

A STUDY OF DATA SELECTION STRATEGIES FOR PRE-TRAINING SELF-SUPERVISED SPEECH MODELS

Ryan Whetten¹, Titouan Parcollet², Marco Dinarelli³, Yannick Estève¹

¹Laboratoire Informatique d'Avignon, Avignon Université, Avignon, France

²University of Cambridge, Cambridge, United Kingdom

³Laboratoire d'Informatique de Grenoble, Université Grenoble Alpes, Grenoble, France

ABSTRACT

Self-supervised learning (SSL) has transformed speech processing, yet its reliance on massive pre-training datasets remains a bottleneck. While robustness is often attributed to scale and diversity, the role of the data distribution is less understood. We systematically examine how curated subsets of pre-training data influence Automatic Speech Recognition (ASR) performance. Surprisingly, optimizing for acoustic, speaker, or linguistic diversity yields no clear improvements over random sampling. Instead, we find that prioritizing the longest utterances achieves superior ASR results while using only half the original dataset, reducing pre-training time by 24% on a large corpora. These findings suggest that for pre-training speech SSL models, data length is a more critical factor than either data diversity or overall data quantity for performance and efficiency, offering a new perspective for data selection strategies in SSL speech processing.

Index Terms— self-supervised learning, speech, automatic speech recognition, data selection

1. INTRODUCTION

Self-supervised learning (SSL) is a framework for training neural networks where pseudo-targets are generated from the input data itself. A neural network model is trained to using these targets in a stage called pre-training, and then can be subsequently fine-tuned using human labeled data for a particular downstream task. SSL in speech processing has led to state of the art performance in various tasks such as automatic speech recognition (ASR), taking advantage of massive amounts of unlabeled data to train a neural network to produce high quality speech representations [1].

To take advantage of this data, SSL speech models are large, typically composed of 12 or more Transformer based layers [2]. While these frameworks have led to high performing models, they require a significant amount of resources in

terms of the amount of memory, time, and data to train these large neural networks.

To improve the efficiency of SSL speech models, researchers have focused on improving model architectures [3, 4], simplifying or adjusting pre-training objectives [5, 6, 4], and proposing efficient evaluation methods [7]. While these methods reduce memory requirements and the time it takes to train and evaluate, high performing SSL speech models such as Google USM [8] and Xeus [9] are still pre-trained on the order of millions of hours of speech data (corresponding centuries of audio), and the effects of changes in the composition of pre-training data has been less explored.

To address this gap, we investigate the impact of simple unsupervised data selection strategies for pre-training. Our first set of methods samples data to promote diversity across acoustic, speaker, or linguistic features, while our second set focuses on sampling based on utterance length.

Unexpectedly, our results show that enforcing diversity in the pre-training data—whether acoustic, speaker, or linguistic—yields no significant improvement in downstream Automatic Speech Recognition (ASR) performance. In contrast, simply selecting the longest utterances for pre-training achieves superior ASR performance while requiring only half of the original dataset and reducing pre-training time by 24% on a large-scale corpus. We make code used in the experiments public for reproducibility¹.

2. RELATED WORK

In this section, we provide an overview of prior work on improving the efficiency of SSL for speech models, as well as studies on data selection for speech processing.

2.1. Efficiency for SSL speech models

Research on addressing the inefficiency of SSL speech models has primarily focused on three directions: modifications to the SSL training framework, architectural innovations, and alternative evaluation methodologies.

This work received funding from the French ANR E-SSL project (N°ANR-22-CE23-0013) and used HPC resources from GENCI-IDRIS (AD011014732 and A0131013821)

¹link to code will be provided for camera ready version

For improving the SSL framework itself, researchers have explored simplifying the generation of pseudo-targets and objectives. For example, BEST-RQ [6] employs a random-projection quantizer and a simple cross-entropy loss, in contrast to wav2vec 2.0 [2], which uses a learned codebook along with a combination of contrastive and diversity losses. Other modifications involve leveraging pre-trained or ASR fine-tuned models to generate higher-quality embeddings, which serve as training targets and yield both performance and efficiency gains [5]. Additionally, practical engineering improvements to the SSL framework have been shown to significantly reduce memory usage and training time such as computing targets on the fly to reduce storage memory as and not encoding the masked tokens [5, 4].

Architectural changes have also played a key role in improving efficiency. In models such as BEST-RQ [6] and Whisper [10], the acoustic feature extractor has been simplified, using mel-filterbanks with only two convolutional layers instead of raw audio inputs with seven convolutional layers. Other work has investigated replacing multi-head self-attention blocks with more efficient alternatives [3]. In parallel, knowledge distillation methods have been explored to reduce model size while maintaining strong performance [11].

Finally, researchers have developed more efficient evaluation approaches for SSL speech models. Recent work demonstrates that the rank of embeddings produced by an SSL model can serve as an indicator of pre-training quality [7, 12]. While these lines of research address efficiency from multiple perspectives, relatively little attention has been given to improving data efficiency in pre-training.

2.2. Data selection for speech models

While substantial work has been done on data selection for speech models, most prior research has focused on active learning. Active learning involves identifying the most informative data for labeling, with the goal of minimizing annotation costs. Typically, this is achieved by using uncertainty or confidence measures, where utterances with the lowest confidence are selected for labeling and then used to train a fully supervised ASR model [13]. In contrast, our work focuses on understanding what constitutes an effective pre-training dataset for SSL speech models, rather than on active learning.

For SSL-specific data selection, researchers have proposed methods that select pre-training data most similar to the target downstream domain [14]. One such approach trains language models on quantized targets of speech from both the pre-training pool and the target domain dataset, and then computes a normalized difference in log-probability between the two models. This score serves as an indicator of how similar a pre-training utterance is to the target domain. While effective for domain-matching, this line of work is distinct from ours, as we instead investigate the general properties of pre-training

data that lead to strong overall performance, independent of any specific target domain.

Most closely related to our study is the work of Berrebbi et al. [15], who analyze the effects of varying the number of speakers and the amount of audio in pre-training datasets. Their results highlight the importance of dataset size for SSL, though their experiments were limited to under one thousand hours of pre-training data and require speaker labels. By contrast, our study explores methods that do not require speaker labels, consider features beyond speaker information, and scale up to 25 thousand hours of audio.

3. METHODS

In this section, we describe the dataset used, the different methods for data selection, and the training environment.

3.1. Data

For our experiments, we use the Loquacious dataset [16], which combines commercially usable English speech corpora. We select this dataset because it is both large-scale and challenging, containing 25,000 hours of diverse speech, including read, spontaneous, conversational, clean, and noisy recordings.

This diversity extends across all splits (train, development, and test), which is crucial for our study. Unlike prior work that focuses on selecting pre-training data to closely match the fine-tuning domain [14], our goal is to investigate which types of pre-training data yield the best overall performance across various speech conditions. For pre-training, we experiment with the medium split (2,500 hours) and the large split (25,000 hours). For fine-tuning Automatic Speech Recognition (ASR) models, we use the small split (250 hours).

3.2. Data selection methods

We consider two baselines. The first, more challenging baseline uses the entirety of the data for pre-training, while the second is a random selection of 50% of the data. We refer to these as the *all* and *random* baselines, respectively.

The next set of methods is based on diversity sampling across different feature types: acoustic, speaker, and linguistic. To capture basic acoustic characteristics of each utterance, we compute the first 13 Mel-Frequency Cepstral Coefficients (MFCCs) along with their first- and second-order derivatives. We then average across frames, producing a 39-dimensional vector per utterance. For speaker features, we extract 256-dimensional speaker embeddings using the WeSpeaker model [17] with pyannote [18], which encodes speaker-specific characteristics from the audio. For linguistic features, we use the SENSE model (Shared Embedding for N-lingual Speech and tExt) [19], a high-performing,

language-agnostic, open-source model that generates semantic embeddings from speech and text. This produces a 1024-dimensional vector representing the semantic content of each utterance.

To obtain diverse and representative samples, we apply stratified sampling guided by k-means clustering. After clustering the dataset using one of the feature representations, we sample a total of N utterances with balanced coverage across all k clusters. To avoid underrepresenting small clusters, we first ensure that each cluster contributes at least one data point. We then iteratively and evenly sample additional utterances from clusters until reaching the target size N .

This approach prevents larger clusters from dominating the sample and guarantees coverage of smaller ones. We set N to approximately 50% of the total utterances, adjusting slightly per method to keep the total audio duration comparable to the random baseline. For the Loquacious medium and large datasets, we use $k = 150$ and $k = 200$, respectively.

Finally, we explore two methods based on utterance length. The first simply selects the longest 50% of utterances, denoted as *length*. The second combines length and speaker diversity: we use the k-means clusters on speaker embeddings from the diversity sampling method, then instead of sampling randomly within clusters, we select the longest utterances from each cluster, denoted as *speaker+len*. Due to limited resources, we only tested this feature combination, as preliminary results showed speaker diversity sampling was slightly more effective. We maintain a total audio duration approximately equal to the random baseline.

3.3. Training settings

For pre-training, we use the BEST-RQ framework and architecture, which employs a random-projection quantizer to generate pseudo-targets [6] and conformer layers as the model backbone. We adopt rotary positional embeddings [20] and configure the model with 12 layers, a hidden size of 640, 8 attention heads, and a feed-forward network dimension of 2048, resulting in approximately 100M parameters. We select BEST-RQ due to its efficiency [21], strong performance [6, 8], and its open-source implementation in SpeechBrain [21, 22]. For fine-tuning for speech recognition, we adopt the fine-tuning SpeechBrain recipe for BEST-RQ which applies a feed-forward neural network with CTC loss. We set vocabulary size is set to 1,024 byte pair encoding tokens following [16].

We use NVIDIA A100 GPUs. To estimate total GPU time, we measure the duration required to complete 50,000 training steps on a system with 8 A100s, and then extrapolate to the full 200,000 steps. We apply dynamic batching, which groups utterances of similar lengths and adjusts batch sizes to maintain a consistent amount of audio per batch. The maximum duration per batch is capped at 800 seconds per GPU, yielding a total batch size of approximately 1.77 hours of audio across 8 GPUs.

Table 1. Automatic Speech Recognition (ASR) results with different pre-training subsets from the Loquacious medium and large splits. All models are fine-tuned on the small split. The *Time* and *Data* columns report the amount GPU hours and pre-training audio used.

Pre-train Split	WER	WER	Time	Data
Selection method	<i>dev</i>	<i>test</i>	(hrs)	(hrs)
Medium Split				
All	18.48	19.39	195	2.50 k
Random	18.80	19.82	202	1.25 k
MFCC	19.05	19.98	202	1.24 k
Speaker	18.78	19.72	204	1.23 k
SENSE	19.23	20.39	199	1.24 k
Length	18.08	19.02*†	202	1.25 k
Speaker+Len	18.16	18.97 *†	205	1.24 k
Large Split				
All	17.12	18.08	263	25.2 k
Random	17.53	18.54	214	12.6 k
MFCC	17.49	18.39	221	12.6 k
Speaker	17.26	17.97*	214	12.6 k
SENSE	17.67	18.42	385	12.6 k
Length	16.76	17.77*†	200	12.6 k
Speaker+Len	16.60	17.42 *†	201	12.5 k
No pre-train [16]	22.30	23.80	-	-

*Significantly lower WER from *random* baseline ($p < 0.05$)

†Significantly lower WER from *all* baseline ($p < 0.05$)

We also employ dynamic chunking, which applies dynamic masks to both the attention mechanism [23] and the convolutional block [24]. This limits left and right context access in a randomized manner, effectively simulating both streaming and non-streaming conditions during training. While our study does not focus on streaming, prior work has shown that pre-training and fine-tuning with dynamic chunking improves performance for both streaming and non-streaming ASR [25]. All reported results are obtained in a non-streaming setting.

We train all models for 200,000 steps with identical hyperparameters, varying only the pre-training subsets described in Section 3.2. To assess statistical reliability, we apply bootstrapping with 1,000 samples to estimate 95% confidence intervals.

4. RESULTS

The results are presented in Table 1. Overall, diversity-based sampling methods did not yield significant improvements over the either baseline. The only exception is the speaker-based method on the large split, which achieved a word error rate of 17.97 on the test set—significantly better than the random baseline (18.54) but not the *all* baseline (18.08).

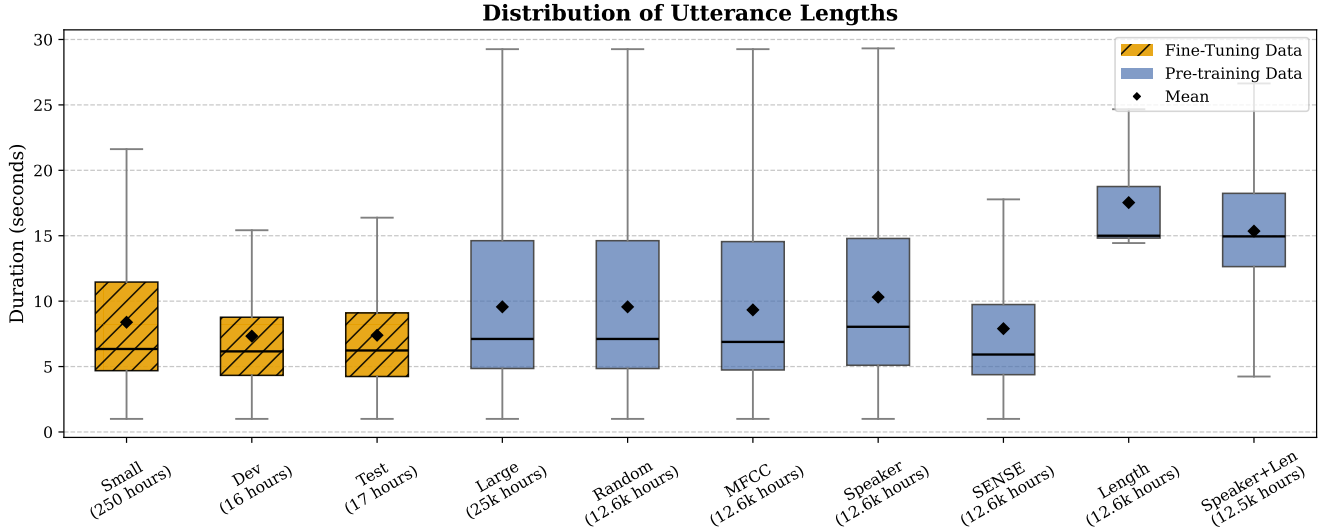


Fig. 1. Box-and-whisker plots of utterance length distributions across different dataset splits. The three plots on the left correspond to the fine-tuning split, while the remaining plots show the various pre-training subsets. Notably, the best performing models were pre-trained on subsets consisting of longer utterances (far right two plots), whose distributions differ most from the fine-tuning data.

In contrast, utterance length-based methods consistently outperformed both baselines. On the medium split, selecting the longest utterances reduced the word error rate to 19.02 on the test set, while combining speaker diversity with utterance length further improved performance to 18.97. On the large split, these same methods reached 17.77 and 17.42 word error rates, respectively, both significantly better than both baselines. Although the combined method performed best overall, the gains over length-only sampling were relatively small.

In terms of efficiency, training times for the medium split were similar across all methods. However, on the large split, length-based sampling substantially reduced pre-training time. Compared to the *all* baseline, these methods trained about 24% faster. This improvement stems from the dynamic batching strategy: with longer utterances, batches contain fewer examples, which lowers per-step computational cost.

In the Loquacious dataset, prior experiments were conducted with models of similar size but using a conformer encoder-decoder architecture trained fully supervised on the small split [16]. While this setup is not directly comparable, it provides a useful reference: the 100M parameter supervised model trained on 250 hours of labeled data. Notably, all of our data selection methods outperform this baseline, despite relying on a less complex architecture. This demonstrates the benefits of self-supervised pre-training.

5. DISCUSSION AND CONCLUSION

From our results, we find that reducing the amount of pre-training data by sampling based on acoustic, speaker, or lin-

guistic features does not yield improvements over a random baseline. In contrast, selecting subsets of longer utterances consistently leads to lower WER, despite these subsets being most out-of-distribution relative to the fine-tuning data. As shown in Figure 1, the best-performing subsets contain utterances with median lengths around 15 seconds and mean lengths above 15 seconds, whereas other subsets fall between 5 and 10 seconds.

While further work is needed to understand the underlying cause of this effect as well as the applicability to other SSL speech models and datasets, our findings suggest that pre-training benefits from longer utterances even when fine-tuning data is dominated by shorter segments. This is potentially due to these utterances having longer richer contexts for learning or more challenging learning signals. In future work, we plan to address these limitations and extend this work to better understand what makes a good pre-training dataset.

In summary, we systematically studied simple unsupervised data selection methods for pre-training SSL speech models. We showed that enforcing diversity across acoustic, speaker, or linguistic features does not improve downstream ASR performance, whereas selecting the longest utterances achieves superior performance while using only half of the original dataset. Moreover, this approach reduces pre-training time by up to 24% on large datasets. These results highlight utterance length as a key factor in constructing efficient pre-training datasets for SSL speech models.

6. REFERENCES

- [1] A. Mohamed, H.-y. Lee, L. Borgholt, J. D. Havtorn, J. Edin, C. Igel, K. Kirchhoff, S.-W. Li, K. Livescu, L. Maaløe, et al., “Self-supervised speech representation learning: A review,” *IEEE Journal of Selected Topics in Signal Processing*, 2022.
- [2] 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. 12449–12460, 2020.
- [3] R. Whetten, T. Parcollet, A. Moumen, M. Dinarelli, and Y. Estève, “An analysis of linear complexity attention substitutes with best-rq,” in *2024 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2024, pp. 169–176.
- [4] A. Baevski, A. Babu, W.-N. Hsu, and M. Auli, “Efficient self-supervised learning with contextualized target representations for vision, speech and language,” in *International Conference on Machine Learning*. PMLR, 2023, pp. 1416–1429.
- [5] W. Chen, X. Chang, Y. Peng, Z. Ni, S. Maiti, and S. Watanabe, “Reducing barriers to self-supervised learning: Hubert pre-training with academic compute,” in *Interspeech 2023*, 2023, pp. 4404–4408.
- [6] C.-C. Chiu, J. Qin, Y. Zhang, J. Yu, and Y. Wu, “Self-supervised learning with random-projection quantizer for speech recognition,” in *Proceedings of the 39th International Conference on Machine Learning*, K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, Eds. 17–23 Jul 2022, vol. 162 of *Proceedings of Machine Learning Research*, pp. 3915–3924, PMLR.
- [7] R. Whetten, L. Maison, T. Parcollet, M. Dinarelli, and Y. Estève, “Towards Early Prediction of Self-Supervised Speech Model Performance,” in *Interspeech 2025*, 2025, pp. 1228–1232.
- [8] Y. Zhang, W. Han, J. Qin, Y. Wang, A. Bapna, Z. Chen, N. Chen, B. Li, V. Axelrod, G. Wang, et al., “Google usm: Scaling automatic speech recognition beyond 100 languages,” *arXiv preprint arXiv:2303.01037*, 2023.
- [9] W. Chen, W. Zhang, Y. Peng, X. Li, J. Tian, J. Shi, X. Chang, S. Maiti, K. Livescu, and S. Watanabe, “Towards robust speech representation learning for thousands of languages,” in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, Eds., Miami, Florida, USA, Nov. 2024, pp. 10205–10224, Association for Computational Linguistics.
- [10] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, “Robust speech recognition via large-scale weak supervision,” in *International conference on machine learning*. PMLR, 2023, pp. 28492–28518.
- [11] H. gun Chi, Z. Aldeneh, T. Likhomanenko, O. Rudovic, T. Higuchi, L.-W. Chen, S. Watanabe, and A. H. Abdelaziz, “DiceHuBERT: Distilling HuBERT with a Self-Supervised Learning Objective,” in *Interspeech 2025*, 2025, pp. 1218–1222.
- [12] Z. Aldeneh, V. Thilak, T. Higuchi, B.-J. Theobald, and T. Likhomanenko, “Towards automatic assessment of self-supervised speech models using rank,” in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, pp. 1–5.
- [13] K. Malhotra, S. Bansal, and S. Ganapathy, “Active learning methods for low resource end-to-end speech recognition,” in *Interspeech 2019*, 2019, pp. 2215–2219.
- [14] Zhiyun Lu and Yongqiang Wang and Yu Zhang and Wei Han and Zhehuai Chen and Parisa Haghani, “Unsupervised Data Selection via Discrete Speech Representation for ASR,” in *Interspeech 2022*, 2022, pp. 3393–3397.
- [15] D. Berrebbi, R. Collobert, N. Jaitly, and T. Likhomanenko, “More speaking or more speakers?,” in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.
- [16] T. Parcollet, Y. Tseng, S. Zhang, and R. C. van Dalen, “Loquacious Set: 25,000 Hours of Transcribed and Diverse English Speech Recognition Data for Research and Commercial Use,” in *Interspeech 2025*, 2025, pp. 4053–4057.
- [17] H. Wang, C. Liang, S. Wang, Z. Chen, B. Zhang, X. Xiang, Y. Deng, and Y. Qian, “Wespeaker: A research and production oriented speaker embedding learning toolkit,” in *ICASSP 2023, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.
- [18] H. Bredin, “pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe,” in *Proc. INTERSPEECH 2023*, 2023, pp. 1983–1987.
- [19] S. Mdhaffar, H. Elleuch, C. Chellaf, H. Nguyen, and Y. Estève, “Sense models: an open source solution for multilingual and multimodal semantic-based tasks,” 2025.
- [20] J. Su, M. Ahmed, Y. Lu, S. Pan, W. Bo, and Y. Liu, “Roformer: Enhanced transformer with rotary position embedding,” *Neurocomputing*, vol. 568, pp. 127063, 2024.
- [21] R. Whetten, T. Parcollet, M. Dinarelli, and Y. Estève, “Open implementation and study of best-rq for speech processing,” in *2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, 2024, pp. 460–464.
- [22] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong, J.-C. Chou, S.-L. Yeh, S.-W. Fu, C.-F. Liao, E. Rastorgueva, F. Grondin, W. Aris, H. Na, Y. Gao, R. D. Mori, and Y. Bengio, “SpeechBrain: A general-purpose speech toolkit,” 2021, arXiv:2106.04624.
- [23] B. Zhang, D. Wu, Z. Yao, X. Wang, F. Yu, C. Yang, L. Guo, Y. Hu, L. Xie, and X. Lei, “Unified streaming and non-streaming two-pass end-to-end model for speech recognition,” *arXiv preprint arXiv:2012.05481*, 2020.
- [24] X. Li, G. Huybrechts, S. Ronanki, J. Farris, and S. Bodapati, “Dynamic chunk convolution for unified streaming and non-streaming conformer asr,” in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.
- [25] J. Duret, S. Mdhaffar, G. Laperrière, R. Whetten, A. Galametz, C. Kobus, M.-C. Martin, J. Oleiwan, and Y. Estève, “In-domain ssl pre-training and streaming asr,” 2025.