Interspeech 2025
17-21 August 2025, Rotterdam, The Netherlands

Efficient Data Selection for Domain Adaptation of ASR Using Pseudo-Labels and Multi-Stage Filtering

Pradeep Rangappa[1], *Andrés Carofilis*[1], *Jeena Prakash*[3], *Shashi Kumar*[1,2], *Sergio Burdisso*[1], *Srikanth Madikeri*[4], *Esaú Villatoro-Tello*[1], *Bidisha Sharma*[3], *Petr Motlicek*[1,5], *Kadri Hacioglu*[3], *Shankar Venkatesan*[3], *Saurabh Vyas*[3], *Andreas Stolcke*[3]

[1]Idiap Research Institute , Switzerland, [3]Uniphore Systems, India & USA,
[2]EPFL, Switzerland, [4]University of Zurich, Switzerland, [5]Brno University, Czech Republic,
{pradeep.rangappa, andres.carofilis,shashi.kumar,esau.villatoro,petr.motlicek}@idiap.ch

Abstract

Fine-tuning pretrained ASR models for specific domains is challenging for small organizations with limited labeled data and computational resources. Here we explore different data selection pipelines and propose a robust approach that improves ASR adaptation by filtering pseudo-labels generated using Whisper (encoder-decoder) and Zipformer (transducer) models. Our approach integrates multiple selection strategies—including word error rate (WER) prediction, named entity recognition (NER), and character error rate (CER) analysis—to extract high-quality training segments. We evaluate our method on Whisper and Zipformer using a 7500-hour baseline, comparing it to a CER-based approach relying on hypotheses from three ASR systems. Fine-tuning on 7500 hours of pseudo-labeled call center data achieves 12.3% WER, while our filtering reduces the dataset to 100 hours (1.4%) with similar performance; a similar trend is observed on Fisher English.

Index Terms: speech recognition, data selection, whisper, zipformers

1. Introduction

Automatic speech recognition (ASR) systems have seen remarkable advancements over the past decade, transitioning from traditional hybrid models [1, 2, 3, 4] to state-of-the-art end-to-end architectures [5, 6, 7, 8, 9, 10, 11]. Notable models such as wav2vec 2.0, XLSR [12], Conformer [6, 13], and Zipformer [14] leverage large-scale speech data to enhance transcription accuracy. Despite these advancements, real-world ASR adaptation remains constrained by the availability of in-domain annotated data. For instance, in our setting, only a few hours of annotated in-domain data are available, compared to a few hundred hours of annotated out-of-domain data and 7500 hours of unlabeled in-domain data. Given computational constraints, utilizing the entire unlabeled dataset for adaptation is impractical.

To address this challenge, recent studies have explored the use of pseudo-labeled speech data for fine-tuning ASR models [15, 16], leveraging large quantities of unlabeled data generated in industrial pipelines. In real-world customer-agent interactions, only a small portion of the data is manually labeled due to the high cost of annotation. As a result, efficient data selection techniques are crucial for leveraging pseudo-labels and fine-tuning ASR models for improved performance [17].

We propose a robust data selection pipeline to optimize ASR adaptation by selecting high-quality training data based on three key heuristics: (1) low predicted WER, (2) high inter-ASR agreement measured via low CER, and (3) the presence of named entity (NE) classes. These heuristics are motivated by the need to mitigate error propagation, leverage the correlation between ASR agreement and transcript accuracy, and prioritize information-rich segments where NEs play a crucial role in domain-specific adaptation [18, 19].

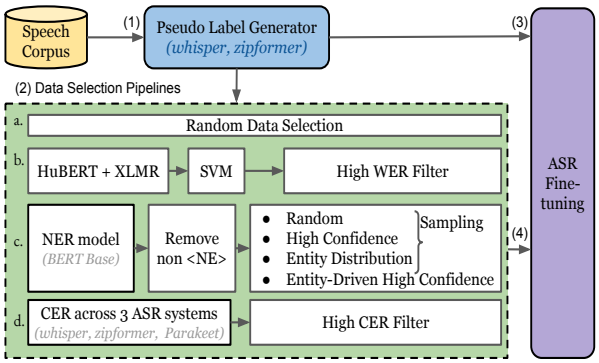

Figure 1: *The functional blocks of the ASR Domain Adaptation process: (1) Pseudo-label (PL) generation using Whisper and Zipformer, where the corresponding ASR model is fine-tuned with the generated pseudo-labels. (2) Data selection pipeline consisting of: (a) Random selection, (b) WER prediction with SVM using HuBERT and XLM-R, (c) NER-based selection using a BERT model, and (d) CER-based filtering, where Nemo's Parkeet model is specifically employed along with whisper and zipformer for selecting data in the CER-based filtering step. ASR fine-tuning is carried out on the (3) full training set using the generated pseudo-labels and (4) the filtered data.*

Our approach integrates WER classification using SVMs trained on HuBERT [20] and XLM-R [21] embeddings, CER-based filtering via majority voting across Whisper Medium, Zipformer, and Parakeet (fast conformer as encoder, RNNT as decoder) models[1], and NE-driven selection using Token2Vec and BERT-base models. In this work, we demonstrate that a tiny, carefully selected subset of pseudo-labeled data—comprising only 1 to 5% of the full dataset—can match or exceed the performance achieved by fine-tuning on the entire dataset when guided by the right selection criteria. By integrating multiple data selection strategies, we ensure computational feasibility while maximizing ASR accuracy. Our results offer empirical insights into the most effective data selection methods for ASR adaptation, providing practical guidance for future model fine-tuning efforts. Furthermore, our dynamic data selection pipeline adapts to evolving acoustic and lexical properties, optimizing both efficiency and accuracy in ASR fine-tuning.

[1]https://huggingface.co/nvidia/parakeet-rnnt-1.1b

10.21437/Interspeech.2025-2580

This scalable approach is particularly beneficial for real-world ASR applications, where manual annotation and computational resources are limited.

The remainder of the paper is structured as follows: Section 2 describes the proposed data selection pipeline in detail. Sections 3 and 4 discuss the experimental setup and results. Finally, Section 5 summarizes the conclusions and suggests future directions.

2. ASR Domain Adaptation on Pseudo Labels

Our proposed ASR domain adaptation approach, illustrated in Figure 1, involves generating pseudo-labels using Whisper and Zipformer models. Initially, voice activity detection (VAD) is performed using PyAnnote [22], which is part of the WhisperX [23] framework[2], to segment the audio. For these VAD-generated segments, pseudo-labels are then extracted separately using both the Whisper and Zipformer ASR models. These pseudo-labels serve as input to the data selection pipeline, which identifies the most suitable segments for ASR finetuning. Further details are provided in Section 2.1. Finally, ASR finetuning is conducted using pretrained Whisper and Zipformer models on the entire training set, a random subset, or the selected pseudo-labeled data from our pipeline.

2.1. Data Selection Pipelines

We propose the following data selection pipelines for efficient ASR finetuning:

1. **WER Prediction-Based Selection**: Inspired by recent studies on WER estimation [24], we train an SVM classifier to categorize speech segments into low- and high-WER classes. Instead of estimating WER directly, we focus on filtering out high-WER segments, prioritizing high-quality data for model finetuning. We define the boundary between low and high WER utterances at 50%, meaning segments with WER $\leq 50\%$ are considered low-WER, while those above this threshold are categorized as high-WER. Each speech segment s is represented by concatenated HuBERT-based acoustic embeddings and XLM-R-based text embeddings, denoted as $\mathbf{f}(s) = [\mathbf{f}_{\text{HuBERT}}(s); \mathbf{f}_{\text{XLM-R}}(s)]$, and classified using an SVM:
$$\hat{y}_s = \text{SVM}(\mathbf{f}(s))$$
where $\hat{y}_s \in \{\text{low-WER}, \text{high-WER}\}$. During testing, all high-WER predicted segments are discarded, and N hours of data are randomly sampled from the remaining low-WER segments:
$$S_{\text{selected}} = \text{RandomSample}(\{s \in S_{\text{test}} \mid \hat{y}_s = \text{low-WER}\}, N) \tag{1}$$
2. **NER-Based Selection**: We employ a lightweight NER model, distilled from BERT [25], to identify segments containing named entities. Specifically, we use a version of DistilBERT [26] finetuned for NER,[3] which achieves F1-scores of 0.99 and 0.94 for identifying samples with and without entities, respectively, on the CoNLL-2003 [27]. Data selection is performed in two stages: first, we identify all the segments containing entities using the NER model; then, we sample from these segments using one of the following strategies:
   - **Random sampling**: Randomly select N hours of data from the filtered segments containing at least one named entity class:
   $$S_{\text{random}} = \text{RandomSample}(S_{\text{NER}}, N) \tag{2}$$
   where S_{NER} is the set of all segments containing named entities, and $\text{RandomSample}(S, N)$ selects N hours randomly.
   - **High Confidence NER score based sampling**: Select the top N hours of data with the highest NER confidence scores:
   $$S_{\text{conf}} = \text{Top-N}(S_{\text{NER}}, N, \text{conf}(s)) \tag{3}$$
   where $\text{conf}(s)$ is the confidence score assigned by the NER model to segment s, and $\text{Top-N}(S, N, \text{conf}(s))$ selects the N most confident segments.
   - **Entity Class Distribution-based Random Sampling**: Maintain entity class balance and randomly select data within each class:
   $$S_{\text{dist-rand}} = \bigcup_{c \in C} \text{RandomSample}(S^c_{\text{NER}}, N_c) \tag{4}$$
   where C is the set of named entity classes, S^c_{NER} is the subset of S_{NER} containing entity class c, and $N_c = P_c \cdot N$, where P_c is the proportion of class c in the full dataset.
   - **Entity Class Distribution-based High Confidence Sampling**: Maintain entity class balance while selecting the most confident segments within each class:
   $$S_{\text{dist-conf}} = \bigcup_{c \in C} \text{Top-N}(S^c_{\text{NER}}, N_c, \text{conf}(s)) \tag{5}$$
   where $\text{Top-N}(S^c_{\text{NER}}, N_c, \text{conf}(s))$ selects the top N_c hours with the highest confidence scores within class c.
3. **CER-based Selection**: The unlabeled speech corpus is transcribed using three ASR models such as Whisper Medium, Zipformer, and Nemo Parakeet. The CER is computed for each segment across all models, and the average CER is used to make the selection. Only segments with an average CER below a predefined threshold τ (e.g., 5%) are retained for finetuning:
$$S_{\text{low-CER}} = \{s \in S_{\text{ASR}} \mid \text{CER}_{\text{avg}}(s) < \tau\} \tag{6}$$
where
$$\text{CER}_{\text{avg}}(s) = \frac{1}{3}\left(\text{CER}_{\text{W-Z}}(s) + \text{CER}_{\text{W-P}}(s) + \text{CER}_{\text{Z-P}}(s)\right) \tag{7}$$
where $\text{CER}_{\text{W-Z}}(s)$ is the CER for segment s from ASR model $whisper$ and $zipformers$.

3. Experiments

3.1. Datasets

We evaluate our proposed approach using two datasets: the Wow dataset, which consists of real-world call center conversations, and the Fisher English dataset, a widely used benchmark for conversational speech. Below, we provide detailed descriptions of both datasets.

- **Wow** : This dataset comprises contact center conversations recorded at a sampling rate of 44.1 kHz in stereo format. The conversations span six domains: automotive, auto insurance,

[2] https://github.com/SYSTRAN/faster-whisper
[3] https://huggingface.co/dslim/bert-base-NER-uncased

Table 1: *Overview of Datasets. The approximated total speech duration is derived from Lhotse cuts[28].*

Dataset	Split	Transcripts	# Conversations	# Segments	Total Duration
Wow	train	✗	70k	2.58M	7500h
	dev	✓	60	5k	4.5h
	test	✓	250	26k	18h
Fisher	train	✗	11.5k	1.88M	1878h
	dev	✓	25	5k	3h
	test	✓	23	5k	3h

medicare, medical, home services, and customer services.[4] Table 1 summarizes the dataset statistics. Notably, the training set does not include ground truth transcripts, and pseudo-labels generated by the Whisper and Zipformer models are used for experimentation, reflecting the real-world scenario of limited annotations.

- **Fisher English**: The Fisher dataset [29] is a large collection of conversational telephone speech, comprising approximately 11.5k conversations across diverse topics and speakers. The dataset is divided into training, development, and test splits, with a total duration of approximately 2,000 hours. To simulate the no-annotation scenario relevant to our study, we discard the original ground truth transcripts for the training set and use pseudo-labels generated as with Wow for experimentation. The development and test sets retain their ground truth transcripts for evaluation purposes.

Since the ground truth WER values doesn't exists for Wow , we used AMI [30] training set for training the WER classifier.

3.2. ASR Models for Fine-tuning

For adapting the ASR system to domain-specific data, we used the following models:

- **Whisper Medium**: A 769M parameter model finetuned on a domain-specific dataset. The model is updated using gradient descent with a learning rate η. Evaluation is done using task-specific metrics.
- **Zipformer**: A pretrained transformer-based model with approximately 70M parameters trained on Gigaspeech dataset is used to finetune the train set. It uses the ScaledAdam optimizer, a learning rate scheduler with warmup and decay, and a combined RNN-T and CTC loss. Training is done for 30 epochs with a peak learning rate of $5.0e^{-2}$ on a single RTX 3090 GPU.

3.3. Histogram of data selected using different Data Selection Pipelines

Figure 2 illustrates the distribution of NER-identified entity durations across four selection strategies applied to 2500 hours of the Wow training dataset.

We observed the average CER distribution across segments in the train set. Around 15-20% of the total hours fall within the 0-5 CER bin, indicating a high concentration of low-error data.

4. Results and Discussion

4.1. Evaluation of the WER Estimator

The WER classification model is evaluated on both the Wow and AMI [30] test sets, with the classification results summarized in Table 2. We observe that the estimator performs well in

[4]For more details, see: `https://wow-ai.com/ots-dataset.html`

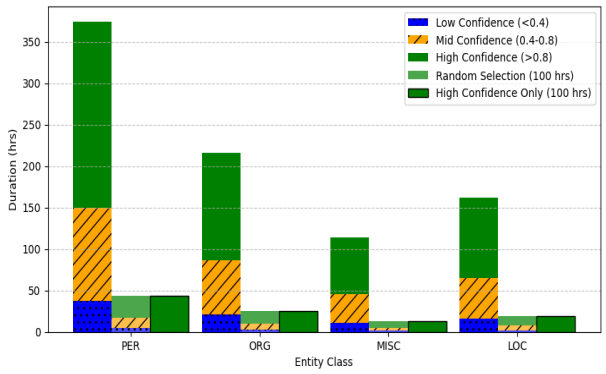

Figure 2: *NER Entity Class Distribution with Confidence Levels. Each bar represents a different filtering method: Total Available Data (leftmost) shows the overall entity class distribution in grayscale, segmented by confidence levels (low, mid, high). Random Selection (100 hrs) maintains the same confidence distribution but selects segments randomly. High Confidence (100 hrs) prioritizes segments with the highest NER confidence scores (>0.8). Entity Class Balanced Selection (rightmost two bars) ensures proportional representation of entity classes while choosing segments either randomly or with high confidence.*

Table 2: *WER Classification Results on Wow and AMI Test Sets*

Dataset	WER Class	Precision	Recall	F1-Score	Support	Accuracy
AMI	Low WER	0.90	0.95	0.92	5977	0.89
	High WER	0.84	0.73	0.78	2340	
Wow	Low WER	0.85	0.95	0.89	18700	0.83
	High WER	0.74	0.48	0.58	5973	

distinguishing between low- and high-WER classes, achieving an overall accuracy of 83% on the Wow test set and 89% on the AMI test set. The model demonstrates higher precision and recall for the low-WER class across both datasets, indicating its robustness in identifying well-recognized speech segments. However, the high-WER class exhibits lower recall, particularly in the Wow test set, suggesting that high-WER utterances are often mistaken for low-WER ones. We use competitive systems for NE detection and CER estimation as utilities to guide data selection, rather than focusing on their absolute performance metrics. Since the WOW and Fisher datasets do not contain ground-truth named entity annotations, we prioritize the utility and integration of these systems within our data selection pipeline over their standalone evaluation accuracy.

4.2. Baseline Performance Without Data Selection

To establish a clear reference point, we first evaluate the performance of ASR models without any data selection strategies. This includes assessing both the performance of pretrained models on unseen data and the results of fine-tuning on the entire dataset, providing an upper bound for model performance.

- *No Fine-Tuning (Pre-Trained Performance)*: This is a scenario where no data is seleced as a part of finetuning. The pretrained Whisper and Zipformer models show WERs of 14.8% and 15.6% on WOW and 17.9% and 16.2% on Fisher, respectively. These results indicate that impact of pretrained

Table 3: *WER (%) comparison across different data selection pipelines for finetuning Whisper and Zipformer models. The table presents results for various selection criteria, including random selection, WER classification, and NER-based strategies, with performance evaluated on the two datasets: WOW and Fisher. For all data selection experiments, 100 hours of data were used. The baseline fine-tuning was performed on the full datasets: 7500 hours for WOW and approximately 1800 hours for Fisher.*

Stage	**Pipeline**	**Selection Criteria**	**Whisper**		**Zipformer**	
			WOW	**Fisher**	**WOW**	**Fisher**
Pre-Train	N/A	N/A	14.8	17.9	15.6	16.2
Fine-tune	Baseline	All Data	**12.3**	**14.8**	**14.3**	**15.4**
	Random	Random	13.3	16.0	14.8	15.9
	WER Clf.	Low WER Segments	13.1	15.1	14.6	15.5
	NER	Random	12.9	15.7	14.6	15.6
		High Conf. NER Scores	13.2	15.4	14.6	15.8
		Entity Dist. + Random	12.8	15.2	14.7	15.8
		Entity Dist. + High Conf.	**12.5**	**15.0**	**14.5**	**15.5**
	CER	Segments Avg CER<5%	**12.2**	**14.1**	**13.3**	**14.6**

models on unseen data, highlighting the need for adaptation.

- *Baseline Fine-Tuning*: Here the ASR adaptation is done on all pseudo-labels of train sets (7500h for WOW and 2000h for Fisher): Fine-tuning on the full dataset significantly improves performance, bringing WERs down to 12.3% (WOW) and 14.8% (Fisher) for Whisper and 14.3% (WOW) and 15.4% (Fisher) for Zipformer. This serves as the upper bound, showing the best possible improvement with extensive labeled and pseudo-labeled data.

4.3. Impact of Data Selection on ASR Finetuning

Fine-tuning was performed independently for Whisper and Zipformer on both Fisher and WOW datasets. A fixed 100-hour subset was selected at the start using different data selection pipelines, ensuring consistency across experiments.

- *Random data selection*: With a fixed seed (42), random selection performs worse than the baseline, with WERs of 13.3% (WOW) and 16.0% (Fisher) for Whisper and 14.8% (WOW) and 15.9% (Fisher) for Zipformer. This suggests that randomly selecting 100 hours is suboptimal, as it doesn't prioritize informative or representative samples.
- *WER-based selection*: Selecting segments with low WER performs better than random, achieving 13.1% (WOW) and 15.1% (Fisher) for Whisper and 14.6% (WOW) and 15.5% (Fisher) for Zipformer. This approach helps in selecting cleaner, more reliable training samples, leading to modest improvements over random selection.
- *NER-based selection*: NER-based selection outperforms WER-based selection in most cases. However, choosing only high-confidence NER segments does not guarantee better performance, sometimes performing worse than WER selection. The best NER-based method is Entity Distribution + High Confidence, which reaches 12.5% (WOW) and 15.0% (Fisher) for Whisper and 14.5% (WOW) and 15.5% (Fisher) for Zipformer—close to the baseline. This shows that a balanced selection approach is crucial rather than relying purely on high-confidence entities.
- *CER-based selection*: Selecting segments with CER $< 5\%$ achieves the lowest WERs among all data selection strategies: 12.2% (WOW) and 14.1% (Fisher) for Whisper, and 13.3% (WOW) and 14.6% (Fisher) for Zipformer. This even surpasses the full-dataset baseline, showing that a carefully chosen subset of clean segments can be more effective than finetuning on large training sets.
- *Impact of NER/CER*: We find that the improvements come from selecting relevant data based on NER class distribution and CER confidence scores, as compared to random selection. As shown in Table 3, pure random and NER-filtered random subsets perform worse than CER and histogram-based NER methods; therefore, what kind of data is selected matters more than the quantity selected.

5. Conclusions

We have explored different data selection pipelines and proposed a robust approach that improves ASR adaptation by filtering pseudo-labels generated using Whisper and Zipformer models. Our approach integrates multiple selection strategies—including WER prediction, NER, and character-level agreement analysis—to extract high-quality pseudo-labels. Our experiments show that finetuning on just 1% to 5% of the pseudo-labeled dataset can achieve performance comparable to that of full dataset finetuning, thus offering a computationally efficient solution for small organizations with limited resources.

6. Acknowledgments

This work was supported by Idiap Research Institute and Uniphore collaboration project. Part of this work was also supported by EU Horizon 2020 project ELOQUENCE (grant number 101070558).

7. References

[1] N. Morgan, H. Bourlard, S. Renals, M. Cohen, and H. Franco, "Hybrid neural network/hidden Markov model systems for continuous speech recognition," in *Advances in Pattern Recognition Systems Using Neural Network Technologies*. World Scientific, 1993, pp. 255–272.

[2] H. A. Bourlard and N. Morgan, *Connectionist speech recognition: a hybrid approach*. Springer Science & Business Media, 1993, vol. 247.

[3] D. Povey, M. Hannemann, G. Boulianne, L. Burget, A. Ghoshal,

M. Janda, M. Karafiát, S. Kombrink, P. Motlíček, Y. Qian *et al.*, "Generating exact lattices in the WFST framework," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2012, pp. 4213–4216.

[4] D. Imseng, P. Motlicek, P. N. Garner, and H. Bourlard, "Impact of deep MLP architecture on different acoustic modeling techniques for under-resourced speech recognition," in *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*. IEEE, 2013, pp. 332–337.

[5] 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.

[6] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented Transformer for Speech Recognition," in *Proc. Interspeech 2020*, 2020, pp. 5036–5040.

[7] V. Pratap, A. Tjandra, B. Shi, P. T. A. B. S. Kundu, A. Elkahky, Z. N. A. V. M. Fazel, Z. A. Baevski, and M. Auli, "Scaling speech technology to 1,000+ languages," *arXiv preprint arXiv:2305.13516*, 2023.

[8] A. Bapna, C. Cherry, Y. Zhang, Y. Jia, M. Johnson, Y. Cheng, S. Khanuja, J. Riesa, and A. Conneau, "MSLAM: Massively multilingual joint pre-training for speech and text," *arXiv preprint arXiv:2202.01374*, 2022.

[9] A. Srinivasamurthy, P. Motlicek, I. Himawan, G. Szaszak, Y. Oualil, and H. Helmke, "Semi-supervised learning with semantic knowledge extraction for improved speech recognition in air traffic control," in *Proceedings of Interspeech 2017*, 2017.

[10] T. Viglino, P. Motlicek, and M. Cernak, "End-to-end accented speech recognition." in *Interspeech*, 2019, pp. 2140–2144.

[11] M. Bhattacharjee, P. Motlicek, S. Madikeri, H. Helmke, O. Ohneiser, M. Kleinert, and H. Ehr, "Minimum effort adaptation of automatic speech recognition system in air traffic management," *European Journal of Transport and Infrastructure Research*, vol. 24, no. 4, pp. 133–153, 2024.

[12] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, "Unsupervised Cross-Lingual Representation Learning for Speech Recognition," in *Proc. Interspeech 2021*, 2021, pp. 2426–2430.

[13] M. Zeineldeen, J. Xu, C. Lüscher, W. Michel, A. Gerstenberger, R. Schlüter, and H. Ney, "Conformer-based hybrid ASR system for switchboard dataset," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7437–7441.

[14] Z. Yao, L. Guo, X. Yang, W. Kang, F. Kuang, Y. Yang, Z. Jin, L. Lin, and D. Povey, "Zipformer: A faster and better encoder for automatic speech recognition," in *The Twelfth International Conference on Learning Representations*, 2024. [Online]. Available: https://openreview.net/forum?id=9WD9KwssyT

[15] L. Lugosch, T. Likhomanenko, G. Synnaeve, and R. Collobert, "Pseudo-labeling for massively multilingual speech recognition," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7687–7691.

[16] G. Vanderreydt, A. Prasad, D. Khalil, S. Madikeri, K. Demuynck, and P. Motlicek, "Parameter-efficient tuning with adaptive bottlenecks for automatic speech recognition," in *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2023, pp. 1–7.

[17] P. Rangappa, J. Zuluaga-Gomez, S. Madikeri, A. Carofilis, J. Prakash, S. Burdisso, S. Kumar, E. Villatoro-Tello, I. Nigmatulina, P. Motlicek *et al.*, "Speech data selection for efficient ASR fine-tuning using domain classifier and pseudo-label filtering," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.

[18] C. S. Anoop, A. Prathosh, and A. Ramakrishnan, "Unsupervised domain adaptation schemes for building ASR in low-resource languages," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2021, pp. 342–349.

[19] S. Wotherspoon, W. Hartmann, M. Snover, and O. Kimball, "Improved data selection for domain adaptation in ASR," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 7018–7022.

[20] W.-N. Hsu, Y.-H. H. Tsai, B. Bolte, R. Salakhutdinov, and A. Mohamed, "HuBERT: How much can a bad teacher benefit ASR pre-training?" in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6533–6537.

[21] Y. K. Wiciaputra, J. C. Young, and A. Rusli, "Bilingual text classification in english and indonesian via transfer learning using XLM-RoBERTa." *International Journal of Advances in Soft Computing & Its Applications*, vol. 13, no. 3, 2021.

[22] H. Bredin, "PyAnnote audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe," in *24th INTERSPEECH Conference (INTERSPEECH 2023)*. ISCA, 2023, pp. 1983–1987.

[23] M. Bain, J. Huh, T. Han, and A. Zisserman, "WhisperX: Time-accurate speech transcription of long-form audio," *arXiv preprint arXiv:2303.00747*, 2023.

[24] C. Park, M. Chen, and T. Hain, "Automatic speech recognition system-independent word error rate estimation," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, and N. Xue, Eds. Torino, Italia: ELRA and ICCL, May 2024, pp. 1979–1987. [Online]. Available: https://aclanthology.org/2024.lrec-main.178/

[25] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 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)*, J. Burstein, C. Doran, and T. Solorio, Eds. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423/

[26] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," 2020. [Online]. Available: https://arxiv.org/abs/1910.01108

[27] E. F. Tjong Kim Sang and F. De Meulder, "Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition," in *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, 2003, pp. 142–147. [Online]. Available: https://aclanthology.org/W03-0419/

[28] P. Żelasko, D. Povey, J. Trmal, S. Khudanpur *et al.*, "Lhotse: a speech data representation library for the modern deep learning ecosystem," *arXiv preprint arXiv:2110.12561*, 2021.

[29] C. Cieri, D. Miller, and K. Walker, "The Fisher corpus: A resource for the next generations of speech-to-text." in *LREC*, vol. 4, 2004, pp. 69–71.

[30] W. Kraaij, T. Hain, M. Lincoln, and W. Post, "The AMI meeting corpus," in *Proc. International Conference on Methods and Techniques in Behavioral Research*, 2005, pp. 1–4.