

Model-Centric Data Selection: Refining End-to-End Speech Recognition

Mengkedalai¹, Yan Meng¹ and Péter Mihajlik^{1,2}

¹*Department of Telecommunications and Media Informatics,
Faculty of Electrical Engineering and Informatics,
Budapest University of Technology and Economics Budapest, Hungary*

²*Hungarian Research Centre for Linguistics,
Budapest, Hungary*

kedalai.meng@edu.bme.hu, yan.meng@edu.bme.hu, mihajlik.peter@vik.bme.hu

Abstract—Data selection can be an important step in pre-processing datasets for Automatic Speech Recognition (ASR) – still its application is not general. In order to handle potential labeling errors and other anomalies in the dataset, we introduced a simple model-centric speech data selection strategy. It discards samples in the dataset that is difficult to recognize by the model, and use a restricted dataset to retrain the model. This technique improved the recognition accuracy of Hungarian ASR both on the BEA-Base and Common Voice (CV) datasets by using the Conformer model architecture. The proposed approach achieved a consistent relative improvement in terms of both Character and Word Error Rates (CER, WER), up to (3%, 2.5%).

Index Terms—Data selection, Dataset Optimization, Machine learning, Automatic Speech Recognition

I. INTRODUCTION

Traditional data selection may refer to data de-weighting, filling in missing values, data standardization and normalization, error and outlier handling, data formatting, data type conversion, data validation, data disaggregation, etc. [1] [2]. These steps are designed to improve the accuracy and completeness of data and are applicable to different types and sources of data.

Conventional data selection methods, however, face limitations when confronted with the challenges posed by large-scale, intricate datasets. These methods often lack the efficiency required for seamless handling of complexity and exhibit a deficiency in automation and context-specific adaptability crucial for contemporary machine learning endeavors. Consequently, in response to the evolving demands of ASR and Machine Learning (ML), there has been a transition from conventional singular approaches to a composite methodology seamlessly integrated into the machine learning workflow. This paradigm shift underscores the imperative that data selection strategies must be intricately connected to the specific objectives of the machine learning application. This alignment ensures that data pre-processing is congruent with the training objectives of the model, ultimately leading to an enhancement in model performance.

In the field of machine learning, the selection of training data significantly impacts the performance of models, especially in scenarios involving small models [3]. Since small models typically have fewer parameters and less complexity,

they are more susceptible to being misled by inaccurately labeled data. This misguidance not only reduces the model's accuracy but also weakens its adaptability and robustness in various situations.

To date, researchers have developed various data selection methods to tackle these challenges. Expanding upon traditional data selecting methodologies, the work of Felix Neutatz et al. explores the paradigm shift from pre-ML data selecting to a more integrated approach in selecting for ML [4]. This approach emphasizes the alignment of data selecting processes with the specific objectives and requirements of ML applications, advocating for a holistic and application-driven perspective in data selecting practices.

Some powerful data selection tools have been proposed by Ki Hyun Tae et al. MLselect, a comprehensive data selecting framework [5] innovatively combines traditional data selecting, unfairness mitigation, and data sanitization to enhance the accuracy and fairness of machine learning models. This framework exemplifies a significant step in big data-AI integration, showcasing the importance of preprocessing in developing robust models.

Each of these methods has its own characteristics, such as identifying and correcting erroneous labels through algorithms or using statistical techniques to filter out anomalous data points. They have demonstrated significant effectiveness in practice, aiding in the enhancement of many machine learning models.

However, despite the achievements of these data selection methods, we believe there is still room for exploration in this field. In particular, for data with specific patterns or structures, traditional methods might not be able to effectively identify and process them. Therefore, we propose a simple model-centric data selection method for Automatic Speech Recognition (ASR), based on the characteristics and learning capabilities of the trained model itself. To the best of our knowledge, this method is not typically used in ASR today.

The core of our method lies in utilizing the model itself to identify data points that are challenging for it to learn. We hypothesize that if some data is difficult for the model to learn or understand, then these data points might be erroneous, anomalous, or unrepresentative. Based on this concept, our

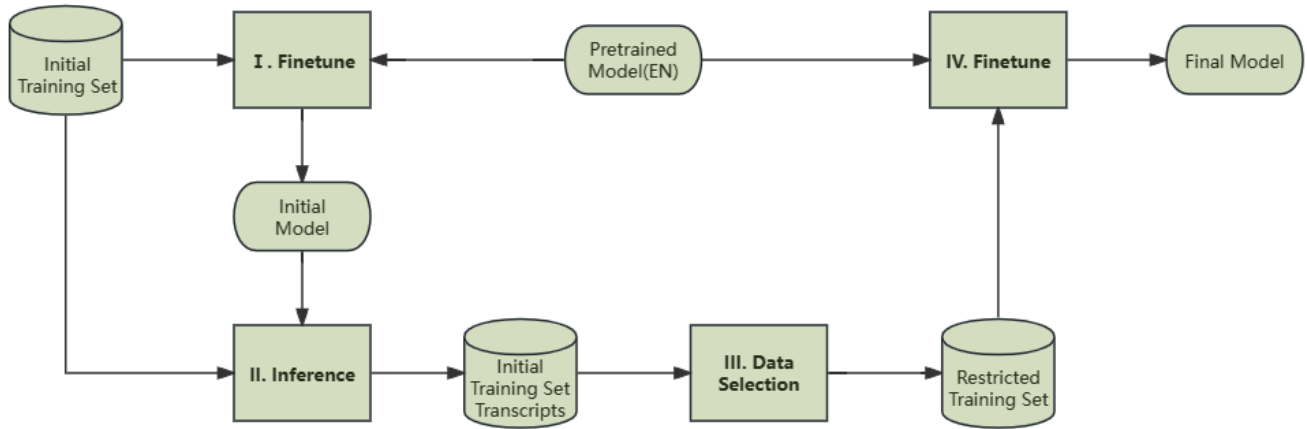


Fig. 1. Data Selection Method Flow Chart. The process involves fine-tuning a pre-trained English model on the initial training set (train-114 from BEA-Base [6]), using it to filter out 100% error data based on transcript prediction WER accuracy, and then re-fine-tuning on the selected (train-114-selected) dataset to yield the final improved model.

method does not attempt to correct these data points, but to discard them from the training set. This approach allows us to create a more refined and higher-quality dataset.

The experimental results are encouraging. We found that Conformer [7] models trained with BEA-Base [6] dataset selected using this method not only improved in accuracy but also exhibited better robustness when dealing with speech data having different acoustic features (various background noises, less controlled speakers, etc.). This discovery indicates that data selection strategies can indeed enhance the performance of small models in complex and challenging tasks. This provides a new perspective on data pre-processing and model training.

II. DATA SELECTION METHOD

Figure 1 shows the flow-chart of our data selection method. The method can be divided into four main steps (I. Finetune, II. Inference, III. Data Selection, IV. Finetune) and three phases (Selection-model Generation, Data Filtering and Post-Select Fine-Tuning Phase).

- **Selection-model Generation Phase**

Initially, preliminary training is conducted using the original training dataset. The goal is to enable the model to learn and master most of the content in the training set. Specifically, during the training process, the model’s predictions should align with the content of the training set. It is crucial not to over-train the model in this phase, to avoid mislearning content that is difficult to learn.

- **Data Filtering Phase**

During this data filtering phase, it is often found that some parts of the data are still unlearnable by the model, indicated by a 100% word error rate. Therefore, the principle of data selection is to use the finetuned initial model to test the initial training dataset (train-114). Find all the data with a 100% word error rate in the transcript

generated by the testing phase. After removing these data, the resulting dataset is the final required restricted training set. At this stage, the Word Error Rate (WER) is calculated using a greedy method.

- **Post-Select Fine-Tuning Phase**

Data identified as difficult to learn or with a high word error rate during the data filtering phase are removed from the original training dataset, forming a selected dataset named train-selected. This refined dataset excludes all data points that are difficult to learn or have high word error rate, thereby providing a higher quality data environment for model training.

Through these three phases, we proactively discarded the data that the model had difficulty learning from, and retrained or fine-tuned the model with a new, selected dataset, expecting to obtain better results on the new test dataset. This data selection method can effectively improve the quality of data, thereby enhancing the performance and robustness of the model in practical applications.

III. DATASET

For the current experiment, the BEA-Base dataset were selected. BEA-Base is an openly available Hungarian language dataset, includes spontaneous and repetitive speech, specifically designed for Automatic Speech Recognition (ASR) task. This dataset has already been utilized in various studies [8], such as those involving end-to-end models like Wav2Vec2 [9] and Conformer. These investigations have demonstrated the dataset’s viability and reliability for training ASR models.

Additionally, to enhance the credibility of the test results, an external test set, Common Voice (CV) [10] was employed. This dataset has different characteristics. A significant aspect of our data selection approach is the evaluation of model performance not only on the native test dataset but also on this external dataset. BEA-spont-eval is completely independent

TABLE I
MAIN CHARACTERISTICS OF DATASETS USED IN THE EXPERIMENTS.

	train-114	BEA-Base dev-spont	eval-spont	CV test
Length [hours]	67.95	4.02	4.91	6.8
Num of speakers	114	10	16	220
Num of segments	69 176	4 893	5 693	4 871
Num of characters	3.1M	154 994	197 738	250 709
Num of words	0.56M	27 939	35 178	35 485

from the training/validation sets. Therefore, Improvement in model performance on both the BEA-Base and Common Voice dataset would substantiate the robustness and reliability of the method.

In our experiments, train-114 was used as the training set, dev-spont as the validation set, eval-spont and Common Voice (v12.0) as the test set. The details of the BEA-Base and Common Voice datasets are presented in Table I.

IV. EXPERIMENT

All experiments are conducted on NVIDIA A6000 graphics cards to maintain a consistent baseline across different tests. Comparative experiments follow suit on the same type of graphics card. The NVIDIA NeMo toolkit v1.6.2¹, featuring a range of speech-related models such as Conformer, Transformer [11], is employed in our experiments. The model of choice is the Conformer model, recognized as one of the most prevalent models in automatic speech recognition. We use the English Conformer model provided by Nvidia as the pre-trained model in our experiments. Within the Nemo framework, Conformer is available in four sizes: Small (13M)², Medium (30M)³, Large (120M)⁴, and X Large (0.6B)⁵. Due to computational resource considerations, the experiments are limited to the Small and Medium models. The experimental configuration includes a regime of 200 epochs, a learning rate set at 0.002, batch-size set as 32 and CTC loss [12] as the chosen loss function.

A. Selecting-model Generation Phase

During this phase, training is conducted using the original BEA-Base training set (train-114), with the dev-spont dataset serving as the validation set. The training is planned for 50 epochs and is halted once a rapid decrease in the Word Error Rate (WER) on the validation set is no longer observed. This approach is designed to ensure that data of lower quality is not over-learned, thereby facilitating the effectiveness of the subsequent data exclusion process. After many training

¹<https://github.com/NVIDIA/NeMo/tree/v1.6.2>

²https://catalog.ngc.nvidia.com/orgs/nvidia/teams/nemo/models/stt_en_conformer_ctc_small

³https://catalog.ngc.nvidia.com/orgs/nvidia/teams/nemo/models/stt_en_conformer_ctc_medium

⁴https://catalog.ngc.nvidia.com/orgs/nvidia/teams/nemo/models/stt_en_conformer_ctc_large

⁵https://catalog.ngc.nvidia.com/orgs/nvidia/teams/nemo/models/stt_en_conformer_ctc_xlarge

attempts, we found that a better selecting model can be obtained when choosing 50 epochs.

B. Data Filtering Phase

Upon completing the initial training, a selecting model is generated. This model is utilized to test the original training dataset (train-114). The output texts generated by the model are compared against the original labels to calculate the Word Error Rate (WER) and Character Error Rate (CER). Subsequently, the accuracy of each predicted text is computed. Texts with a persisting word error rate of 100% are identified, and the corresponding training data are removed from the original training dataset (train-114). This process results in the creation of a new, refined training dataset, designated as train-114-selected. The comparison between the selected data and the original data shows in Table II. This methodology ensures the exclusion of potentially misleading data that could adversely impact model training and performance.

C. Post-Select Fine-Tuning Phase

Following the aforementioned procedure, a new training dataset, train-114-selected, is established. The validation set remains unchanged, utilizing the same dev-spont dataset as in the pre-training phase. For this round of training, the model and hyperparameters are maintained identically to those of the baseline, encompassing aspects such as learning rate (0.002) and the number of epochs (200). This approach ensures consistency and comparability in evaluating the efficacy of the data selection method on model performance.

D. Evaluation Phase

In assessing the performance of these models, a greedy decoding algorithm [13] was implemented at first. This algorithm, renowned for its computational efficiency, operates by selecting the most probable prediction at each sequential step, thereby facilitating the calculation of the Word Error Rate (WER) and the Character Error Rate (CER). Complementing this approach, a SentencePiece tokenizer, trained on the train-114 dataset with 128 unigram units, was utilized to construct an intricate subword 6-gram language model using the HGC-SPOK data corpus by using the KenLM [14] tool. This evaluation framework not only underscores the model’s real-time decoding efficiency but also provides a foundational analysis of its transcriptional accuracy in processing Hungarian speech.

E. Result Analysis

As a result of the above experiments, Table III and Table IV detail the Word Error Rate (WER) and Character Error Rate (CER) for the small model, while Table V and Table VI respectively showcase these metrics for the medium model. Each table also delineates the impact of incorporating a 6-gram language model on these performance indicators.

The findings reveal that the adoption of our model-centric data selection strategy significantly improved the performance of both small and medium models across tests on external and

TABLE II
CHARACTERISTICS COMPARISON BETWEEN DATASETS IN THE DATA FILTERING PHASE.

train-114	Length [hours]	Num of segments	Num of characters	Num of words
Initial	67.95	69 176	3 310 493	555 322
Small-Restricted	67.71	68 613	3 305 826	554 548
Small-Discarded	0.24	563	4 667	774
Medium-Restricted	67.74	68 950	3 305 043	554 485
Medium-Discarded	0.21	226	5 450	837

local datasets. Notably, in the case of the small model tested on the external CommonVoice dataset, there was a notable 3% relative decrease in CER, amounting to an absolute reduction of 0.58%. This shows the efficacy of the proposed strategy in enhancing the accuracy of speech recognition models.

TABLE III
WER (%) RESULTS WITH CONFORMER SMALL

Model	LM	BEA-Base eval-spont	CV test
Conformer-Small-Initial	-	25.42	46.66
Conformer-Small-Restricted	-	24.80	46.08
Conformer-Small-Initial	6-gram	22.06	41.49
Conformer-Small-Restricted	6-gram	21.68	40.09

TABLE IV
CER (%) RESULTS WITH CONFORMER SMALL

Model	LM	BEA-Base eval-spont	CV test
Conformer-Small-Initial	-	8.20	12.79
Conformer-Small-Restricted	-	7.96	12.50
Conformer-Small-Initial	6-gram	7.29	11.54
Conformer-Small-Restricted	6-gram	7.20	11.25

TABLE V
WER(%) RESULTS WITH CONFORMER MEDIUM

Model	LM	BEA-Base eval-spont	CV test
Conformer-Medium-Initial	-	23.47	42.58
Conformer-Medium-Restricted	-	23.49	42.14
Conformer-Medium-Initial	6-gram	20.74	38.57
Conformer-Medium-Restricted	6-gram	20.56	38.36

TABLE VI
CER(%) RESULTS WITH CONFORMER MEDIUM

Model	LM	BEA-Base eval-spont	CV test
Conformer-Medium-Initial	-	7.36	11.17
Conformer-Medium-Restricted	-	7.41	11.14
Conformer-Medium-Initial	6-gram	6.76	10.24
Conformer-Medium-Restricted	6-gram	6.72	10.17

V. CONCLUSION

This study introduces a simple but effective model-centric data selection strategy anchored in the intrinsic characteristics of the model itself. Experimental results conducted on

the Hungarian language dataset BEA-Base indicate that this methodology, when applied to small and medium-sized speech recognition models based on the Conformer architecture, can achieve notable performance enhancements, even with a reduced training data volume. This improvement is evident in both the homologous (eval-spont) dataset and the heterologous (CommonVoice) test set.

This study hypothesizes that these data might possess certain attributes that are detrimental to learning, potentially impacting the robustness of the models to a certain degree.

These findings offer a new perspective on the methodology of data selection in speech recognition. However, this approach has not yet been tested on a broader range of models and datasets. Future research endeavors will aim to apply this strategy to a more diverse set of models, including large and XL sized Conformers, thereby further exploring its general applicability and efficacy across various settings.

ACKNOWLEDGMENT

This research benefited greatly from the support provided by the Hungarian Linguistic Research Center in the development of the BEA-Base dataset. This work was supported partially by NKFIH-828- 2/2021(MILab), by the NVIDIA Academic Hardware Grant and by the NKFIH K143075 and K135038 projects of the NRDI Fund. Thanks are also extended to the Budapest University of Technology and Economics and NVIDIA Academic Hardware Grant for their vital contribution including but not limited to hardware support.

REFERENCES

- [1] S. García, J. Luengo, and F. Herrera, *Data preprocessing in data mining*. Springer, 2015, vol. 72.
- [2] G. Y. Lee, L. Alzamil, and B. Doskenov, "A survey on data cleaning methods for improved machine learning model performance," *arXiv preprint arXiv:2109.07127*, 2021.
- [3] G. Y. Lee, L. Alzamil, B. Doskenov, and A. Termehchy, "A survey on data cleaning methods for improved machine learning model performance," *arXiv preprint arXiv:2109.07127*, 2021.
- [4] F. Neutatz, B. Chen, Z. Abedjan, and E. Wu, "From cleaning before ml to cleaning for ml." *IEEE Data Eng. Bull.*, vol. 44, no. 1, pp. 24–41, 2021.
- [5] K. H. Tae, Y. Roh, Y. H. Oh, H. Kim, and S. E. Whang, "Data cleaning for accurate, fair, and robust models: A big data-ai integration approach," in *Proceedings of the 3rd International Workshop on Data Management for End-to-End Machine Learning*, 2019, pp. 1–4.
- [6] P. Mihajlik, A. Balog, T. E. Graczi, A. Kohari, B. Tarján, and K. Mady, "BEA-base: A benchmark for ASR of spontaneous Hungarian," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, Jun. 2022, pp. 1970–1977. [Online]. Available: <https://aclanthology.org/2022.lrec-1.211>

- [7] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu *et al.*, “Conformer: Convolution-augmented transformer for speech recognition,” *arXiv preprint arXiv:2005.08100*, 2020.
- [8] P. Mihajlik, M. S. Kádár, G. Dobsinszki, Y. Meng, M. Kedalai, J. Linke, T. Fegyó, and K. Mády, “What kind of multi-or cross-lingual pre-training is the most effective for a spontaneous, less-resourced asr task?” *SIGUL 2023-Satellite Workshop of Interspeech 2023*, 2023.
- [9] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, “wav2vec 2.0: A framework for self-supervised learning of speech representations,” *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.
- [10] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, “Common voice: A massively-multilingual speech corpus,” *arXiv preprint arXiv:1912.06670*, 2019.
- [11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” *Advances in neural information processing systems*, vol. 30, 2017.
- [12] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, “Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks,” in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 369–376.
- [13] Z. Zhang, S. Schwartz, L. Wagner, and W. Miller, “A greedy algorithm for aligning dna sequences,” *Journal of Computational biology*, vol. 7, no. 1-2, pp. 203–214, 2000.
- [14] K. Heafield, “Kenlm: Faster and smaller language model queries,” in *Proceedings of the sixth workshop on statistical machine translation*, 2011, pp. 187–197.