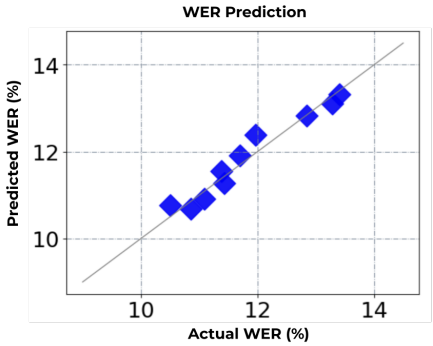

Figure 4: Actual WER vs Predicted WER, showing alignment of fitting OT to predict WER.

We verify that the fitted function based on optimal transport in Equations from Section 2.1 can effectively predict the model's performance, i.e., average WER in our case. Indeed, we visually observe in Figure 4 that our predicted WER values align well with the actual WER performance.

Optimal Transport Computation. To compute the OT distances of the audio clips, we use the Wav2Vec2 [20] as the embedding space and measure the distances between the extracted features averaged over attention layers. The optimal transport distance is computed with the help of the open-source Python library, POT: Python Optimal Transport [21].

Additional explanation of results. In our experiments, we want to emphasize the effectiveness of our method for selecting a fine-tuning data ratio and compare it with the baseline method. We show that our data ratios are non-obvious and outperform the baseline selection by improving the model's WER performance.

Orthogonal to our focus, we observe a non-trivial behavior of fine-tuned models when increasing the budget. We conjecture that the addition of TED-LIUM3 fine-tuning data to the LibriSpeech data causes the LibriSpeech SOTA performance of the pretrained model to weaken since the TED-LIUM3 domain naturally diverges from the LS domain. With more TED-LIUM3 data, the model weights optimized for LibriSpeech will start to degrade. To mitigate this problem, we can lower the learning rate, which can decrease the overfitting and the catastrophic forgetting effects when fine-tuning the model.

We hope our work can attract the community to further explore and develop automatic speech recognition models on multiple, diverse tasks by incorporating strong data selection methods.

Budget (Hr)	Method	Train Data L:T3 (Hr)	Test WER (%) L Clean	T3	L Other
40	Base	20:20	5.8	18.0	13.8
	Ours	6:34	6.0	14.6	13.5
50	Base	25:25	4.9	18.1	12.0
	Ours	8:42	5.0	13.9	12.0
100	Base	50:50	6.8	17.2	15.4
	Ours	18:82	6.5	14.2	15.2

Table 3: Comparison between the heuristic selection and the OT-guided selection under different fine-tuning data budgets for the fine-tuning tasks. L and T3 represent LibriSpeech and TED-LIUM3 data, respectively. Evaluation on Test LibriSpeech Clean (L Clean), Test TED-LIUM3 (T3), and Test LibriSpeech Other (L Other), respectively. Preliminary Results.