Speech Data Selection for Efficient ASR Fine-Tuning using Domain Classifier and Pseudo-Label Filtering

Pradeep Rangappa[1], Juan Zuluaga-Gomez[1,2], Srikanth Madikeri[1], Andres Carofilis[1], Jeena Prakash[4], Sergio Burdisso[1], Shashi Kumar[1,2], Esaú Villatoro-Tello[1], Iuliia Nigmatulina[1,3], Petr Motlicek[1,5], Karthik Pandia[4] and Aravind Ganapathiraju[4]

[1] Idiap Research Institute, Martigny, Switzerland
[2] EPFL, Lausanne, Switzerland
[3] University of Zurich, Switzerland
[4] Uniphore Software Systems, India
[5] Brno University of Technology, Czech Republic

***Abstract*—In real-world speech data processing, the scarcity of annotated data and the abundance of unlabelled speech data present a significant challenge. To address this, we propose an efficient data selection pipeline for fine-tuning ASR models by generating pseudo-labels using WhisperX pipeline and selecting efficient labels for fine-tuning. In our work, we propose a domain classifier system developed with a computationally inexpensive TFIDF and classical machine learning algorithm. Later, we filter data from the classifier output using a novel metric that assesses word ratio and perplexity distribution. The filtered pseudo labels are then used for fine-tuning standard encoder-decoder Whisper models and Zipformer. Our proposed data selection pipeline reduces the dataset size by approximately 1/100th while maintaining performance comparable to the full dataset, outperforming random domain-independent selection strategies.**

***Index Terms*—Automatic Speech Recognition (ASR), Data Selection, Domain Classification, WhisperX, Zipformer**

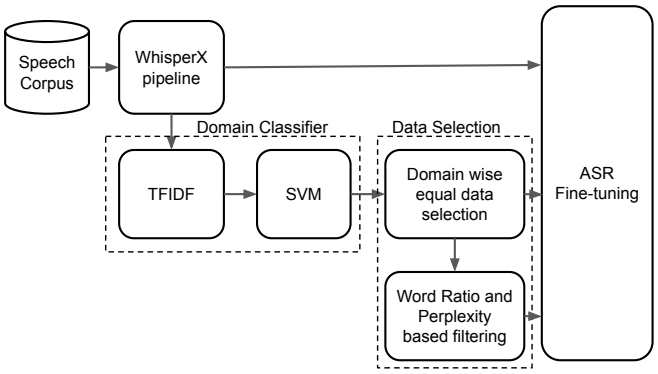


Fig. 1: Functional blocks of the proposed data selection pipeline.

I. Introduction

Automatic Speech Recognition (ASR) systems have undergone significant advancements in recent years [1], moving from traditional hybrid-based models [2]–[5] to cutting-edge end-to-end architectures [6]–[13]. These innovations include models such as wav2vec 2.0, cross lingual speech representations (XLSR) [14], Conformer [7], [15], and the latest Zipformer [16], all of which leverage large-scale speech data to improve transcription accuracy. Most of these advancements rely on supervised learning methods, where massive annotated datasets are used to fine-tune models [17]. While this approach has been highly effective, the availability of annotated data remains a major bottleneck in scaling ASR solutions to broader, real-world applications.

Past works suggest improving ASR performance in target domain by exploiting manually transcribed data from other domains, e.g., through feature mapping approaches [18]. More recent works explore semi-supervised learning as a cost-effective solution to this problem [19]–[23], utilizing pseudo labeled speech data available for target domain to fine-tune ASR models. This approach allows models to benefit from the vast quantities of unlabeled data flowing into industrial pipelines. At Uniphore, an AI-native company that unifies voice, video, text, and data with AI to enhance customer and employee experiences, these challenges are particularly pronounced. Large volumes of customer-agent interaction data are collected daily, but only a fraction of it is manually labeled due to high annotation costs. This makes it crucial to explore efficient data selection methods, especially when applying pseudo labels.

Finetuning ASR models faces challenges in selecting relevant data. Traditional methods rely on static datasets [24] and fixed domain labels [25], which fail to adapt to evolving real-world data like customer-agent interactions. The absence of domain labels and the high cost of training on large datasets [24] further complicate the process. We propose a dynamic data selection pipeline that uses pseudo labels for domain classification and filters high-quality data for efficient ASR fine-tuning.

Our proposed approach (see Figure 1) comprises four key stages: (1) pseudo label generation using WhisperX pipeline, (2) training a domain classifier to classify into 6 domains, (3) our data selection pipeline which includes (a) selection of an equal amount of data from different domains to ensure balanced fine-tuning across diverse contexts, (b) introduce a novel metric to automatically filter hallucinated or low-quality

data, further enhancing model performance and (4) adapting whisper and zipformer models on the selected pseudo labels. This pipeline is designed to optimize both the efficiency and accuracy of fine-tuning ASR models in real-world applications, providing a scalable solution for evolving speech data.

In this paper, we enhance data selection for ASR fine-tuning by overcoming the constraints of using fixed seed-labeled distributions. First, we leverage pseudo labels to effectively fine-tune ASR models on large, unlabeled datasets, overcoming the challenge of lacking ground truth labels. Second, we introduce a novel domain classifier trained directly from pseudo labels, eliminating the need for prior domain knowledge. Third, we present a dynamic data selection pipeline that adapts to evolving acoustic and lexical properties by utilizing text-based features such as word ratio and perplexity. Lastly, our approach demonstrates a significant improvement in ASR fine-tuning performance through effective data selection based on pseudo labels.

The paper is organized as follows: Section II briefly describes the proposed data selection pipeline. Section III and IV discusses experimental design and results. Section V concludes the paper, suggesting future directions.

II. Proposed Speech Data Selection Pipeline

Our proposed approach is in Figure 1 in the introduction. The functional blocks of Figure 1 are illustrated below:

1) *Speech Corpus*: the speech corpus at Uniphore includes a large amount of untranscribed conversations between agents and customers, with only a small portion transcribed by Uniphore's annotation partners.
2) *WhisperX pipeline*: In this task, we aim at transcribing large unlabeled audio-only corpora with state-of-the-art foundational speech models such as WhisperX [26]. WhisperX consist of i) an audio pre-processing step, including voice activity detection (VAD) and optional speaker diarization (SD) using PyAnnote [27]; ii) segment batching with cut & merge for efficient inference; and finally, iii) a transcription phase with one of the pre-trained Whisper models [28]. All in all, WhisperX is a friendly pipeline to generate low-effort large-scale PL datasets. In addition, it loads the models with faster-whisper [1], which uses quantized pre-trained models from CTranslate2 [2], e.g., whisper-large-v2.
3) *Domain Classification*: The Uniphore dataset consists of six domains: auto insurance, automotive, customer service, home service, medical, and medicare. We propose domain classification on this dataset using Term Frequency-Inverse Document Frequency (TFIDF) features, employing machine learning algorithms such as logistic regression and SVM [29]. The domain classifier is integrated as the front end of the data selection pipeline to enhance its efficiency.
4) *Data Filtering*: There are two steps followed in the data filtering block:
   - Filtering data based on domain: Here we select equal amount of speech data based on the outputs of the domain classifier.
   - Filtering using word ratio (WR) and perplexity (PPL): In all our experiments, the ASR output generated by WhisperX is processed through two filtering mechanisms: (1) WR (2) PPL. WR is defined as the total number of words that are present in the speech segment divided by the total duration of that particular segment. It is calculated across all segments of the ASR outputs using the following formula:

$$WR = \frac{\sum_{i=1}^{N} W_i}{\sum_{i=1}^{N} D_i} \tag{1}$$

     where:
     - W_i is the number of words in segment i,
     - D_i is the duration of segment i,
     - N is the total number of segments.

     Filtering based on WR helps detect hallucinations, a known issue in Whisper models where excessive words are generated for short speech segments [30]. WR measures the ratio of words to segment duration, with higher values indicating potential hallucinations. In this paper, we analyze the WR distribution and set an empirical threshold to filter out these segments, improving pseudo-label quality.

     An another metric that we utilize for data selection is via PPL and is calculated as:

$$\text{Perplexity}(P) = 2^{-\frac{1}{N}\sum_{i=1}^{N} \log_2 P(w_i|w_1,w_2,\ldots,w_{i-1})} \tag{2}$$

     where:
     - $P(w_i \mid w_1, w_2, \ldots, w_{i-1})$ is the probability assigned by the GPT-2 [31] language model to the word given the preceding words.
     - N is the total number of words in the sequence.

     GPT-2, a transformer-based model, uses self-attention to capture dependencies between the current word and previous words, modeling complex relationships and context across multiple layers for more accurate predictions. In our approach, we calculate PPL on WhisperX segments and analyze its distribution to select a representative subset of data. This method helps prioritize more informative and challenging segments, improving model fine-tuning by focusing on data that enhances performance and reduces redundancy.
5) *ASR Finetuning*: The pre-trained Whisper medium model [32], with 769M parameters, is fine-tuned using (a) the entire training set, (b) a randomly selected subset, and (c) pseudo-labels from our data selection pipeline. The pseudo-labels are further refined using Zipformers [33]. More details on model fine-tuning are provided in Section III.

[1] https://github.com/SYSTRAN/faster-whisper
[2] https://github.com/OpenNMT/CTranslate2/

III. Experiments

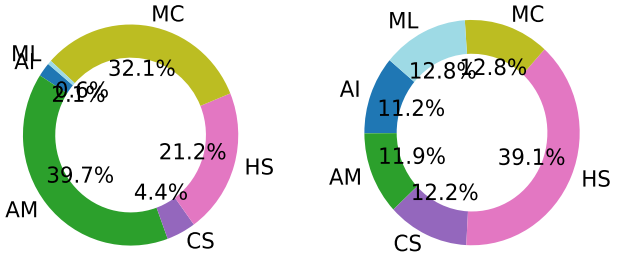
The following section outlines the datasets and provides details of the experimental setup used for our various ablations.

A. Data

The dataset contains industry conversations recorded at 44.1 kHz stereo across six domains: automotive, auto insurance, medicare, medical, home services, and customer services. Table I shows the duration of data in training, development, and test sets. Total speech duration is derived from Lhotse cuts [34]. The training set does not contain any ground truth transcripts, whereas the test and development sets have them available. Figure 2 (a) and (b) shows the distribution of domains across train and test data set.

TABLE I: Overview of Conversational Speech Dataset

Dataset	# Conversations	# Segments	Total speech duration (hrs)
Train	67911	1.35 M	7057
Dev	64	6516	5
Test	248	26649	20



(a) Train (b) Test

Fig. 2: Distribution of domains across (a) Train and (b) Test set. The domain abbreviations are: HS (Home Services), AI (Auto Insurance), AM (Automotive), ML (Medical), MC (Medicare), and CS (Customer Service).

In our domain classification experiment, we used only 1% of the training set (approximately 72 hours), ensuring that the dataset was balanced with 12 hours of data from each domain, totaling 72 hours.

B. Models used for different stages of data selection pipeline

1) Domain Classification Models: We utilize two machine learning models for domain classification: Logistic Regression and Support Vector Machines (SVM). Logistic Regression provides a straightforward approach to classify domains based on conversation features, offering simplicity and interpretability. In contrast, SVMs are effective in high-dimensional feature spaces and enhance classification accuracy by maximizing the margin between domain classes.

2) ASR Model Adaptation: For adapting the Automatic Speech Recognition (ASR) system to domain-specific data, we use:

- **Whisper Medium** – A model with 769 million parameters, designed for precise speech recognition. Adaptation involves fine-tuning a pretrained model on a domain-specific dataset D_{adapt}. The adaptation process uses the loss function $\mathcal{L}_{adapt} = -\sum_{(x,y)\in D_{adapt}} \log P(y|x,\theta_{adapt})$, where (x,y) are input-target pairs and θ_{adapt} are the updated model parameters. Fine-tuning is performed by minimizing $\mathcal{L}_{adapt}$ using gradient descent: $\theta_{adapt} \leftarrow \theta_{adapt} - \eta\nabla_{\theta_{adapt}}\mathcal{L}_{adapt}$, with η as the learning rate. The adapted model is then evaluated using task-specific metrics.
- **Zipformer** – A transformer-based model for managing long-range dependencies and varying speech characteristics, optimized for domain-specific fine-tuning. We train Transformer-Transducer models from scratch using the Zipformer encoder [16] with the Icefall Transducer recipe. Training employs the *ScaledAdam* optimizer [35], a learning rate scheduler with 500-step warmup and decay phases [36], and a combined RNN-T and CTC loss [6], [37], [38]. The loss function is defined as $\mathcal{L} = (1-\lambda)\cdot\mathcal{L}_{RNN-T} + \lambda\cdot\mathcal{L}_{CTC}$ with $\lambda = 0.1$. Training is performed for 30 epochs with a peak learning rate of $5.0e^{-2}$ on a single RTX 3090 GPU.

C. Evaluation

The domain classifier's performance is measured using F1 scores per domain and overall accuracy. For the fine-tuned ASR model, thee standard word error rate (WER) is used for evaluation. WER $= \frac{S+D+I}{N}$ where S, D, and I represent substitutions, deletions, and insertions, respectively, and N is the total number of words in the reference.

IV. Results and Discussion

A. Domain Classification Results

Table II presents the performance metrics of two classifiers, Logistic Regression and SVM, across various domains. It can be observed that the top four domains such as auto insurance, automotive, customer service, and home services achieve an average F1 score above 95%, reflecting strong performance. In contrast, the medical and medicare domains have comparatively lower F1 scores due to the use of overlapping medical terminology, which poses challenges for classification.

TABLE II: Performance of Classifiers by Domain on the test data

Domain	Classifier	P	R	F1	Support
auto_insurance	Logistic Regression	0.92	0.94	0.93	35
	SVM	0.92	0.94	0.93	
automotive	Logistic Regression	0.90	0.97	0.94	37
	SVM	0.95	0.97	0.96	
customer_service	Logistic Regression	1.00	0.92	0.96	38
	SVM	1.00	0.92	0.96	
homeservices	Logistic Regression	0.98	0.98	0.98	122
	SVM	0.98	0.98	0.98	
medical	Logistic Regression	0.90	0.45	0.60	40
	SVM	0.95	0.45	0.61	
medicare	Logistic Regression	0.63	0.93	0.75	40
	SVM	0.62	0.95	0.75	

TABLE III: WER Performance Across Different Data Selection Strategies for ASR Model Fine-Tuning

Model	Data Selection Stages for Training	Data Selection for Training	Duration of Train set (hh:mm:ss)	WER (↓)
Pretrained (whisper medium)	N/A	N/A	N/A	17.53
Finetuned (whisper medium)	Baseline	All	7056:32:49	15.26
	Fixed	Single Domain (home services)	95:59:58	19.37
	Random	Random (seed=42)	95:59:59	18.51
		Random (seed=111)	95:59:47	17.14
		Random (seed=2024)	95:59:45	16.30
	Domain Dependent (Filtering 01)	Equal data selection (seed=42)	95:58:52	16.57
		Equal data selection (seed=111)	95:58:34	16.44
		Equal data selection (seed=2024)	95:58:54	17.28
	Domain Dependent (Filtering 02)	Perplexity	95:22:60	17.08
		Word Ratio	94:49:60	16.26
Pretrained (zipformers)	N/A	N/A	N/A	23.52
Finetuned (zipformers)	Filtering 01 + Filtering 02	Equal data selection + Word ratio	94:49:60	15.60

B. ASR Finetuning based on Data Selection Pipeline

Our comparison of data selection methods is currently limited to random selection, as presented in Table III. Unlike recent approaches [25], [39], which rely on seed models trained on manually labeled data (unavailable for real-world customer-agent interactions), our strategy leverages perplexity and word ratio distributions for selection. This study highlights (1) comparable fine-tuning results between 7000 hours and 95 hours of data and (2) a framework tailored to industrial needs.

1) Results using Whisper Adaptation: Table III summarizes the WER performance across various data selection strategies for fine-tuning the Whisper medium ASR model. The pretrained model had a WER of 17.53%, which improved to 15.26% after fine-tuning on the full training set. However, fine-tuning on a single domain (home services) increased WER to 19.37%, indicating domain bias when evaluated across all domains. Random selection of Lhotse segments (96 hours) resulted in WERs ranging from 16.30% to 18.51%, with an average of 17.32%. Domain-balanced random selection gave WERs between 16.57% and 17.28%, averaging 16.76%. Further filtering based on perplexity and word ratio reduced WER to 17.08% and 16.26%, respectively, demonstrating the effectiveness of targeted data selection.

2) Results using Zipformers Adaptation: Table III (bottom part) compares the WER performance of whisper and zipformer models, both pretrained and fine-tuned using the proposed Data Selection Pipeline (DSP). For the Whisper model, the pretrained version achieves a WER of 17.53%, while fine-tuning using data selection pipeline (DSP) over 94 hours of training reduces the WER to 16.26%. In contrast, the zipformer pretrained model exhibits a higher initial WER of 23.52%. When trained from scratch for 94 hours, its WER increases significantly to 34.23%, demonstrating the challenge of training from scratch on a small amount of data. However, applying the DSP for fine-tuning zipformer over 94 hours improves the WER considerably, lowering it to 15.60%, outperforming Whisper and highlighting the effectiveness of the proposed DSP in fine-tuning both models.

C. Effectiveness and Practical Implications

The proposed data selection pipeline has been deployed in a real-world production environment at Uniphore, addressing key challenges in adapting ASR models to evolving datasets. High-quality data is selected on a weekly basis, ensuring continuous improvements in transcription accuracy. The computationally intensive domain classifier ensures data is classified across six domains, enabling effective fine-tuning, while the fast and efficient WhisperX pipeline generates pseudo labels, maintaining scalability. As a next step, we plan to streamline the process by passing only critical data to WhisperX, optimizing both data selection and ASR adaptation.

V. Conclusion

In this paper, we address the challenge of limited annotated data and an excess of unlabeled speech data by introducing an efficient data selection pipeline for ASR model fine-tuning. Our approach integrates four key stages: generating pseudo-labels with the WhisperX pipeline, training a domain classifier with 90% accuracy using a TFIDF-based classical machine learning algorithm, applying a novel metric for data filtering, and adapting Whisper and Zipformer models. Our pipeline effectively reduces the dataset from 7,000 hours to just 94 hours by ensuring balanced domain coverage and filtering low-quality data. The results demonstrate that this streamlined approach maintains performance comparable to using the full dataset, significantly surpassing traditional random selection methods.

Acknowledgment

This work was supported by Idiap Research Institute and Uniphore collaboration project. Part of this work was also supported by EU Horizon 2020 project ELOQUENCE5 (grant number 101070558).

References

[1] Amandeep Singh Dhanjal and Williamjeet Singh, "A comprehensive survey on automatic speech recognition using neural networks," *Multimedia Tools and Applications*, vol. 83, no. 8, pp. 23367–23412, 2024.

[2] Nelson Morgan, Herve Bourlard, Steve Renals, Michael Cohen, and Horacio Franco, "Hybrid neural network/hidden markov model systems for continuous speech recognition," in *Advances in Pattern Recognition Systems Using Neural Network Technologies*, pp. 255–272. World Scientific, 1993.

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

[4] Daniel Povey, Mirko Hannemann, Gilles Boulianne, Lukáš Burget, Arnab Ghoshal, Miloš Janda, Martin Karafiát, Stefan Kombrink, Petr Motlíček, Yanmin 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.

[5] David Imseng, Petr Motlicek, Philip N Garner, and Hervé 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.

[6] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen 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.

[7] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang, "Conformer: Convolution-augmented Transformer for Speech Recognition," in *Proc. Interspeech 2020*, 2020, pp. 5036–5040.

[8] Florian Mai, Juan Zuluaga-Gomez, Titouan Parcollet, and Petr Motlicek, "HyperConformer: Multi-head HyperMixer for Efficient Speech Recognition," in *Proc. Interspeech*, 2023, pp. 2213–2217.

[9] Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello Arun Babu Sayani Kundu, Ali Elkahky, Zhaoheng Ni Apoorv Vyas Maryam Fazel, Zarandi Alexei Baevski, and Michael Auli, "Scaling speech technology to 1,000+ languages," *arXiv preprint arXiv:2305.13516*, 2023.

[10] Ankur Bapna, Colin Cherry, Yu Zhang, Ye Jia, Melvin Johnson, Yong Cheng, Simran Khanuja, Jason Riesa, and Alexis Conneau, "MSLAM: Massively multilingual joint pre-training for speech and text," *arXiv preprint arXiv:2202.01374*, 2022.

[11] Ajay Srinivasamurthy, Petr Motlicek, Ivan Himawan, Gyorgy Szaszak, Youssef Oualil, and Hartmut Helmke, "Semi-supervised learning with semantic knowledge extraction for improved speech recognition in air traffic control," in *Proceedings of Interspeech 2017*, 2017.

[12] Thibault Viglino, Petr Motlicek, and Milos Cernak, "End-to-end accented speech recognition.," in *Interspeech*, 2019, pp. 2140–2144.

[13] Mrinmoy Bhattacharjee, Petr Motlicek, Srikanth Madikeri, Hartmut Helmke, Oliver Ohneiser, Matthias Kleinert, and Heiko 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.

[14] Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli, "Unsupervised Cross-Lingual Representation Learning for Speech Recognition," in *Proc. Interspeech 2021*, 2021, pp. 2426–2430.

[15] Mohammad Zeineldeen, Jingjing Xu, Christoph Lüscher, Wilfried Michel, Alexander Gerstenberger, Ralf Schlüter, and Hermann 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.

[16] Zengwei Yao, Liyong Guo, Xiaoyu Yang, Wei Kang, Fangjun Kuang, Yifan Yang, Zengrui Jin, Long Lin, and Daniel Povey, "Zipformer: A faster and better encoder for automatic speech recognition," in *The Twelfth International Conference on Learning Representations*, 2024.

[17] Juan Zuluaga-Gomez, Amrutha Prasad, Iuliia Nigmatulina, Seyyed Saeed Sarfjoo, Petr Motlicek, Matthias Kleinert, Hartmut Helmke, Oliver Ohneiser, and Qingran Zhan, "How does pre-trained wav2vec 2.0 perform on domain-shifted asr? an extensive benchmark on air traffic control communications," in *2022 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2023, pp. 205–212.

[18] Ivan Himawan, Petr Motlicek, David Imseng, Blaise Potard, Namhoon Kim, and Jaewon Lee, "Learning feature mapping using deep neural network bottleneck features for distant large vocabulary speech recognition," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Apr. 2015, pp. 4540–4544.

[19] David Imseng, Blaise Potard, Petr Motlicek, Alexandre Nanchen, and Hervé Bourlard, "Exploiting un-transcribed foreign data for speech recognition in well-resourced languages," in *Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing*. 2014, pp. 2322 – 2326, IEEE.

[20] Banriskhem Khonglah, Srikanth Madikeri, Subhadeep Dey, Hervé Bourlard, Petr Motlicek, and Jayadev Billa, "Incremental semi-supervised learning for multi-genre speech recognition," in *Proceedings of ICASSP 2020*, 2020.

[21] Loren Lugosch, Tatiana Likhomanenko, Gabriel Synnaeve, and Ronan 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.

[22] Dan Berrebbi, Ronan Collobert, Samy Bengio, Navdeep Jaitly, and Tatiana Likhomanenko, "Continuous pseudo-labeling from the start," in *The Eleventh International Conference on Learning Representations*, 2022.

[23] Han Zhu, Dongji Gao, Gaofeng Cheng, Daniel Povey, Pengyuan Zhang, and Yonghong Yan, "Alternative pseudo-labeling for semi-supervised automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

[24] Shannon Wotherspoon, William Hartmann, Matthew Snover, and Owen 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.

[25] Nikolaos Lagos and Ioan Calapodescu, "Unsupervised multi-domain data selection for asr fine-tuning," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 10711–10715.

[26] Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman, "Whisperx: Time-accurate speech transcription of long-form audio," *arXiv preprint arXiv:2303.00747*, 2023.

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

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

[29] Denis Eka Cahyani and Irene Patasik, "Performance comparison of tf-idf and word2vec models for emotion text classification," *Bulletin of Electrical Engineering and Informatics*, vol. 10, no. 5, pp. 2780–2788, 2021.

[30] Allison Koenecke, Anna Seo Gyeong Choi, Katelyn X Mei, Hilke Schellmann, and Mona Sloane, "Careless whisper: Speech-to-text hallucination harms," in *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, 2024, pp. 1672–1681.

[31] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever, "Language models are unsupervised multitask learners," 2019.

[32] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever, "Robust speech recognition via large-scale weak supervision," *ArXiv*, vol. abs/2212.04356, 2022.

[33] Zengwei Yao, Liyong Guo, Xiaoyu Yang, Wei Kang, Fangjun Kuang, Yifan Yang, Zengrui Jin, Long Lin, and Daniel Povey, "Zipformer: A faster and better encoder for automatic speech recognition," in *The Twelfth International Conference on Learning Representations*, 2023.

[34] Piotr Żelasko, Daniel Povey, Jan Trmal, Sanjeev Khudanpur, et al., "Lhotse: a speech data representation library for the modern deep learning ecosystem," *arXiv preprint arXiv:2110.12561*, 2021.

[35] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[37] Fangjun Kuang, Liyong Guo, Wei Kang, Long Lin, Mingshuang Luo, Zengwei Yao, and Daniel Povey, "Pruned rnn-t for fast, memory-efficient asr training," *arXiv preprint arXiv:2206.13236*, 2022.

[38] Alex Graves, "Sequence transduction with recurrent neural networks," *arXiv preprint arXiv:1211.3711*, 2012.

[39] Ghimire Rupak Raj, Bal Bal Krishna, and Poudyal Prakash, "Active learning approach for fine-tuning pre-trained asr model for a low-resourced language: A case study of nepali," in *Proceedings of the 20th International Conference on Natural Language Processing (ICON)*, 2023, pp. 82–89.